ENCYCLOPEDIA OF IMAGING SCIENCE AND TECHNOLOGY, VOLUME 2

Joseph P. Hornak

John Wiley & Sons, Inc.

ENCYCLOPEDIA OF

IMAGING SCIENCE AND TECHNOLOGY

VOLUME 2

ENCYCLOPEDIA OF IMAGING SCIENCE AND TECHNOLOGY

Editor
Joseph P. Hornak
Rochester Institute of Technology

Editorial Board
Christian DeMoustier
Scripps Institution of Oceanography
William R. Hendee
Medical College of Wisconsin
Jay M. Pasachoff
Williams College
William Philpot
Cornell University
Joel Pokorny
University of Chicago
Edwin Przyblowicz
Eastman Kodak Company
John Russ
North Carolina State University
Kenneth W. Tobin
Oak Ridge National Laboratory
Mehdi Vaez-Iravani
KLA-Tencor Corporation

Editorial Staff
Executive Publisher: **Janet Bailey**
Publisher: **Paula Kepos**
Executive Editor: **Jacqueline I. Kroschwitz**
Senior Managing Editor: **John Sollami**
Senior Associate Managing Editor: **Shirley Thomas**
Editorial Assistant: **Susanne Steitz**

ENCYCLOPEDIA OF

IMAGING SCIENCE AND TECHNOLOGY

VOLUME 2

Joseph P. Hornak
Rochester Institute of Technology
Rochester, New York

The *Encyclopedia of Imaging Science and Technology* is available Online in full color at www.interscience.wiley.com/eist



A Wiley-Interscience Publication
John Wiley & Sons, Inc.

This book is printed on acid-free paper. ♾

Copyright © 2002 by John Wiley & Sons, Inc., New York. All rights reserved.

Published simultaneously in Canada.

No part of this publication may be reproduced, stored in a retrieval system or transmitted in any form or by any means, electronic, mechanical, photocopying, recording, scanning or otherwise, except as permitted under Sections 107 or 108 of the 1976 United States Copyright Act, without either the prior written permission of the Publisher, or authorization through payment of the appropriate per-copy fee to the Copyright Clearance Center, 222 Rosewood Drive, Danvers, MA 01923, (978) 750-8400, fax (978) 750-4744. Requests to the Publisher for permission should be addressed to the Permissions Department, John Wiley & Sons, Inc., 605 Third Avenue, New York, NY 10158-0012, (212) 850-6011, fax (212) 850-6008, E-Mail: PERMREQ@WILEY.COM.

For ordering and customer service, call 1-800-CALL-WILEY.

Library of Congress Cataloging in Publication Data:

Encyclopedia of imaging science and technology/[edited by Joseph P. Hornak].
p. cm.
"A Wiley-Interscience publication."
Includes index.
ISBN 0-471-33276-3 (cloth:alk.paper)
1. Image processing–Encyclopedias. 2. Imaging systems–Encyclopedias. I. Hornak, Joseph P.
TA1632.E53 2001
621.36′7′03–dc21 2001046915

Printed in the United States of America.

10 9 8 7 6 5 4 3 2 1

ENCYCLOPEDIA OF

IMAGING SCIENCE AND TECHNOLOGY

VOLUME 2

L

LASER-INDUCED FLUORESCENCE IMAGING

STEPHEN W. ALLISON
WILLIAM P. PARTRIDGE
Engineering Technology Division
Oak Ridge National Laboratory
Knoxville, TN

INTRODUCTION

Fluorescence imaging is a tool of increasing importance in aerodynamics, fluid flow visualization, and nondestructive evaluation in a variety of industries. It is a means for producing two-dimensional images of real surfaces or fluid cross-sectional areas that correspond to properties such as temperature or pressure. This article discusses three major laser-induced fluorescence imaging techniques:

- Planar laser-induced fluorescence
- Phosphor thermography
- Pressure-sensitive paint

Since the 1980s, planar laser-induced fluorescence (PLIF) has been used for combustion diagnostics and to characterize gas- and liquid-phase fluid flow. Depending on the application, the technique can determine species concentration, partial pressure, temperature, flow velocity, or flow distribution/visualization. Phosphor thermography (PT) is used to image surface temperature distributions. Fluorescence imaging of aerodynamic surfaces coated with phosphor material for thermometry dates back to the 1940s, and development of the technique continues today. Imaging of fluorescence from pressure-sensitive paint (PSP) is a third diagnostic approach to aerodynamic and propulsion research discussed here that has received much attention during the past decade. These three methodologies are the primary laser-induced fluorescence imaging applications outside medicine and biology. As a starting point for this article, we will discuss PLIF first because it is more developed than the PT or PSP applications.

PLANAR LASER-INDUCED FLUORESCENCE

Planar laser-induced fluorescence (PLIF) in a fluid medium is a nonintrusive optical diagnostic tool for making temporally and spatially resolved measurements. For illumination, a laser beam is formed into a thin sheet and directed through a test medium. The probed volume may contain a mixture of various gaseous constituents, and the laser may be tuned to excite fluorescence from a specific component. Alternatively, the medium may be a homogenous fluid into which a fluorescing tracer has been injected. An imaging system normal to the plane of the imaging sheet views the laser-irradiated volume. Knowledge of the laser spectral characteristics, the spectroscopy of the excited material, and other aspects of the fluorescence collection optics is required for quantifying the parameter of interest.

A typical PLIF setup is shown schematically in Fig. 1. In this example, taken from Ref. 1, an ultraviolet laser probes a flame. A spherical lens of long focal length and a cylindrical lens together expand the beam and form it into a thin sheet. The spherical lens is specified to achieve the desired sheet thickness and depth of focus. This relates to the Rayleigh range, to be discussed later. An alternate method for planar laser imaging is to use the small diameter, circular beam typically emitted by the laser and scan it. Alternate sheet formation methods include combining the spherical lens with a scanned-mirror system and other scanning approaches. Fluorescence excited by the laser is collected by a lens or lens system, sometimes by intervening imaging fiberoptics, and is focused onto a camera's sensitive surface. In the example, this is performed by a gated intensified charge-coupled device (ICCD).

Background

Since its conception in the early 1980s, PLIF has become a powerful and widely used diagnostic technique. The PLIF diagnostic technique evolved naturally out of early imaging research based on Raman scattering (2), Mie scattering, and Rayleigh scattering along with 1-D LIF research (3). Planar imaging was originally proposed by Hartley (2), who made planar Raman-scattering measurements and termed the process Ramanography. Two-dimensional LIF-based measurements were made by Miles et al. (4) in 1978. Some of the first applications

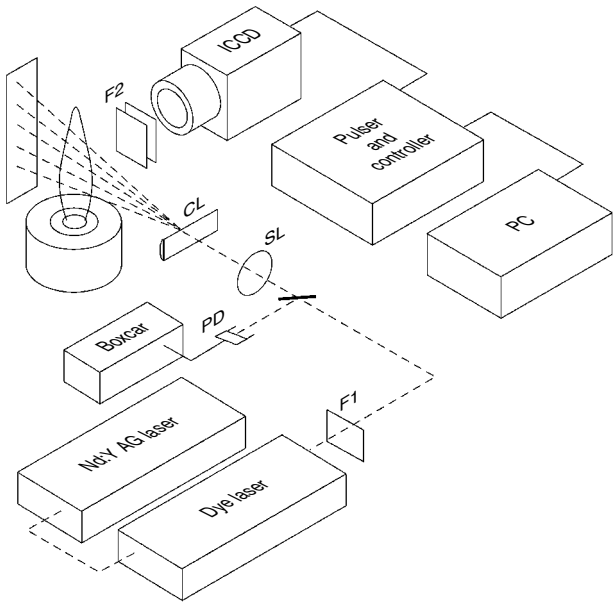


Figure 1. Representative PLIF configuration.

of PLIF, dating to the early 1980s, involved imaging the hydroxyl ion, OH^-, in a flame. In addition to its use for species imaging, PLIF has also been employed for temperature and velocity imaging. General reviews of PLIF have been provided by Alden and Svanberg (3) and Hanson et al. (5). Reference 6 also provides recent information on this method, as applied to engine combustion. Overall, it is difficult to state, with a single general expression, the range and limits of detection of the various parameters, (e.g., temperature, concentration, etc.), because there are so many variations of the technique. Single molecules can be detected and temperature measured from cryogenic to combustion ranges, depending on specific applications.

General PLIF Theory

The relationship between the measured parameter (e.g., concentration, temperature, pressure) and the fluorescent signal is unique to each measured parameter. However, the most fundamental relationship between the various parameters is provided by the equation that describes LIF or PLIF concentration measurements. Hence, this relationship is described generally here to clarify the different PLIF measurement techniques that derive from it. The equation for the fluorescent signal in volts (or digital counts on a per-pixel basis for PLIF measurements) is formulated as

$$S_D = (V_C f_B N_T) \cdot (\Gamma_{12,L} B_{12} I_\nu^o) \cdot \Phi \cdot \left(\eta \frac{\Omega}{4\pi}\right) \cdot GR\Delta t_L \qquad (1)$$

where

$$\Phi = \frac{A_{N,F}}{A_N + Q_e + W_{12} + W_{21} + Q_P}$$

and

where S_D: Measured fluorescent signal
V_C: Collection volume, i.e. portion of laser irradiated volume viewed by detection system
f_B: Boltzmann fraction in level 1.
N_T: Total number density of probe species
$\Gamma_{12,L}$: Overlap fraction (i.e., energy level line width divided by laser line width)
B_{12}: Einstein coefficient for absorption from energy level l to level 2
I_ν^o: Normalized laser spectral irradiance
Φ: Fluorescent quantum yield
η: Collection optics efficiency factor
Ω: Solid angle subtended by the collection optics
G: gain of camera (applicable if it is an intensified CCD)
R: CCD responsivity
Δt_L: Temporal full width half-maximum of the laser pulse
$A_{N,F}$: Spectrally filtered net spontaneous emission rate coefficient
A_N: Net spontaneous emission rate coefficient
Q_e: Fluorescence quenching rate coefficient
Q_p: Predissociation rate coefficient
W_{12}: Absorption rate coefficient
W_{21}: Stimulated emission rate coefficient

The individual terms in Eq. (1) have been grouped to provides a clear physical interpretation of the actions represented by the individual groups. Moreover, the groups have been arranged from left to right in the natural order that the fluorescent measurement progresses. The first parenthetical term in Eq. (1) is the number of probe molecules in the lower laser-coupled level. This is the fraction of the total number of probe molecules that are available for excitation. The second parenthetical term in Eq. (1) is the probability per unit time that one of the available molecules will absorb a laser photon and become electronically excited. Hence, following this second parenthetical term, a fraction of the total number of probed molecules has become electronically excited and has the potential to fluoresce. More detailed explanation is contained in Ref. 1.

The fluorescent quantum yield Φ represents the probability that one of the electronically excited probe molecules will relax to the ground electronic state by spontaneously emitting a fluorescent photon within the spectral bandwidth of the detection system. This fraction reflects the fact that spectral filtering is applied to the total fluorescent signal and that radiative as well as nonradiative (e.g., spontaneous emission and quenching, respectively) decay paths are available to the excited molecule. In the linear fluorescent regime and in the absence of other effects such as predissociation, the fluorescent yield essentially reduces to

$$\Phi \approx \frac{A_{N,F}}{(A_N + Q_E)}$$

so that the fluorescent signal is adversely affected by the quenching rate coefficient.

Within the third parenthetical term in Eq. (1) represents the net efficiency of the collection optics. This term accounts for reflection losses which occur at each optical surface. The next term, $\Omega/4\pi$, is the fraction of the fluorescence emitted by the electronically excited probe molecules that impinges on the detector surface (in this case, an ICCD) Ω is the finite solid angle of the collection optics. This captured fluorescence is then passed through an optical amplifier where it receives a gain G. The amplified signal is then detected by a given spectral responsivity R. The detection process in Eq. (1) produces a time-varying voltage or charge (depending on whether a PMT or ICCD detector is used.) This time-varying signal is then integrated over a specific gate time to produce the final measured fluorescent signal. Using Eq. (1), the total number density N_T, of the probed species, can be determined via a PLIF measurement of S_D provided that the remaining unknown parameters can be calculated or calibrated.

Investigation of the different terms of Eq. 1 suggests possible schemes for PLIF measurements of temperature, velocity, and pressure. For a given experimental setup (i.e., constant optical and timing parameters) and total number density of probe molecules, all of the terms in Eq. (1) are constants except for f_B, $\Gamma_{12,L}$, and Q_e. The Boltzmann fraction f_B varies in a known manner with temperature. The degree and type of variation

with temperature is unique to the lower laser-coupled level chosen for excitation. The overlap fraction $\Gamma_{12,L}$ varies with changes in the spectral line shape(s) of the absorption transition and/or the laser. Changes in velocity and pressure produce varying degrees of Doppler and pressure shift, respectively, in the absorption spectral profile (7–9). Hence, variations in these parameters will, in turn, produce changes in the overlap fraction. The electronic quenching rate coefficient varies with temperature, pressure, and major species concentrations. Detailed knowledge of the relationship between the variable of interest (i.e., temperature, pressure, or velocity) and the Boltzmann fraction f_B and/or the overlap fraction $\Gamma_{12,L}$ can be used in conjunction with Eq. (1) to relate the PLIF signal to the variable of choice. Often ratiometric techniques can be used to allow canceling of terms in Eq. (1) that are constant for a given set of experiments. Specific examples of different PLIF measurement schemes are given in the following review of pertinent literature.

PLIF Temperature Measurements

The theory behind PLIF thermometric measurements is the same as that developed for point LIF. Laurendeau (10) gives a review of thermometric measurements from a theoretical and historical perspective. Thermometric PLIF measurement schemes may be generally classified as monochromatic or bichromatic (two-line). Monochromatic methods employ a single laser. Bichromatic methods require two lasers to excite two distinct molecular rovibronic transitions simultaneously. In temporally stable environments (e.g., laminar flows), it is possible to employ bichromatic methods with a single laser by systematically tuning the laser to the individual transitions.

In bichromatic PLIF thermometric measurements, the ratio of the fluorescence from two distinct excitation schemes is formed pixel-by-pixel. If the two excitation schemes are chosen so that the upper laser-coupled level (i.e., exited state) is the same, then the fluorescent yields (Stern–Volmer factors) are identical. This is explained by Eckbreth in Ref. 11, an essential reference book for LIF and other laser-based flow and combustion diagnostic information. Hence, as evident from Eq. (1), the signal ratio becomes a sole function of temperature through the ratio of the temperature-dependent Boltzmann fractions for the two lower laser-coupled levels of interest.

Monochromatic PLIF thermometry is based on either the thermally assisted fluorescence (THAF) or the absolute fluorescence (ABF) methods. In THAF-based techniques, the temperature is related to the ratio of the fluorescent signals from the laser-excited level and from another higher level collisionally coupled to the laser-excited level. Implementing of this method requires detailed knowledge of the collisional dynamics, that occur in the excited level (9). In ABF-based techniques, the field of interest is uniformly doped or seeded, and fluorescence is monitored from a single rovibronic transition. The temperature-independent terms in Eq. (1) (i.e., all terms except f_B, $\Gamma_{12,L}$, and Φ) are determined through calibration. The temperature field may then be determined from the fluorescent field by assuming a known dependence of the Boltzmann fraction, the overlap fraction, and the quenching rate coefficient on temperature.

PLIF Velocity and Pressure Measurements

PLIF velocity and pressure measurements are based on changes in the absorption line-shape function of a probed molecule under the influence of variations in velocity, temperature, and pressure. In general, the absorption line-shaped function is Doppler-shifted by velocity, Doppler-broadened (Gaussian) by temperature, and collisionally broadened (Lorentzian) and shifted by pressure (10). These influences on the absorption line-shape function and consequently on the fluorescent signal via the overlap fraction of Eq. (1) provide a diagnostic path for velocity and pressure measurements.

The possibility of using a fluorescence-based Doppler-shift measurement to determine gas velocity was first proposed by Measures (12). The measurement strategy involved seeding a flow with a molecule that is excited by a visible, narrow-bandwidth laser. The Doppler shift could be determined by tuning the laser over the shifted absorption line and comparing the spectrally resolved fluorescence to static cell measurements. By probing the flow in two different directions, the velocity vector along each propagative direction could be determined from the resulting spectrally resolved fluorescence. For another early development, Miles et al. (4) used photographs to resolve spatially the fluorescence from a sodium-seeded, hypersonic nonreacting helium flow to make velocity and pressure measurements. The photographs of the fluorescence at each tuning position of a narrow-bandwidth laser highlighted those regions of the flow that had a specific velocity component. Although this work used a large diameter beam rather than a sheet for excitation, it evidently represents the first two-dimensional, LIF-based imaging measurement.

Another important method that is commonly used for visualizing flow characteristics involves seeding a flow with iodine vapor. The spectral properties are well characterized for iodine, enabling pressure and velocity measurements (13).

PLIF Species Concentration Measurements

The theory for PLIF concentration measurements is similar to that developed for linear LIF using broadband detection. The basic measurement technique involves exciting the specific rovibronic transition of a probe molecule (seeded or naturally occurring) and determining the probed molecule concentration from the resulting broadband fluorescence. Unlike ratiometric techniques, the fluorescent signal from this single-line method retains its dependence on the fluorescent yield (and therefore the electronic quenching rate coefficient). Hence, the local fluorescent signal depends on the number density the local probe molecule, of the Boltzmann fraction, the overlap fraction, and the electronic quenching rate coefficient. Furthermore, the Boltzmann fraction depends on the local temperature; the overlap fraction depends on the local temperature and pressure; and the electronic quenching rate coefficient depends on the local temperature, pressure,

and composition. This enhanced dependence of the fluorescent signal complicates determining of probed species concentrations from PLIF images. The difficulty in accurately determining the local electronic quenching rate coefficient, particularly in reacting environments, is the primary limitation to realizing quantitative PLIF concentration imaging (5). Nevertheless, methodologies for PLIF concentration measurements in quenching environments, based on modeling (1) and secondary measurements (2), have been demonstrated.

Useful fundamental information can be obtained from uncorrected, uncalibrated PLIF "concentration" images. Because of the species specificity of LIF, unprocessed PLIF images can be used to identify reaction zones, mixing regimes, and large-scale structures of flows. For instance, qualitative imaging of the formation of pollutant in a combustor can be used to determine optimum operating parameters.

The primary utility of PLIF concentration imaging remains its ability to image relative species distributions in a plane, rather than providing quantitative field concentrations. Because PLIF images are immediately quantitative in space and time (due to the high temporal and spatial resolution of pulsed lasers and ICCD cameras, respectively), qualitative species images may be used effectively to identify zones of species localization, shock wave positions, and flame-front locations (5).

The major experimental considerations limiting or pertinent to the realization of quantitative PLIF are

1. spatial cutoff frequency of the imaging system;
2. selection of imaging optics parameters (e.g., f number and magnification) that best balance spatial resolution and signal-level considerations;
3. image corrections implemented via postprocessing to account for nonuniformities in experimental parameters such as pixel responsivity and offset and laser sheet intensity; and
4. spatial variation in the fluorescent yield due to the electronic quenching rate coefficient.

Laser Beam Control

A distinctive feature of planar LIF is that the imaging resolution is controlled by the camera and its associated collection optics and also by the laser beam optics. For instance, the thinner a laser beam is focused, the higher the resolution. This section is a simple primer for lens selection and control of beam size.

The most important considerations for the choice of lenses are as follows. A simple lens will process light, to a good approximation, according to the thin lens equation,

$$\frac{1}{s_o} + \frac{1}{s_i} = \frac{1}{f}, \tag{2}$$

where s_o is the distance from an object to be imaged to the lens, $s_{i,}$, is the distance from the lens to where an image is formed, and f is the focal length of the lens, as shown in Fig. 2. In practice, this relationship is useful for imaging laser light from one plane (such as the position of an aperture or template) to another desired position

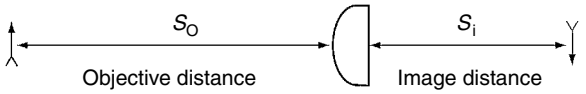


Figure 2. Simple lens imaging.

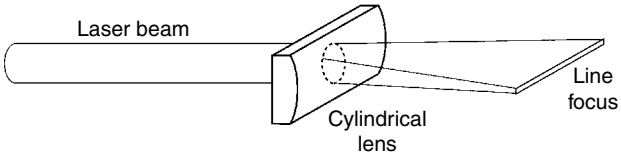


Figure 3. Line focus using a cylindrical lens.

in space whose magnification $M = -s_i/s_o$. If two lenses are used, the image of the first lens becomes the object distance for the second. For a well-collimated beam, the object distance is considered infinity, and thus the image distance is simply the focal length of the lens. There is a limit on how small the beam may be focused, and this is termed the diffraction limit. This minimum spot size w is given in units of length as $w = (1.22f\lambda)/D$, where λ is the wavelength of the light and D is the collimated beam diameter. If the laser beam is characterized by a divergence α, then the minimum spot size is $w = f\alpha$.

To form a laser beam into a sheet, sometimes termed "planarizing", a combination of two lenses, one spherical and the other cylindrical, is used. The spherical lens controls the spread, and the cylindrical lens controls the sheet thickness. The result is illustrated in Fig. 3. A laser sheet may be formed by combining spherical and cylindrical lenses; the cylindrical lens is used to achieve the desired sheet height, and a spherical lens is used to achieve the desired sheet thickness and Rayleigh range. Rayleigh range, a term that describes Gaussian beams (e.g., see Ref. 9), is the propagative distance required on either side of the beam waist to achieve a radius of $\sqrt{2}$ times the waist radius. The Rayleigh range z_o, is defined as $\pi \cdot w_o^2/\lambda$, where w_o is the waist radius that is used as a standard measurement of the waist-region length (i.e., length of the region of minimum and uniform sheet thickness). In general, longer focal length lenses produce longer Rayleigh ranges. In practice, lens selection is determined by the need to make the Rayleigh range greater than the lateral imaged distance. In general, because longer focal length lenses produce wider sheet-waist thicknesses, the specified sheet thickness and lateral image extent must be balanced.

PHOSPHOR THERMOGRAPHY

Introduction

As conceived originally, phosphor thermography was intended foremost to be a means of depicting two-dimensional temperature patterns on surfaces. In fact, during its first three decades of existence, the predominant use of the technique was for imaging applications in aerodynamics (14). The method was termed "contact thermometry" because the phosphor was in contact with the surface to be monitored. The overall approach, however,

has largely been overshadowed by the introduction of modern infrared thermal imaging techniques, several of which have evolved into commercial products that are used in a wide range of industrial and scientific applications. Yet, phosphor thermography (PT) remains a viable method for imaging and discrete point measurements.

A comprehensive survey of fluorescence-based thermometry is provided in Ref. 14 and 15. The former emphasizes noncontact phosphor applications, and the latter includes the use of fluorescent crystals, glasses, and optical fibers as temperature sensors, as well as phosphors. Phosphor thermography exploits the temperature dependence of powder materials identical or similar to phosphors used commercially in video and television displays, fluorescent lamps, X-ray scintillating screens, etc. Typically, a phosphor is coated onto a surface whose temperature is to be measured. The coating is illuminated by an ultraviolet source, which induces fluorescence. The emitted fluorescence may be captured by either a nonimaging or an imaging detector. Several fluorescent properties are temperature-dependent. The fluorescence may change in magnitude and/or spectral distribution due to a change in temperature. Figure 4 shows a spectrum of Gd_2O_2S:Tb, a representative phosphor. The emission from this material originates from atomic transitions of the rare-earth activator Tb. At ambient temperatures, the ratio of emission intensities at 410 and 490 nm changes drastically with temperature from ambient to about 120 F. The other emission lines in the figure do not change until much higher temperatures are achieved. Thus the ratio indicates temperature in the said range, as shown in Fig. 5. Figure 6 shows a typical setup that depicts illumination either with laser light emerging from a fiber or an ultraviolet lamp. If the illumination source is pulsed, fluorescence will persist for a period of time after the illumination is turned off. The intensity I decreases, ideally according to $I = e^{-t/\tau}$, where the time required for decreasing by $1/e$

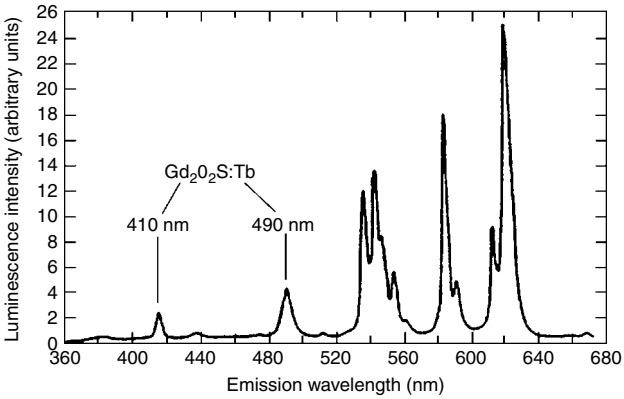


Figure 4. Gd_2O_2S:Tb spectrum.

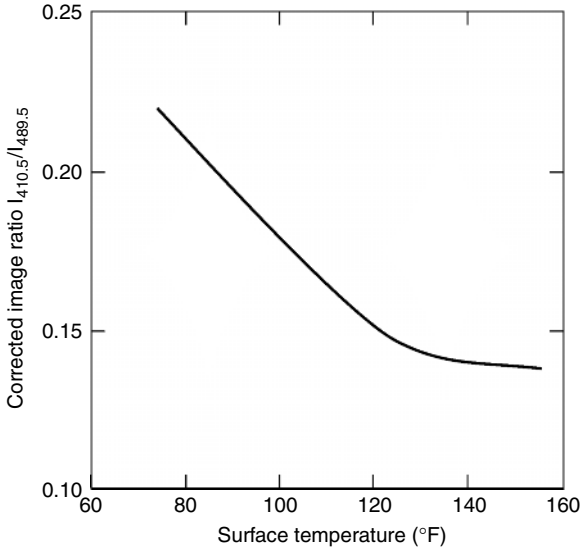


Figure 5. Intensity ratio versus temperature.

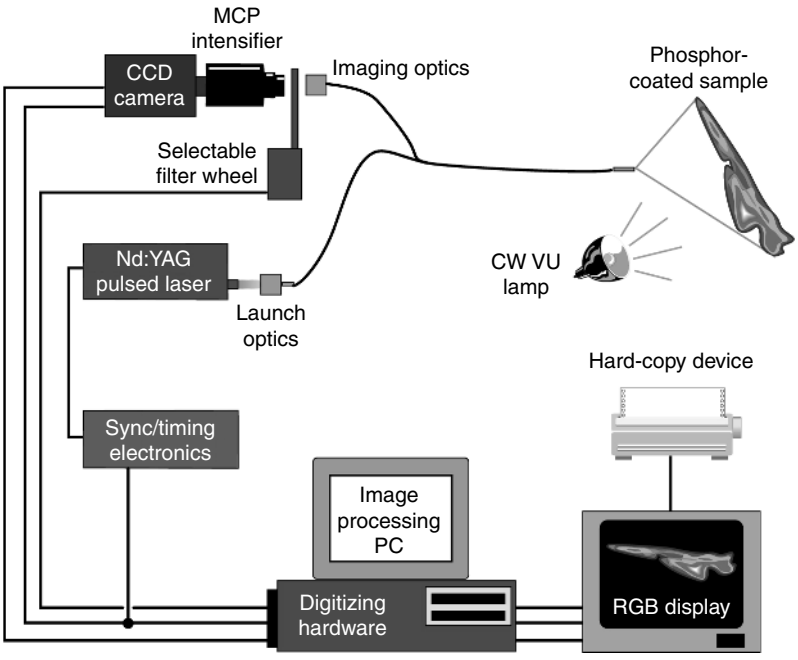


Figure 6. A phosphor imaging system.

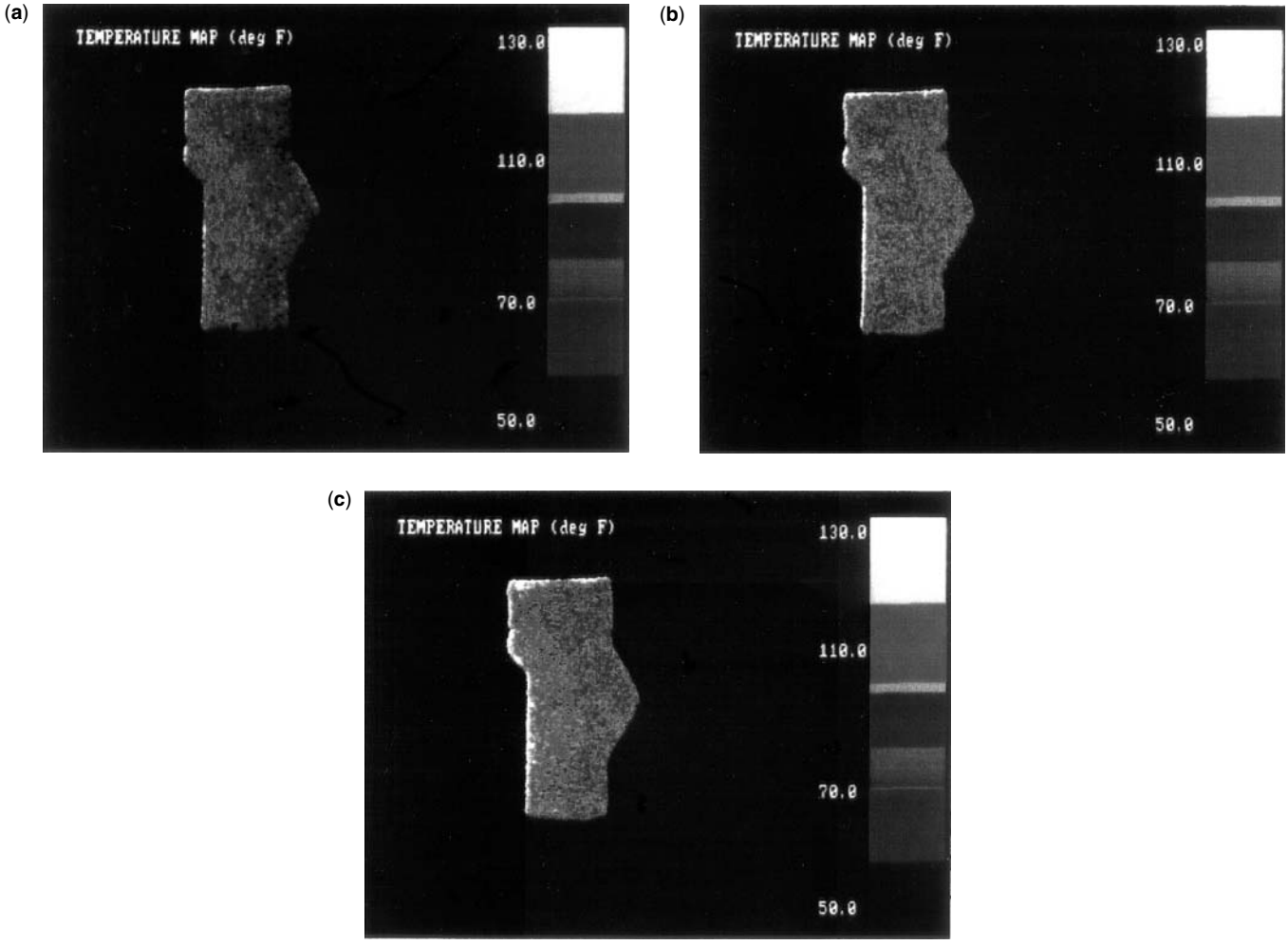


Figure 7. False-color thermograph of heated turbine blade.

is termed the characteristic decay time τ, also known as lifetime. The decay time is very temperature-dependent and in most nonimaging applications, the decay time is measured to ascertain temperature. For imaging, it is usually easier to implement the ratio method (16). Figure 7 shows false color images of a heated turbine blade (17). Temperature can be measured from about 12 K to almost 2,000 K. In some cases, a temperature resolution of less than 0.01 K has been achieved.

Applications

Why use phosphor thermometry when infrared techniques work so well for many imaging applications? As noted by Bizzak and Chyu, conventional thermometric methods are not satisfactory for temperature and heat transfer measurements that must be made in the rapidly fluctuating conditions peculiar to a microscale environment (18). They suggested that thermal equilibrium on the atomic level might be achieved within 30 ns and, therefore, the instrumentation system must have a very rapid response time to be useful in microscale thermometry. Moreover, its spatial resolution should approach the size of an individual phosphor particle. This can be specified and may range from $<1\,\mu$m to about 25 μm. In their effort to develop a phosphor imaging system based on La_2O_2S:Eu, they used a frequency-tripled Nd:YAG laser. The image was split, and the individual beams were directed along equal-length paths to an intensified CCD detector. The laser pulse duration was 4 ns, and the ratio of the $^5D_2$ to $^5D_0$ line intensities yielded the temperature. A significant finding is that they were able to determine how the measurement accuracy varied with measurement area. The maximum error found for a surface the size of a 1×1 pixel in their video system was 1.37 °C, and the average error was only 0.09 °C.

A particularly clever conception by Goss et al. (19) illustrates another instance where the method is better suited than infrared emission methods. It involved the visualization through a flame produced by condensed-phase combustion of solid rocket propellant. They impregnated the fuel under test with YAG:Dy, and used the ratio of an F-level band at 496 nm to a G-level band at 467 nm as the signal of interest. Because of thermalization, the intensity of the 467-nm band increased

in comparison with the 496-nm band across the range from ambient to 1,673 K, the highest temperature they were able to attain. At that temperature, the blackbody emission introduces a significant background component into the signal, even within the narrow passband of the spectrometer that was employed. To mitigate this, they used a Q-switched Nd:YAG laser (frequency-tripled to 355 nm). The detectors in this arrangement included an intensified (1,024-element) diode array and also an ICCD detector. They were gated for a period of 10 μs. To simulate combustion in the laboratory, the phosphor was mixed with a low melting point (400 K) plastic, and the resulting blend was heated with a focused CO_2 laser which ignited a flame that eroded the surface. A time history of the disintegrating plastic surface was then obtained from the measurement system. Because of the short duration of the fluorescence, the power of the laser, and the gating of the detector, they were able to measure temporally resolved temperature profiles in the presence of the flame.

Krauss, Laufer, and colleagues at the University of Virginia used laser-induced phosphor fluorescent imaging for temperatures as high as 1,100 °C. Their efforts during the past decade have included a simultaneous temperature and strain sensing method, which they pioneered (20,21). For this, they deposit closely spaced thin stripes of phosphor material on the test surface. A camera views laser-induced fluorescence from the stripes. An image is acquired at ambient, unstressed conditions and subsequently at temperature and under stress. A digital moiré pattern of the stripes is produced by comparing the images before and after. The direction and magnitude of the moiré pattern indicates strain. The ratio of the two colors of the fluorescence yields temperature.

PRESSURE-SENSITIVE PAINT

Background

Pressure-sensitive paints are coatings that use luminescing compounds that are sensitive to the presence of oxygen. References 22–26 are reviews of the subject. There several varieties of PSPs discussed in the literature. Typically, they are organic compounds that have a metal ligand. Pressure-sensitive paint usually consists of a PSP compound mixed with a gas-permeable binder. On the molecular level, a collision of an oxygen molecule with the compound prevents fluorescence. Thus, the greater the oxygen concentration, the less the fluorescence. This application of fluorescence is newer than planar laser-induce fluorescence and phosphor thermography, but it is a field of rapidly growing importance. The utility of imaging pressure profiles of aerodynamic surfaces in wind tunnels and inside turbine engines, as well as for flights in situ has spurred this interest.

For the isothermal case, the luminescent intensity I and decay time τ of a pressure-sensitive paint depend on oxygen pressure as

$$\frac{\tau_0}{\tau} = \frac{I_0}{I} = 1 + K \cdot P_o = 1 + k_q \tau_o P_o, \tag{3}$$

where $K = k_q \cdot \tau_0$ and the subscript 0 refers to the respective values at zero oxygen pressure (vacuum). In practice, rather than performing the measurements under vacuum, a reference image is taken at atmospheric conditions where pressure and temperature are well established. The common terminology is "wind-on" and "wind-off" where the latter refers to reference atmospheric conditions. The equations may then be rearranged to obtain (where $A(T)$ and $B(T)$ are functions of temperature)

$$\frac{I_{REF}}{I} = A(T) + B(T)\frac{P}{P_{REF}}. \tag{4}$$

Pressure-sensitive paint installations typically use an imaging system that has a CCD camera connected to a personal computer and some type of frame digitizing hardware. Temperature also plays a role in the oxygen quenching because the collision rate is temperature-dependent. The precise temperature dependence may be further complicated because there may be additional thermal quenching mechanisms and the absorption coefficient can be somewhat affected by temperature. Therefore, for unambiguous pressure measurement, the temperature should either be uniform or measured by other means. One way to do this is to incorporate the fluorescent material in an oxygen-impermeable host. The resulting coating is termed a temperature-sensitive paint (TSP) because oxygen quenching is prevented in this case. Another means is to mix a thermographic phosphor with the PSP. There are advantages and disadvantages to these methods. For example, the fluorescent emissions of the PSP and TP or TSP should be distinguishable from each other. This is an active area of research, and other approaches are being investigated.

A PSP may be illuminated by any of a variety of pulsed or continuous mercury or rare-gas discharge lamps. Light sources that consist of an array of bright blue LEDs are of growing importance for this application. Laser beams may be used which are expanded to illuminate the object fully. Alternatively, the lasers are scanned as described earlier in the planar LIF section.

Applications

This is a young but rapidly changing field. Some of the most important initial work using oxygen-quenched materials for aerodynamic pressure measurements was done by Russian researchers in the early 1980s (26). In the United States, the method was pioneered by researchers at the University of Washington and collaborators (28,29) In the early 1990s, PSP became a topic for numerous conference presentations that reported a wide range of low-speed, transonic, and supersonic aerodynamic

applications. Measurement of rotating parts is important and is one requirement that drives the search for short decay time luminophors. The other is the desire for fast temporal response of the sensor. Laboratory setups can be compact, simple, and inexpensive. On the other hand, a 16-ft diameter wind tunnel at the Air Force's Arnold Engineering Development Center has numerous cameras that viewing the surface from a variety of angles (30). In this application, significant progress has been achieved in computational modeling to remove various light scattering effects that can be significant for PSP work (31).

The collective experience from nonimaging applications of PSPs and thermographic phosphors shows that either decay time or phase measurement usually presents the best option for determining pressure (or temperature) due to immunity from various noise sources and inherently better sensitivity. Therefore, phase-sensitive imaging and time-domain imaging are approaches that are being explored and could prove very useful (32,33).

Research into improved PSP material is proceeding in a variety of directions. Various biluminophor schemes are being investigated for establishing temperature as well as pressure. Phosphors and laser dyes are receiving attention because they exhibit temperature dependence but no pressure dependence. Not only is the fluorescing material important but the host matrix is as well. The host's permeability to oxygen governs the time response to pressure changes. Thus, there is a search for new host materials to enable faster response. Now, the maximum time response rate is about 100 kHz. This is a fast moving field, as evidenced by the fact that only two years ago, one of the authors was informed that the maximum response rate was about 1 kHz. In aerodynamic applications, the method for coating surfaces is very important, especially for scaled studies of aerodynamic heating and other effects that depend on model surface properties. Work at NASA Langley on scaled models uses both phosphors and pressure-sensitive paints (34,35). Scaling considerations demand a very smooth surface finish.

FUTURE ADVANCES FOR THESE TECHNIQUES

Every advance in spectroscopy increases the number of applications for planar laser-induced fluorescence. One of the main drivers for this is laser technology. The variety of lasers available to the user continues to proliferate. As a given type of laser becomes smaller and less expensive, its utility in PLIF applications is expanded and sometimes facilitates the movement of a technique out of the lab and into the field. New laser sources always enable new types of spectroscopies which produce new information on various spectral properties that can be exploited by PLIF. Improvements in producing narrower linewidths, shorter pulse lengths, higher repetition rates, better beam quality, and wider frequency range, as the case may be, will aid PLIF.

In contrast, phosphor thermometry and pressure-sensitive paint applications usually require a laser only for situations that require high intensity due to remote distances, or the need to use fiber optics to access difficult to reach surfaces. Those situations usually do not involve imaging. Improvements to incoherent light sources are likely to have a greater impact on PT and PSP. For example, blue LEDs are sufficiently bright for producing useful fluorescence and are available commercially in arrays for spectroscopic applications. The trend will be to increased output and shorter wavelengths. However, one area of laser technology that could have a significant impact is the development of inexpensive blue and ultraviolet diode lasers.

The field of PSP application is the newest of the technologies discussed here and it has been growing the fastest. Currently, applications are limited to pressures of a few atmospheres. Because PSPs used to date are organic materials, the temperatures at which they can operate and survive are limited to about 200 °C. Development of inorganic salts and other materials are underway to increase the temperature and pressure range accessible to the technique.

One PSP material will not serve all possible needs. There may eventually be hundreds of PSP materials that will be selected on the basis of (1) chemical compatibility in the intended environment, (2) the need to match excitation and emission spectral characteristics with available light sources, (3) decay time considerations that are important for moving surfaces, (4) pressure and temperature range, (5) frequency response required, and (6) the specifics of adhesion requirements in the intended application. The PSP materials of today are, to our knowledge, all based on oxygen quenching. However, materials will be developed that are sensitive to other substances as well.

Acknowledgments

Oak Ridge National Laboratory is managed by UT-Battelle, LLC, for the U.S. Dept. of Energy under contract DE-AC05-00OR22725.

BIBLIOGRAPHY

1. W. P. Partridge, Doctoral Dissertation, Purdue University, 1996.
2. D. L. Hartley, M. Lapp and C. M. Penney, eds., *Laser Raman Gas Diagnostics*, Plenum Press, NY, 1974.
3. M. Alden and S. Svanberg, *Proc. Laser Inst. Am.* **47**, 134–143 (1984).
4. R. B. Miles, E. Udd, and M. Zimmerman, *Appl. Phys. Lett.* **32**, 317–319 (1978).
5. R. K. Hanson, J. M. Seitzman, and P. H. Paul, *Appl. Phys. B* **50**, 441–454 (1990).
6. P. H. Paul, J. A. Gray, J. L. Durant, and J. W. Thoman, *AIAA J.* **32**, 1,670–1,675 (1994).
7. P. Partridge and N. M. Laurendeau, to be published in *Appl. Phy. B.*
8. A. M. K. P Taylor, ed., *Instrumentation for Flows with Combustion*, Academic Press, London, England 1993.
9. W. Demtroder, *Laser Spectroscopy, Basic Concepts and Instrumentations*, Springer-Verlag, NY, 1988.
10. N. M. Laurendeau, *Prog. Energy Combust. Sci.* **14**, 147–170 (1988).

11. A. C. Eckbreth, *Laser Diagnostics for Combustion Temperature and Species*, Abacus Press, Cambridge, CA, 1988.
12. R. M. Measures, *J. Appl. Phys.* **39**, 5,232–5,245 (1968).
13. Flow Visualization VII: Proc. Seventh Int. Symp. Flow Visualization J. P. Crowder, ed., 1995.
14. S. W. Allison and G. T. Gillies, *Rev. Sci. Instrum.* **68**(7), 1–36 (1997).
15. K. T. V. Grattan and Z. Y. Zhang, *Fiber Optic Fluorescence Thermometry*, Chapman & Hall, London, 1995.
16. K. W. Tobin, G. J. Capps, J. D. Muhs, D. B. Smith, and M. R. Cates, *Dynamic High-Temperature Phosphor Thermometry*. Martin Marietta Energy Systems, Inc. Report No. ORNL/ATD-43, August 1990.
17. B. W. Noel, W. D. Turley, M. R. Cates, and K. W. Tobin, Two_Dimensional Temperature Mapping using Thermographic Phosphors. Los Alamos National Laboratory Technical Report No. LA_UR_90_1534, May 1990.
18. D. J. Bizzak and M. K. Chyu, *Rev. Sci. Instrum.* **65**, 102 (1994).
19. P. Goss, A. A. Smith, and M. E. Post, *Rev. Sci. Instrum.* **60**(12), 3,702–3,706 (1989).
20. R. P. Marino, B. D. Westring, G. Laufer, R. H. Krauss, and R. B. Whitehurst, *AIAA J.* **37**, 1,097–1,101 (1999).
21. A. C. Edge, G. Laufer, and R. H. Krauss, *Appl. Opt.* **39**(4), 546–553 (2000).
22. B. C. Crites, in *Measurement Techniques Lecture Series 1993–05*, von Karman Institute for Fluid Dynamics, 1993.
23. B. G. McLachlan, and J. H. Bell, *Exp. Thermal Fluid Sci.* **10**(4), 470–485 (1995).
24. T. Liu, B. Campbell, S. Burns, and J. Sullivan, *Appl. Mech. Rev.* **50**(4), 227–246 (1997).
25. J. H. Bell, E. T. Schairer, L. A. Hand, and R. Mehta, to be published in *Annu. Rev. Fluid Mech.*
26. V. Mosharov, V. Radchenko, and S. Fonov, *Luminescent Pressure Sensors in Aerodynamic Experiments*. Central Aerohydrodynamic Institute and CW 22 Corporation, Moscow, Russia, 1997.
27. M. M. Ardasheva, L. B. Nevshy, and G. E. Pervushin, *J. Appl. Mech. Tech. Phys.* **26**(4), 469–474 (1985).
28. M. Gouterman, *J. Chem. Ed.* **74**(6), 697–702 (1997).
29. J. Kavandi and J. P. Crowder, *AIAA* Paper 90-1516, 1990.
30. M. E. Sellers and J. A. Brill, *AIAA* Paper 94-2481, 1994.
31. W. Ruyten, *Rev. Sci. Instrum.* **68**(9), 3,452–3,457 (1997).
32. C. W. Fisher, M. A. Linne, N. T. Middleton, G. Fiechtner, and J. Gord, *AIAA* Paper 99-0771.
33. P. Hartmann and W. Ziegler, *Anal. Chem.* **68**, 4,512–4,514 (1996).
34. Quantitative Surface Temperature Measurement using Two-Color Thermographic Phosphors and Video Equipment, US Pat. 4,885,633 December 5, 1989 G. M. Buck.
35. G. M. Buck, *J. Spacecraft Rockets* **32**(5), 791–794 (1995).

LIDAR

P. S. Argall
R. J. Sica
The University of Western Ontario
London, Ontario, Canada

INTRODUCTION

Light detection and ranging (lidar) is a technique in which a beam of light is used to make range-resolved remote measurements. A lidar emits a beam of light, that interacts with the medium or object under study. Some of this light is scattered back toward the lidar. The backscattered light captured by the lidar's receiver is used to determine some property or properties of the medium in which the beam propagated or the object that caused the scattering.

The lidar technique operates on the same principle as radar; in fact, it is sometimes called laser radar. The principal difference between lidar and radar is the wavelength of the radiation used. Radar uses wavelengths in the radio band whereas lidar uses light, that is usually generated by lasers in modern lidar systems. The wavelength or wavelengths of the light used by a lidar depend on the type of measurements being made and may be anywhere from the infrared through the visible and into the ultraviolet. The different wavelengths used by radar and lidar lead to the very different forms that the actual instruments take.

The major scientific use of lidar is for measuring properties of the earth's atmosphere, and the major commercial use of lidar is in aerial surveying and bathymetry (water depth measurement). Lidar is also used extensively in ocean research (1–5) and has several military applications, including chemical (6–8) and biological (9–12) agent detection. Lidar can also be used to locate, identify, and measure the speed of vehicles (13). Hunters and golfers use lidar-equipped binoculars for range finding (14,15).

Atmospheric lidar relies on the interactions, scattering, and absorption, of a beam of light with the constituents of the atmosphere. Depending on the design of the lidar, a variety of atmospheric parameters may be measured, including aerosol and cloud properties, temperature, wind velocity, and species concentration.

This article covers most aspects of lidar as it relates to atmospheric monitoring. Particular emphasis is placed on lidar system design and on the Rayleigh lidar technique. There are several excellent reviews of atmospheric lidar available, including the following:

Lidar for Atmospheric Remote Sensing (16) gives a general introduction to lidar; it derives the lidar equation for various forms of lidar including Raman and differential absorption lidar (DIAL). This work includes details of a Raman and a DIAL system operated at NASA's Goddard Space Flight Center. *Lidar Measurements: Atmospheric Constituents, Clouds, and Ground Reflectance* (17) focuses on the differential absorption and DIAL techniques as well as their application to monitoring aerosols, water vapor, and minor species in the troposphere and lower stratosphere. Descriptions of several systems are given, including the results of measurement programs using these systems. *Optical and Laser Remote Sensing* (18) is a compilation of papers that review a variety of lidar techniques and applications. *Lidar Methods and Applications* (19) gives an overview of lidar that covers all areas of atmospheric monitoring and research, and emphasizes

the role lidar has played in improving our understanding of the atmosphere. *Coherent Doppler Lidar Measurement of Winds* (20) is a tutorial and review article on the use of coherent lidar for measuring atmospheric winds. *Lidar for Atmospheric and Hydrospheric Studies* (21) describes the impact of lidar on atmospheric and to a lesser extent oceananic research particularly emphasizing work carried out during the period 1990 to 1995. This review details both the lidar technology and the environmental research and monitoring undertaken with lidar systems.

Laser Remote Sensing (22) is a comprehensive text that covers lidar. This text begins with chapters that review electromagnetic theory, which is then applied to light scattering in the atmosphere. Details, both theoretical and practical, of each of the lidar techniques are given along with many examples and references to operating systems.

HISTORICAL OVERVIEW

Synge in 1930 (23) first proposed the method of determining atmospheric density by detecting scattering from a beam of light projected into the atmosphere. Synge suggested a scheme where an antiaircraft searchlight could be used as the source of the beam and a large telescope as a receiver. Ranging could be accomplished by operating in a bistatic configuration, where the source and receiver were separated by several kilometres. The receiver's field-of-view (FOV) could be scanned along the searchlight beam to obtain a height profile of the scattered light's intensity from simple geometric considerations. The light could be detected by using a photoelectric apparatus. To improve the signal level and thus increase the maximum altitude at which measurements could be made, Synge also suggested that a large array of several hundred searchlights could be used to illuminate the same region of the sky.

The first reported results obtained using the principles of this method are those of Duclaux (24) who made a photographic recording of the scattered light from a searchlight beam. The photograph was taken at a distance of 2.4 km from the searchlight using an $f/1.5$ lens and an exposure of 1.5 hours. The beam was visible on the photograph to an altitude of 3.4 km. Hulbert (25) extended these results in 1936 by photographing a beam to an altitude of 28 km. He then made calculations of atmospheric density profiles from the measurements.

A monostatic lidar, the typical configuration for modern systems, has the transmitter and receiver at the same location, (Fig. 1). Monostatic systems can be subdivided into two categories, coaxial systems, where the laser beam is transmitted coaxially with the receiver's FOV, and biaxial systems, where the transmitter and receiver are located adjacent to each other. Bureau (26) first used a monostatic system in 1938. This system was used for determining cloud base heights. As is typical with a monostatic system, the light source was pulsed, thereby enabling the range at which the scattering occured to be determined from the round-trip time of the scattered light pulse, as shown in Fig. 2.

By refinements of technique and improved instrumentation, including electrical recording of backscattered light

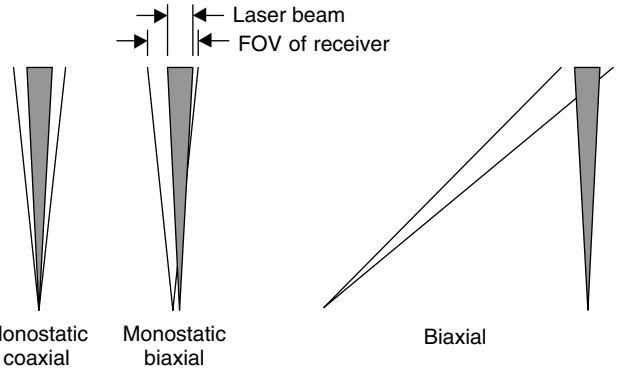


Figure 1. Field of view arrangements for lidar laser beam and detector optics .

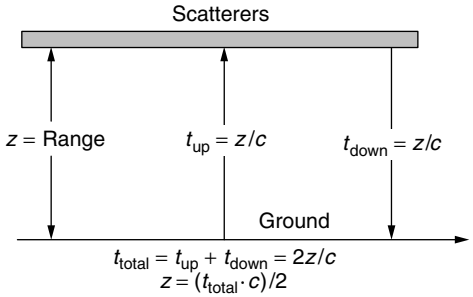


Figure 2. Schematic showing determination of lidar range.

intensity, Elterman (27) calculated density profiles up to 67.6 km. He used a bistatic system where the transmitter and receiver were 20.5 km apart. From the measured density profiles, Elterman calculated temperature profiles using the Rayleigh technique.

Friedland et al. (28) reported the first pulsed monostatic system for measuring atmospheric density in 1956. The major advantage of using a pulsed monostatic lidar is that for each light pulse fired, a complete altitude-scattering profile can be recorded, although commonly many such profiles are required to obtain measurements that have a useful signal-to-noise ratio. For a bistatic lidar, scattering can be detected only from a small layer in the atmosphere at any one time, and the detector must be moved many times to obtain an altitude profile. The realignment of the detector can be difficult due to the large separations and the strict alignment requirements of the beam and the FOV of the detector system. Monostatic lidar inherently averages the measurements at all altitudes across exactly the same period, whereas a bistatic system takes a snapshot of each layer at a different time.

The invention of the laser (29) in 1960 and the giant pulse or Q-switched laser (30) in 1962 provided a powerful new light source for lidar systems. Since the invention of the laser, developments in lidar have been closely linked to advances in laser technology. The first use of a laser in a lidar system was reported in 1962 by Smullins and Fiocco (31), who detected laser light scattered from the lunar surface using a ruby laser that fired 0.5-J pulses at 694 nm. In the following year, these same workers

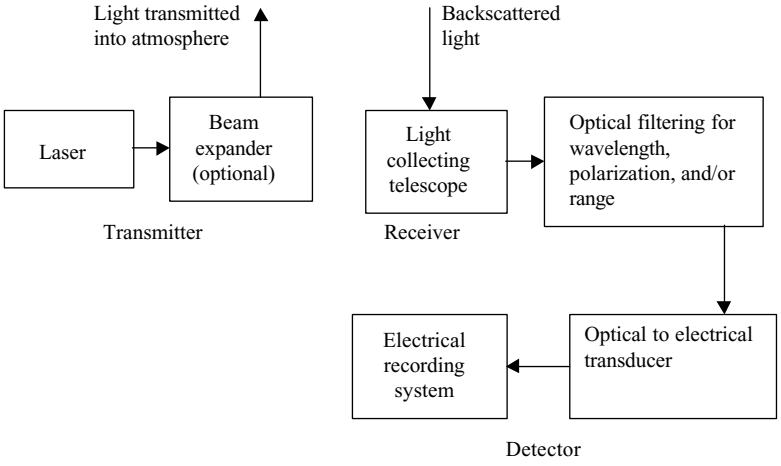


Figure 3. Block diagram of a generic lidar system.

reported the detection of atmospheric backscatter using the same laser system (32).

LIDAR BASICS

The first part of this section describes the basic hardware required for a lidar. This can be conveniently divided into three components: the transmitter, the receiver, and the detector. Each of these components is discussed in detail. Figure 3, a block diagram of a generic lidar system, shows how the individual components fit together.

In the second part of this section, the lidar equation that gives the signal strength measured by a lidar in terms of the physical characteristics of the lidar and the atmosphere is derived.

Transmitter

The purpose of the transmitter is to generate light pulses and direct them into the atmosphere. Figure 4 shows the laser beam of the University of Western Ontario's Purple Crow lidar against the night sky. Due to the special characteristics of the light they produce, pulsed lasers are ideal as sources for lidar systems. Three properties of a pulsed laser, low beam divergence, extremely narrow spectral width, and short intense pulses, provide significant advantages over white light as the source for a lidar.

Generally, it is an advantage for the detection system of a lidar to view as small an area of the sky as possible as this configuration keeps the background low. Background is light detected by the lidar that comes from sources other than the transmitted laser beam such as scattered or direct sunlight, starlight, moonlight, airglow, and scattered light of anthropogenic origin. The larger the area of the sky that the detector system views, that is, the larger the FOV, the higher the measured background. Therefore, it is usually preferable for a lidar system to view as small an area of the sky as possible. This constraint is especially true if the lidar operates in the daytime (33–35), when scattered sunlight becomes the major source of background. Generally, it is also best if the entire laser beam falls within the FOV of the detector system as this configuration gives maximum system efficiency. The divergence of the laser beam should be sufficiently small, so that it remains within the FOV of the receiver system in all ranges of interest.



Figure 4. Laser beam transmitted from the University of Western Ontario's Purple Crow lidar. The beam is visible from several kilometers away and often attracts curious visitors. See color insert.

A simple telescope arrangement can be used to decrease the divergence of a laser beam. This also increases the diameter of the beam. Usually, only a small reduction in the divergence of a laser beam is required in a lidar system, because most lasers have very low divergence. Thus, a small telescope, called a beam expander, is usually

all that is required to obtain a sufficiently well-collimated laser beam for transmission into the atmosphere.

The narrow spectral width of the laser has been used to advantage in many different ways in different lidar systems. It allows the detection optics of a lidar to spectrally filter incoming light and thus selectively transmit photons at the laser wavelength. In practice, a narrowband interference filter is used to transmit a relatively large fraction of the scattered laser light (around 50%) while transmitting only a very small fraction of the background white light. This spectral selectivity means that the signal-to-background ratio of the measurement will be many orders of magnitude greater when a narrowband source and a detector system interference filter are used in a lidar system.

The pulsed properties of a pulsed laser make it an ideal source for a lidar, as this allows ranging to be achieved by timing the scattered signal. A white light or a continuous-wave (cw) laser can be mechanically or photo electrically chopped to provide a pulsed beam. However, the required duty cycle of the source is so low that most of the energy is wasted. To achieve ranging, the length of the laser pulses needs to be much shorter than the required range resolution, usually a few tens of meters. Therefore, the temporal length of the pulses needs to be less than about 30 ns. The pulse-repetition frequency (PRF) of the laser needs to be low enough that one pulse has time to reach a sufficient range, so that it no longer produces any signal before the next pulse is fired. This constraint implies a maximum PRF of about 20 kHz for a lidar working at close range. Commonly, much lower laser PRFs are used because decreasing the PRF reduces the active observing time of the receiver system and therefore, reduces the background. High PRF systems do have the distinct advantage of being able to be made "eye-safe" because the energy transmitted in each pulse is reduced (36).

Using the values cited for the pulse length and the PRF gives a maximum duty cycle for the light source of about 0.06%. This means that a chopped white light or cw laser used in a lidar would have effective power of less than 0.06% of its actual power. However, for some applications, it is beneficial to use cw lasers and modulation code techniques for range determination (37,38).

The type of laser used in a lidar system depends on the physical quantity that the lidar has been designed to measure. Some measurements require a very specific wavelength (i.e., resonance–fluorescence) or wavelengths (i.e., DIAL) and can require complex laser systems to produce these wavelengths, whereas other lidars can operate across a wide wavelength range (i.e., Rayleigh, Raman and aerosol lidars). The power and pulse-repetition frequency of a laser must also match the requirements of the measurements. There is often a compromise of these quantities, in addition to cost, in choosing from the types of lasers available.

Receiver

The receiver system of a lidar collects and processes the scattered laser light and then directs it onto a photodetector, a device that converts the light to an electrical signal. The primary optic is the optical element that collects the light scattered back from the atmosphere and focuses it to a smaller spot. The size of the primary optic is an important factor in determining the effectiveness of a lidar system. A larger primary optic collects a larger fraction of the scattered light and thus increases the signal measured by the lidar. The size of the primary optic used in a lidar system may vary from about 10 cm up to a few meters in diameter. Smaller aperture optics are used in lidar systems that are designed to work at close range, for example, a few 100 meters. Larger aperture primary optics are used in lidar systems that are designed to probe the middle and upper regions of the Earth's atmosphere where the returned signal is a much smaller fraction of the transmitted signal (39,40). Smaller primary optics may be lenses or mirrors; the larger optics are typically mirrors. Traditional parabolic glass telescope primary mirrors more than about a half meter in diameter are quite expensive, and so, some alternatives have been successfully used with lidar systems. These alternatives include liquid-mirror telescopes (LMTs) (36,41) (Fig. 5), holographic elements (42,43), and multiple smaller mirrors (44–46).

After collection by the primary optic, the light is usually processed in some way before it is directed to the detector system. This processing can be based on wavelength, polarization, and/or range, depending on the purpose for which the lidar has been designed.

The simplest form of spectral filtering uses a narrowband interference filter that is tuned to the laser wavelength. This significantly reduces the background, as described in the previous section, and blocks extraneous signals. A narrowband interference filter that is typically around 1 nm wide provides sufficient rejection of background light for a lidar to operate at nighttime. For daytime use, a much narrower filter is usually employed (47–49). Complex spectral filtering schemes



Figure 5. Photograph of the 2.65-m diameter liquid mercury mirror used at the University of Western Ontario's, Purple Crow lidar. See color insert.

are often used for Doppler and high-spectral-resolution lidar (50–54).

Signal separation based on polarization is a technique that is often used in studying atmospheric aerosols, including clouds, by using lidar systems (55–58). Light from a polarized laser beam backscattered by aerosols will generally undergo a degree of depolarization, that is, the backscattered light will not be plane polarized. The degree of depolarization depends on a number of factors, including the anisotropy of the scattering aerosols. Depolarization of backscattered light also results from multiple scattering of photons.

Processing of the backscattered light based on range is usually performed in order to protect the detector from the intense near-field returns of higher power lidar systems. Exposing a photomultiplier tube (PMT) to a bright source such as a near-field return, even for a very short time, produces signal-induced noise (SIN) that affects the ability of the detection system to record any subsequent signal accurately (59,60). This protection is usually achieved either by a mechanical or electroptical chopper that closes the optical path to the detector during and immediately after the laser fires or by switching the detector off during this time, called gating.

A mechanical chopper used for protecting the detector is usually a metal disk that has teeth on its periphery and is rotated at high speed. The laser and chopper are synchronized so that light backscattered from the near field is blocked by the chopper teeth but light scattered from longer ranges is transmitted through the spaces between the teeth. The opening time of the chopper depends on both the diameter of the optical beam that is being chopped and the speed at which the teeth move. Generally, opening times around 20–50 ms corresponding to a lidar range of between a few and several kilometers are required. Opening times of this order can be achieved by using a beam diameter of a few millimeters and a 10-cm diameter chopper rotating at several thousand revolutions per minute (61).

Signal Detection and Recording

The signal detection and recording section of a lidar takes the light from the receiver system and produces a permanent record of the measured intensity as a function of altitude. The signal detection and recording system in the first lidar experiments was a camera and photographic film (24,25).

Today, the detection and recording of light intensity is done electronically. The detector converts the light into an electrical signal, and the recorder is an electronic device or devices, that process and record this electrical signal.

Photomultiplier tubes (PMTs) are generally used as detectors for incoherent lidar systems that use visible and UV light. PMTs convert an incident photon into an electrical current pulse (62) large enough to be detected by sensitive electronics. Other possibilities for detectors (63) for lidar systems include multianode PMTs (64), MCPs (65), avalanche photodiodes (66,67), and CCDs (68,69). Coherent detection is covered in a later section.

The output of a PMT has the form of current pulses that are produced both by photons entering the PMT and the thermal emission of electrons inside the PMT. The output due to these thermal emissions is called dark current.

The output of a PMT can be recorded electronically in two ways. In the first technique, photon counting, the pulses are individually counted; in the second technique, analog detection, the average current due to the pulses is measured and recorded. The most appropriate method for recording PMT output depends on the rate at which the PMT produces output pulses, which is proportional to the intensity of the light incident on the PMT. If the average rate at which the PMT produces output pulses is much less that the average pulse width, then individual pulses can be easily identified, and photon counting is the more appropriate recording method.

Photon Counting

Photon counting is a two-step process. First, the output of the PMT is filtered using a discriminator to remove a substantial number of the dark counts. This is possible because the average amplitude of PMT pulses produced by incident photons is higher that the average amplitude of the pulses produced by dark counts. A discriminator is essentially a high-speed comparator whose output changes state when the signal from the PMT exceeds a preset level, called the discriminator level. By setting the discriminator level somewhere between the average amplitude of the signal count and dark count levels, the discriminator can effectively filter out most of the dark counts. Details of operating a photomultiplier in this manner can be found in texts on optoelectronics (62,70,71).

The second step in photon counting involves using a multichannel counter, often called a multichannel scaler (MCS). A MCS has numerous memory locations that are accessed sequentially and for a fixed time after the MCS receives a signal indicating that a laser pulse has been fired into the atmosphere. If the output from the discriminator indicates that a count should be registered, then the MCS adds one to the number in the currently active memory location. In this way, the MCS can count scattered laser photons as a function of range. An MCS is generally configured to add together the signals detected from a number of laser pulses. The total signal recorded by the MCS, across the averaging period of interest, is then stored on a computer. All of the MCS memory locations are then reset to zero, and the counting process is restarted.

If a PMT produces two pulses that are separated by less that the width of a pulse, they are not resolved, and the output of the discriminator indicates that only one pulse was detected. This effect is called pulse pileup. As the intensity of the measured light increases, the average count rate increases, pulse pileup becomes more likely, and more counts are missed. The loss of counts due to pulse pileup can be corrected (39,72), as long as the count rate does not become excessive. In extreme cases, many pulses pileup, and the output of the PMT remains above the discriminator level, so that no pulses are counted.

Analog Detection

Analog detection is appropriate when the average count rate approaches the pulse-pair resolution of the detector system, usually of the order of 10 to 100 MHz depending on the PMT type, the speed of the discriminator, and the MCS. Analog detection uses a fast analog-to-digital converter to convert the average current from the PMT into digital form suitable for recording and manipulation on a computer (73).

Previously, we described a method for protecting a PMT from intense near-field returns using a rotating chopper. An alternative method for protecting a PMT is called blanking or gating (74–76). During gating, the PMT is effectively turned off by changing the distribution of voltages on the PMT's dynode chain. PMT gating is simpler to implement and more reliable than a mechanical chopper system because it has no moving parts. However, it can cause unpredictable results because gating can cause gain variations and a variable background that complicates the interpretation of the lidar returns.

Coherent Detection

Coherent detection is used in a class of lidar systems designed for velocity measurement. This detection technique mixes the backscattered laser light with light from a local oscillator on a photomixer (77). The output of the photomixer is a radio-frequency (RF) signal whose frequency is the difference between the frequencies of the two optical signals. Standard RF techniques are then used to measure and record this signal. The frequency of the measured RF signal can be used to determine the Doppler shift of the scattered laser light, which in turn allows calculation of the wind velocity (78–82).

Coherent lidar systems have special requirements for laser pulse length and frequency stability. The advantage of coherent detection for wind measurements is that the instrumentation is generally simpler and more robust than that required for incoherent optical interferometric detection of Doppler shifts (20).

An Example of a Lidar Detection System

Many lidar systems detect light at multiple wavelengths and/or at different polarization angles. The Purple Crow lidar (39,83) at the University of Western Ontario detects scattering at four wavelengths (Fig. 6). A Nd:YAG laser operating at the second-harmonic frequency (532 nm) provides the light source for the Rayleigh (532 nm) and the two Raman shifted channels, N_2 (607 nm) and H_2O (660 nm). The fourth channel is a sodium resonance-fluorescence channel that operates at 589 nm. Dichroic mirrors are used to separate light collected by the parabolic mirror into these four channels before the returns are filtered by narrowband interference filters and imaged onto the PMTs.

A rotating chopper is incorporated into the two high-signal-level channels, Rayleigh and sodium, to protect the PMTs from intense near-field returns. The chopper operates at a high speed, 8,400 rpm, and is comprised of a rotating disk that has two teeth on the outside edge. This chopper blocks all scatter from below 20 km and is fully open by 30 km. The signal levels in the two Raman channels are sufficiently small that the PMTs do not require protection from near-field returns.

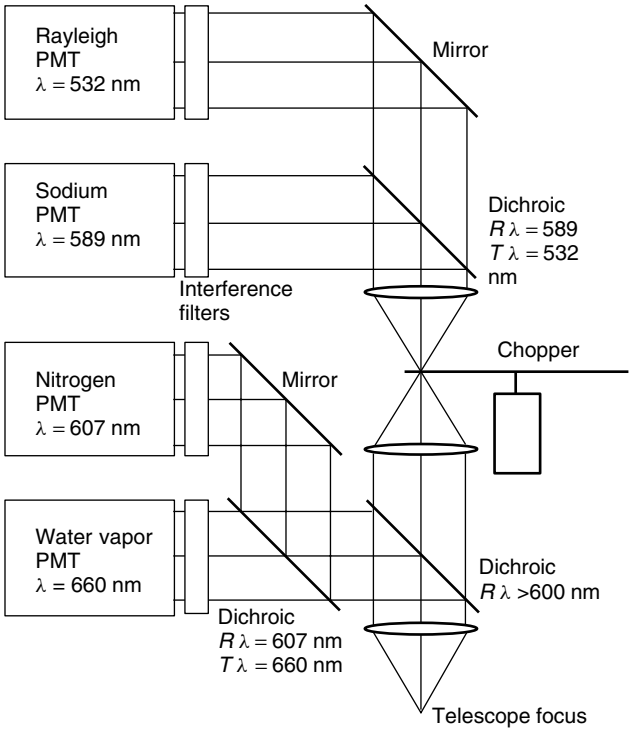


Figure 6. Schematic of the detection system of the Purple Crow lidar at the University of Western Ontario.

The two Raman channels are used for detecting H_2O and N_2 in the troposphere and stratosphere and thus allow measurement of water vapor concentration and temperature profiles. Measurements from the Rayleigh and sodium channels are combined to provide temperature profiles from 30 to 110 km.

THE LIDAR EQUATION

The lidar equation is used to determine the signal level detected by a particular lidar system. The basic lidar equation takes into account all forms of scattering and can be used to calculate the signal strength for all types of lidar, except those that employ coherent detection. In this section, we derive a simplified form of the lidar equation that is appropriate for monostatic lidar without any high-spectral resolution components. This equation is applicable to simple Rayleigh, vibrational Raman, and DIAL systems. It is not appropriate for Doppler or pure rotational Raman lidar, because it does not include the required spectral dependencies.

Let us define P as the total number of photons emitted by the laser in a single laser pulse at the laser wavelength λ_l and τ_t as the transmission coefficient of the lidar transmitter optics. Then the total number of photons transmitted into the atmosphere by a lidar system in a single laser pulse is given by

$$P\tau_t(\lambda_l). \tag{1}$$

The number of photons available to be scattered in the range interval r to $r + dr$ from the lidar is

$$P\tau_t(\lambda_l)\tau_a(r, \lambda_l)\, dr, \tag{2}$$

where $\tau_a(r, \lambda_l)$ is the optical transmission of the atmosphere at the laser wavelength, along the laser path to the range r. Note that range and altitude are equivalent only for a vertically pointing lidar.

The number of photons backscattered, per unit solid angle due to scattering of type i, from the range interval R_1 to R_2, is

$$P\tau_t(\lambda_l) \int_{R_1}^{R_2} \tau_a(r, \lambda_l)\sigma_\pi^i(\lambda_l)N^i(r)\, dr, \tag{3}$$

where $\sigma_\pi^i(\lambda_l)$ is the backscatter cross section for scattering of type i at the laser wavelength and $N^i(r)$ is the number density of scattering centers that cause scattering of type i at range r.

Range resolution is most simply and accurately achieved if the length of the laser pulse is much shorter than the length of the range bins. If this condition cannot be met, the signal can be deconvolved to obtain the required range resolution (84,85). The effectiveness of this deconvolution depends on a number of factors, including the ratio of the laser pulse length to the length of the range bins, the rate at which the signal changes over the range bins, and the signal-to-noise ratio of the measurements.

The number of photons incident on the collecting optic of the lidar due to scattering of type i is

$$P\tau_t(\lambda_l)A \int_{R_1}^{R_2} \frac{1}{r^2}\tau_a(r, \lambda_l)\tau_a(r, \lambda_s)\zeta(r)\sigma_\pi^i(\lambda_l)N^i(r)\, dr, \tag{4}$$

where A is the area of the collecting optic, λ_s is the wavelength of the scattered light, and $\zeta(r)$ is the overlap factor that takes into account the intensity distribution across the laser beam and the physical overlap of the transmitted laser beam on the FOV of the receiver optics. The term $1/r^2$ arises in Eq. (4) due to the decreasing illuminance of the telescope by the scattered light, as the range increases.

For photon counting, the number of photons detected as pulses at the photomultiplier output per laser pulse is

$$P\tau_t(\lambda_l)A\tau_r(\lambda_s)Q(\lambda_s) \int_{R_1}^{R_2} \frac{1}{r^2}\tau_a(r, \lambda_l)\tau_a(r, \lambda_s)\zeta(r)\sigma_\pi^i(\lambda_l)N^i(r)\, dr, \tag{5}$$

where $\tau_r(\lambda_s)$ is the transmission coefficient of the reception optics at λ_s and $Q(\lambda_s)$ is the quantum efficiency of the photomultiplier at wavelength λ_s.

For analog detection, the current recorded can be determined by replacing the quantum efficiency of the photomultiplier $Q(\lambda_s)$ by the gain $G(\lambda_s)$ of the photomultiplier combined with the gain of any amplifiers used.

In many cases, approximations allow simplification of Eq. (5). For example, if none of the range-dependent terms, $\tau_a(r, \lambda_l)$, $\tau_a(r, \lambda_s)$, $N^i(r)$, and $\zeta(r)$, varies significantly throughout individual range bins, then the range integral may be removed, and Eq. 5 becomes

$$P\tau_t(\lambda_l)A\tau_r(\lambda_s)Q(\lambda_s)\tau_a(R, \lambda_l)\tau_a(R, \lambda_s)\frac{1}{R^2}\zeta(R)\sigma_\pi^i(\lambda_l)N^i(R)\delta R \tag{6}$$

where R is the range of the center of the scattering volume and $\delta R = R_2 - R_1$, is the length of the range bin.

This form of the lidar equation can be used to calculate the signal strength for Rayleigh, vibrational Raman lidar, and DIAL as long as the system does not incorporate any filter whose spectral width is of the same order or smaller than the width of the laser output or the Doppler broadening function. For high-resolution spectral lidar, where a narrow-spectral-width filter or tunable laser is used, the variations in the individual terms of Eq. (6) with wavelength need to be considered. To calculate the measurement precision of a lidar that measures the Doppler shift and broadening of the laser line for wind and temperature determination, computer simulation of the instrument may be necessary.

LIGHT SCATTERING IN THE ATMOSPHERE AND ITS APPLICATION TO LIDAR

The effect of light scattering in the Earth's atmosphere, such as blue skies, red sunsets, and black, grey, and white clouds, is easily observed and reasonably well understood (86–89). Light propagating through the atmosphere is scattered and absorbed by the molecules and aerosols, including clouds that form the atmosphere. Molecular scattering takes place via a number of different processes and may be either elastic, where there is no exchange of energy with the molecule, or inelastic, where an exchange of energy occurs with the molecule. It is possible to calculate, by at least a reasonable degree of accuracy, the parameters that describe these molecular scattering processes.

The theory of light scattering and absorption by spherical aerosols, usually called Mie (90) theory, is well understood, though the application of Mie theory to lidar can be difficult in practice. This difficulty arises due to computational limits encountered when trying to solve atmospheric scattering problems where the variations in size, shape, and refractive index of the aerosol particles can be enormous (91–97). However, because aerosol lidars can measure average properties of aerosols directly, they play an important role in advancing our understanding of the effect of aerosols on visibility (98–101) as well as on climate (102,103).

Molecules scatter light by a variety of processes; there is, however, an even greater variety of terms used to describe these processes. In addition, researchers in different fields have applied the same terms to different processes. Perhaps the most confused term is Rayleigh scattering, which has been used to identify at least three different spectral regions of light scattered by molecules (104–106).

RAYLEIGH SCATTER AND LIDAR

Rayleigh theory describes the scattering of light by particles that are small compared to the wavelength of the incident radiation. This theory was developed by Lord Rayleigh (107,108) to explain the color, intensity distribution, and polarization of the sky in terms of scattering by atmospheric molecules.

In his original work on light scattering, Rayleigh used simple dimensional arguments to arrive at his well-known equation. In later years, Rayleigh (109,110) and others (22,87,111,112) replaced these dimensional arguments with a more rigorous mathematical derivation of the theory. Considering a dielectric sphere of radius r in a parallel beam of linearly polarized electromagnetic radiation, one can derive the scattering equation. The incident radiation causes the sphere to become an oscillating dipole that generates its own electromagnetic field, that is, the scattered radiation. For this derivation to be valid, it is necessary for the incident field to be almost uniform across the volume of the scattering center. This assumption leads to the restriction of Rayleigh theory to scattering by particles that are small compared to the wavelength of the incident radiation. It can be shown (113) that when $r < 0.03\lambda$, the differences between results obtained with Rayleigh theory and the more general Mie (90) theory are less than 1%.

Rayleigh theory gives the following equation for the scattered intensity from a linearly polarized beam by a single molecule:

$$I_{\mathrm{m}}(\phi) = E_0^2 \frac{9\pi^2 \varepsilon_0 c}{2N^2\lambda^4} \left(\frac{n^2-1}{n^2+2}\right)^2 \sin^2\phi, \tag{7}$$

where r is the radius of the sphere, n is the index of refractive of the sphere relative to that of the medium, that is, $n = n_{\mathrm{molecule}}/n_{\mathrm{medium}}$, N is the number density of the scattering centers, ϕ is the angle between the dipole axis and the scattering direction, and E_0 is the maximum value of the electrical field strength of the incident wave (22,87). From Eq. (7), we see that the intensity of the scattered light varies as λ^{-4}. However, because the refractive index may also have a small wavelength dependence, the scattered intensity is in fact not exactly proportional to λ^{-4}. Middleton (114) gives a value of $\lambda^{-4.08}$ for wavelengths in the visible.

A useful quantity in discussion is the differential-scattering cross section (22), which is also called the angular scattering cross section (87). The differential-scattering cross section is the fraction of the power of the incident radiation that is scattered, per unit solid angle, in the direction of interest. The differential-scattering cross section is defined by

$$\frac{d\sigma(\phi)}{d\Omega} I_0 = I(\phi), \tag{8}$$

where $I_0 = 1/2c\varepsilon_0 E_0^2$ is the irradiance of the incident beam.

By substituting Eq. (7) in (8), it can be seen that the differential scattering cross section for an individual molecule illuminated by plane polarized light, is

$$\frac{d\sigma_{\mathrm{m}}(\phi)}{d\Omega} = \frac{9\pi^2}{N^2\lambda^4} \left(\frac{n^2-1}{n^2+2}\right)^2 \sin^2\phi. \tag{9}$$

If we assume that $n \approx 1$, then Eq. (9) can be approximated as

$$\frac{d\sigma_{\mathrm{m}}(\phi)}{d\Omega} = \frac{\pi^2(n^2-1)^2}{N^2\lambda^4} \sin^2\phi. \tag{10}$$

For a gas, the term $(n^2 - 1)$ is approximately proportional to the number density N (115), so Eq. (10) has only a very slight dependence on N. For air, the ratio $(n^2 - 1)/N$ varies less than 0.05% in the range of N between 0 and 65 km in altitude.

When Rayleigh theory is extended to include unpolarized light, the angle ϕ no longer has any meaning because the dipole axis may lie along any line in the plane perpendicular to the direction of propagation. The only directions that can be uniquely defined are the direction of propagation of the incident beam and the direction in which the scattered radiation is detected; we define θ as the angle between these two directions. The differential-scattering cross section for an individual molecule that is illuminated by a parallel beam of unpolarized light is

$$\frac{d\sigma_{\mathrm{m}}(\theta)}{d\Omega} = \frac{\pi^2(n^2-1)^2}{2N^2\lambda^4}(1 + cos^2\theta). \tag{11}$$

Figure 7 shows the intensity distribution for Rayleigh scattered light from an unpolarized beam. The distribution has peaks in the forward and backward directions, and the light scattered at right angles to the incident beam is plane polarized. Because of the anisotropy of molecules, which moves the molecules dipole moment slightly out of alignment with the incident field, scattering by molecules causes some depolarization of the scattered light. This results in some light whose a polarization is parallel to the incident beam being detected at a scattering angle of 90°.

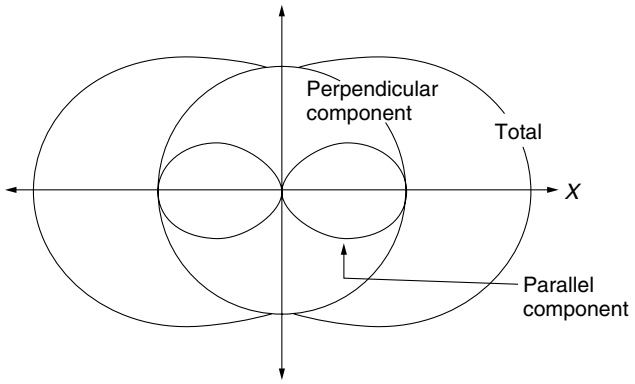


Figure 7. Intensity distribution pattern for Rayleigh scatter from an unpolarized beam traveling in the x direction. The perpendicular component refers to scattering of radiation whose electric vector is perpendicular to the plane formed by the direction of propagation of the incident beam and the direction of observation.

The depolarization ratio δ_n^t is defined as

$$\delta_n^t = \frac{I_\parallel}{I_\perp}, \tag{12}$$

where the parallel and perpendicular directions are taken with respect to the direction of the incident beam. The subscript n denotes natural (unpolarized) incident light and the superscript t denotes total molecular scattering. The depolarization is sometimes defined in terms of polarized incident light and/or for different spectral components of molecular scattering. There is much confusion about which is the correct depolarization to use under different circumstances, a fact evident in the literature. The reader should take great care to understand the terminology used by each author.

Young (104) gives a brief survey of depolarization measurements for dry air and concludes that the effective value of δ_n^t is 0.0279. He also gives a correction factor for the Rayleigh differential-scattering cross section, which, when applied to Eq. (11) gives

$$\frac{d\sigma_m(\theta)}{d\Omega} = \frac{\pi^2(n^2-1)^2}{2N^2\lambda^4} \frac{1+\delta_n^t+(1-\delta_n^t)\cos^2\theta}{1-\left(\frac{7}{6}\right)\delta_n^t} \tag{13}$$

Most lidar applications work with direct backscatter, i.e. $\theta = \pi$, and the differential-backscatter cross section per molecule for scattering from an unpolarized beam is

$$\frac{d\sigma_m(\theta=\pi)}{d\Omega} = \frac{\pi^2(n^2-1)^2}{2N^2\lambda^4}\left(\frac{12}{6-7\delta_n^t}\right) \tag{14}$$

The correction factor for backscatter is independent of the polarization state of the incident beam (111). This means that the correction factor and thus, the backscatter cross section per molecule are independent of the polarization characteristics of the laser used in a backscatter lidar.

The Rayleigh molecular-backscatter cross section for an altitude less than 90 km and without the correction factor is given by Kent and Wright (116) as $4.60 \times 10^{-57}/\lambda^4\ \mathrm{m^2\,sr^{-1}}$. When the correction factor is applied, with $\delta_n^t = 0.0279$, this result becomes

$$\frac{d\sigma_m(\theta=\pi)}{d\Omega} = \frac{4.75\times10^{-57}}{\lambda^4}\ \mathrm{m^2\,sr^{-1}} \tag{15}$$

Collis et al. (117) gives a value of the constant in Eq. (15) as $4.99 \times 10^{-57}\ \mathrm{m^6\,sr^{-1}}$.

Fiocco (118) writes Eq. (15) in the form

$$\frac{d\sigma_m(\theta=\pi)}{d\Omega} = \frac{4.73\times10^{-57}}{\lambda^{4.09}}\ \mathrm{m^2\,sr^{-1}} \tag{16}$$

Here, the wavelength exponent takes into account dispersion in air. Equations (15) and (16) are applicable to the atmosphere at altitudes less than 90 km. Above this altitude, the concentration of atomic oxygen becomes significant and changes the composition and thus, the refractive index. Equations (15) and (16), used in conjunction with the lidar equation [Eq. (6)] can be used to determine the backscatter intensity of a particular Rayleigh lidar.

Rayleigh Lidar

Rayleigh lidar is the name given to the class of lidar systems that measure the intensity of the Rayleigh backscatter from an altitude of about 30 km up to around 100 km. The measured backscatter intensity can be used to determine a relative density profile; this profile is used to determine an absolute temperature profile. Rayleigh scattering is by far the dominant scattering mechanism for light above an altitude of about 30 km, except in the rare case where noctilucent clouds exist. At altitudes below about 25–30 km, light is elastically scattered by aerosols in addition to molecules. Only by using high-spectral-resolution techniques can the scattering from these two sources be separated (119). Thus, most Rayleigh lidar systems cannot be used to determine temperatures below the top of the stratospheric aerosol layer. The maximum altitude of the stratospheric aerosol layer varies with the season and is particularly perturbed after major volcanic activity.

Above about 90 km, changes in composition, due mainly to the increasing concentration of atomic oxygen, cause the Rayleigh backscatter cross-section and the mean molecular mass of air to change with altitude. This leads to errors in the temperatures derived by using the Rayleigh technique that range from a fraction of a degree at 90 km to a few degrees at 110 km. For current Rayleigh systems, the magnitude of this error is significantly smaller than the uncertainties from other sources, such as the photocount statistics, in this altitude range. Low photocount rates give rise to large statistical uncertainties in the derived temperatures at the very top of Rayleigh lidar temperature profiles (Fig. 8a). Additional uncertainties in the temperature retrieval algorithm, due to the estimate of the pressure at the top of the density profile which is required to initiate temperature integration (120), can be significant and are difficult to quantify.

The operating principle of a Rayleigh lidar system is simple. A pulse of laser light is fired up into the atmosphere, and any photons that are backscattered and collected by the receiving system are counted as a function of range. The lidar equation [Eq. (6)] can be directly applied to a Rayleigh lidar system to calculate the expected signal strength. This equation can be expressed in the form

$$\text{Signal strength} = K\left(\frac{1}{R^2}\right)N_a\delta R \tag{17}$$

where K is the product of all of the terms that can be considered constants between 30 and 100 km in Eq. (6) and N_a is the number density of air. This result assumes that there is insignificant attenuation of the laser beam as it propagates from 30 to 100 km, that is, the atmospheric transmission $\tau_a(r, \lambda_l)$ is a constant for $30 < r < 100$ km. If there are no aerosols in this region of the atmosphere and the laser wavelength is far from the absorption lines of any molecules, then the only attenuation of the laser beam is due to Rayleigh scatter and possibly

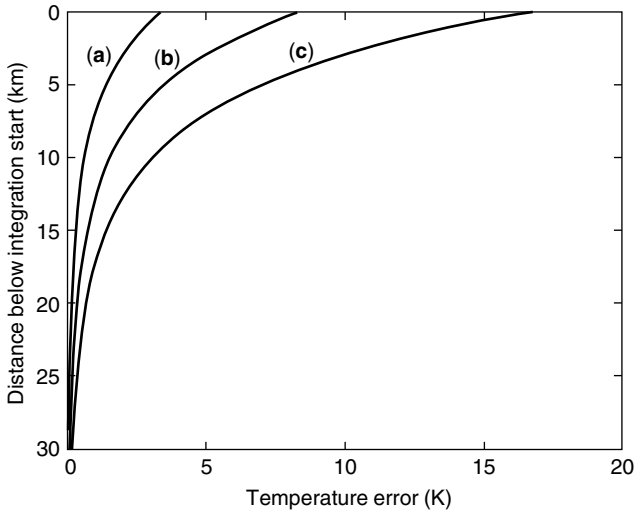


Figure 8. The propagation of the error in the calculated temperature caused by a (**a**) 2%, (**b**) 5% and (**c**) 10% error in the initial estimate of the pressure.

ozone absorption. Using Rayleigh theory, it can be shown that the transmission of the atmosphere from 30 to 100 km is greater than 99.99% in the visible region of the spectrum.

Equation (17) shows that after a correction for range R, the measured Rayleigh lidar signal between 30 and 100 km is proportional to the atmospheric density. K cannot be determined due to the uncertainties in atmospheric transmission and instrumental parameters [see Eq. (6)]. Hence, Rayleigh lidar can typically determine only relative density profiles. A measured relative density profile can be scaled to a coincident radiosonde measurement or model density profile, either at a single altitude or across an extended altitude range.

This relative density profile can be used to determine an absolute temperature profile by assuming that the atmosphere is in hydrostatic equilibrium and applying the ideal gas law. Details of the calculation and an error analysis for this technique can be found in both Chanin and Hauchecorne (120) and Shibata (121). The assumption of hydrostatic equilibrium, the balance of the upward force of pressure and the downward force of gravity, can be violated at times in the middle atmosphere due to instability generated by atmospheric waves, particularly gravity waves (122,123). However, sufficient averaging in space (e.g., 1 to 3 km) and in time (e.g., hours) minimizes such effects.

Calculating an absolute temperature profile begins by calculating a pressure profile. The first step in this process is to determine the pressure at the highest altitude range-bin of the measured relative density profile. Typically, this pressure is obtained from a model atmosphere. Then, using the density in the top range-bin, the pressure at the bottom of this bin is determined using hydrostatic equilibrium. This integration is repeated for the second to top density range-bin and so on down to the bottom of the density profile. Because atmospheric density increases as altitude decreases, the choice of pressure at the top range-bin becomes less significant in the calculated pressures, as the integration proceeds. A pressure profile calculated in this way is a relative profile because the density profile from which it was determined is a relative profile. However, the ratio of the relative densities to the actual atmospheric densities will be exactly the same as the ratio of the relative pressures to the actual atmospheric pressures:

$$N_{\mathrm{rel}} = K' N_{\mathrm{act}}$$

and

$$P_{\mathrm{rel}} = K' P_{\mathrm{act}}, \tag{18}$$

where N_{rel} is the relative density and N_{act} is the actual atmospheric density, similarly for the pressure P, and K' is the unknown proportionality constant. The ideal gas law can then be applied to the relative density and pressure profiles to yield a temperature profile. Because the relative density and relative pressure profiles have the same proportionality constant [see Eq. (18)], the constants cancel, and the calculated temperature is absolute.

The top of the temperature profile calculated in this scheme is influenced by the choice of initial pressure. Figure 8 shows the temperature error as a function of altitude for a range of pressures used to initiate the pressure integration algorithm. Users of this technique are well advised to ignore temperatures from at least the uppermost 8 km of the retrieval because the uncertainties introduced by the seed pressure estimate are not easily

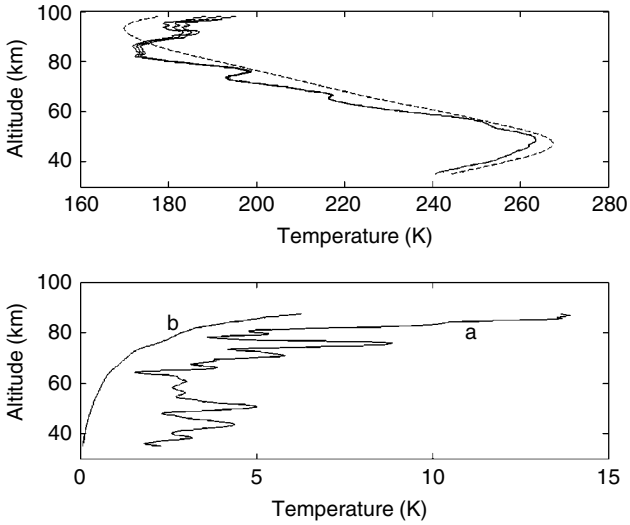


Figure 9. Top panel shows the average temperature (middle of the three solid lines) for the night of 13 August 2000 as measured by the PCL. The two outer solid lines represent the uncertainty in the temperature. Measurements are summed across 288 m in altitude and 8 hours in time. The temperature integration algorithm was initiated at 107.9 km; the top 10 km of the profile has been removed. The dashed line is the temperature from the Fleming model (289) for the appropriate location and date. Bottom panel shows (**a**) the rms deviation from the mean temperature profile for temperatures calculated every 15 minutes at the same vertical resolution as before. (**b**) is the average statistical uncertainty in the individual temperature profiles used in the calculation of the rms and is based on the photon counting statistics.

quantified, unless an independent determination of the temperature is available.

The power–aperture product is the typical measure of a lidar system's effectiveness. The power–aperture product is the mean laser power (watts) multiplied by the collecting area of the receiver system (m^2). This result is, however, a crude metric because it ignores both the variations in Rayleigh-scatter cross section and atmospheric transmission with transmitter frequency, as well as the efficiency of the system.

The choice of a laser for use in Rayleigh lidar depends on a number of factors, including cost and ease of use. The best wavelengths for a Rayleigh lidar are in the blue–green region of the spectrum. At longer wavelengths, for example, the infrared, the scattering cross section is smaller, and thus, the return signal is reduced. At shorter wavelengths, for example, the ultraviolet, the scattering cross section is higher, but the atmospheric transmission is lower, leading to an overall reduction in signal strength. Most dedicated Rayleigh lidars use frequency-doubled Nd:YAG lasers that operate at 532 nm (green light). Other advantages of this type of laser are that it is a well-developed technology that provides a reliable, "turnkey," light source that can produce pulses of short duration with typical average powers of 10 to 50 W. Some Rayleigh lidar systems use XeF excimer lasers that operate at about 352 nm. These systems enjoy the higher power available from these lasers, as well as a Rayleigh-scatter cross section larger than for Nd:YAG systems, but the atmospheric transmission is lower at these wavelengths. In addition, excimer lasers are generally considered more difficult and expensive to operate than Nd:YAG lasers.

An example of a temperature profile from The University of Western Ontario's Purple Crow lidar Rayleigh (40) system is shown in Fig. 9. The top panel of the figure shows the average temperature during the night's observations, including statistical uncertainties due to photon counting. The bottom panel shows the rms deviation of the temperatures calculated at 15-minute intervals. The rms deviations are a measure of the geophysical variations in temperature during the measurement period. Also included on the bottom panel is the average statistical uncertainty due to photon counting in the individual 15-minute profiles.

Rayleigh lidar systems have been operated at a few stations for several years building up climatological records of middle atmosphere temperature (60,124,125). The lidar group at the Service d'Aeronomie du CNRS, France has operated a Rayleigh lidar at the Observatory of Haute-Provence since 1979 (120,125–128). The data set collected by this group provides an excellent climatological record of temperatures in the middle and upper stratosphere and in the lower mesosphere.

Lidar systems designed primarily for sodium and ozone measurements have also been used as Rayleigh lidar systems for determining stratospheric and mesospheric temperatures (129–131). Rayleigh-scatter lidar measurements can be used in conjunction with independent temperature determinations to calculate molecular nitrogen and molecular oxygen mixing ratios in the mesopause region of the atmosphere (132).

Rayleigh lidar systems cannot operate when clouds obscure the middle atmosphere from their view. Most Rayleigh systems can operate only at nighttime due to the presence of scattered solar photons during the day. However, the addition of a narrow band-pass filter in the receiver optics allows daytime measurements (35,133).

Doppler Effects

Both random thermal motions and bulk-mean flow (e.g., wind) contribute to the motion of air molecules. When light is scattered by molecules, it generally undergoes a change in frequency due to the Doppler effect that is proportional to the molecules line of sight velocity. If we consider the backscattered light and the component of velocity of the scattering center in the direction of the scatter, then the Doppler shift, that is, the change in frequency $\Delta\nu$ of the laser light is given by (134)

$$\Delta\nu = \nu' - \nu \approx 2\nu\frac{v}{c} \tag{19}$$

where ν is the frequency of the incident photon, ν' is the frequency of the scattered photon, and v is the component of the velocity of the scattering center in the direction of scatter (e.g., backscatter).

The random thermal motions of the air molecules spectrally broaden the backscattered light, and radial wind causes an overall spectral shift. The velocity distribution function due to thermal motion of gas molecules in thermal equilibrium is given by Maxwell's distribution. For a single direction component x, the probability that a molecule has velocity v_x is (135)

$$P(v_x)dv_x = \left(\frac{M}{2\pi kT}\right)^{1/2} \exp\left(-\frac{Mv_x^2}{2\,kT}\right) dv_x \tag{20}$$

where M is molecular weight, k is Boltzmann's constant, T is temperature, and v_x is the component of velocity in the x direction.

Using Eqs. (19) and (20), it can be shown that when monochromatic light is backscattered by a gas, the frequency distribution of the light is given by

$$P(\nu') = \frac{1}{2\pi^{1/2}\sigma} \exp\left[-\frac{1}{2}\left(\frac{\nu' - \nu}{\sigma}\right)^2\right], \tag{21}$$

where

$$\sigma = \frac{\nu}{c}\left(\frac{2kT}{M}\right)^{1/2}. \tag{22}$$

The resulting equation for $P(\nu')$ is a Gaussian distribution whose full width at half maximum is equal to $2\sigma\sqrt{2\ln 2}$.

Equations (21) and (22) are strictly true only if all the atoms (molecules) of the gas have the same atomic (molecular) weight. However, air contains a number of molecular and atomic species, and therefore the frequency distribution function for Rayleigh backscattered light $P_a(\nu')$ is the weighted sum of Gaussian functions for each constituent. The major constituents of air, N_2 and O_2, have similar molecular masses which allows the function $P_a(\nu')$ to be fairly well approximated by a single Gaussian

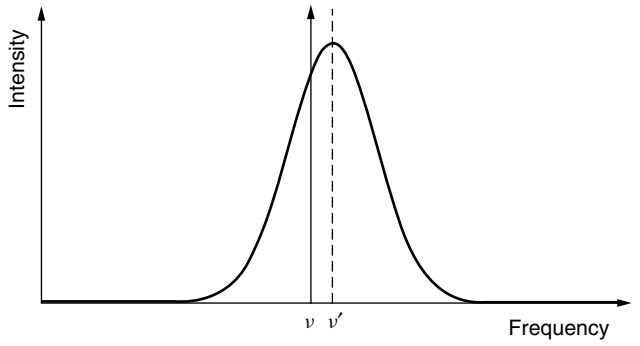


Figure 10. The frequency distribution function for Rayleigh backscattering from a clean dry atmosphere (i.e., no water vapor or aerosols), for monochromatic incident radiation of frequency ν. The broadening is due to random thermal motions and the shift is due to wind.

calculated for a gas whose a molecular mass is equal to the mean molecular mass of air.

Wind, the bulk motion of the air, causes the distribution function $P_a(\nu')$ to shift frequency while maintaining its shape. The frequency shift can be calculated directly from Eq. (19), which shows that the shift is directly proportional to the component of the wind velocity in the direction of scattering, the radial wind velocity. Figure 10 shows how the spectrum of a narrow bandwidth laser is changed due to scattering by molecules in the atmosphere.

In principle, it is possible to determine both the radial wind velocity and temperature by measuring the spectral shape of the light backscattered from air molecules in the middle atmosphere. However, using this Doppler technique, the signal-to-noise ratio requirements for temperature measurement are much higher than that for measuring winds (136), and so in practice, Rayleigh–Doppler temperature measurements are quite difficult. The advantage of this method of temperature determination is that the true kinetic temperature of the atmosphere is obtained without the need for the assumptions required by the Rayleigh technique. The group at the Observatory Haute-Provence (54,137) has demonstrated the Doppler technique for measuring middle atmosphere winds. They used a Fabry–Perot interferometer as a narrowband filter to measure the intensity of the lidar returns in a pair of wavelength ranges centered on the laser wavelength (54). Tepley et al. used a scanning interferometer to make similar measurements (136).

AEROSOL SCATTERING AND LIDAR

The theory of scattering that was developed by Mie (90) in the early 1900's is a general solution that covers the scattering of electromagnetic radiation by a homogeneous sphere for all wavelengths of radiation and spheres of all sizes and refractive indexes. A parameter that is basic to the Mie theory is the size parameter α. This parameter is a measure of the relative size of the scattering particle to the wavelength of the radiation:

$$\alpha = \frac{2\pi a}{\lambda} \tag{23}$$

where a is the radius of the scattering particle and λ is the wavelength of the incident radiation. When the particle size is small compared to the wavelength of the incident radiation (i.e., α is small), Mie theory reduces to Rayleigh theory.

Mie theory is general enough to cover the range of α's for which Rayleigh and geometrical optics also apply, but it is mathematically more complex than Rayleigh theory and geometrical optics. This complexity has led to the common use of Mie scattering to imply scattering from particles larger than those to which Rayleigh theory applies and smaller than those to which geometrical optics applies. Mie theory solves Maxwell's equations for the boundary conditions imposed by a homogeneous sphere whose refractive index is different from that of the surrounding medium. Since Mie first published the solution to this problem, others have extended the calculations to include different shapes (e.g., infinite cylinders and paraboloids) and have provided methods for finding solutions for irregular shapes and nonhomogenous particles (112,138–140).

The atmosphere contains particles that have an infinite variety of shapes, sizes and refractive indexes. The measurement of the properties of atmospheric aerosols is also complicated by the composition and size of these particles (87,141–143). Evaporation, condensation, coagulation, absorption, desorption, and chemical reactions change the atmospheric aerosol composition on short timescales. Care must be taken with direct sampling methods that the sampling process allows correct interpretation of the properties of the aerosols collected.

Aerosol concentrations in the atmosphere vary widely with altitude, time, and location. The vertical structure of aerosol concentration profiles is complex and ever changing (144–148). There is a layer of aerosols in the atmosphere from about 15 to 23 km that is known as the stratospheric aerosol layer or the Junge (149) layer. The Junge is primarily volcanic in origin. Lidar measurements have shown that the altitude range and density of the aerosols in this layer vary widely depending on recent volcanic activity (150–154).

Extinction cross sections given by the Mie theory for size parameters corresponding to atmospheric aerosols and visible light are generally larger than extinction cross sections due to molecular scattering (87). In the atmospheric boundary layer, where the aerosol concentrations are high, the extinction of a beam of visible light is much greater than that due solely to Rayleigh scattering. Tropospheric aerosols can be a mixture of natural and anthropogenic aerosols. The effects of clouds are difficult to quantify due to the great variability they exhibit in their optical properties and in their distribution in time and space.

Atmospheric aerosols, including clouds, play an important role in the earth's radiation budget. A full understanding of the role of aerosols is important for improving weather forecasting and understanding climate change. Aerosols scatter and absorb both incoming solar radiation and outgoing terrestrial radiation. The amount of radiation that is scattered and the directions of scatter, as well as the amount or radiation absorbed, varies with aerosol

composition, size, and shape. Thus, the physical properties of aerosols determine whether they contribute net heating or cooling to the Earth's climate. Lidar provides a method of directly measuring the optical properties of atmospheric aerosol distributions and is playing an important role in current work to better quantify the atmospheric radiation budget (148,155–160).

Aerosol Lidar

Since the early 1960s, a large number of lidar systems have been built that are designed to study aerosols, including clouds, in the troposphere and lower stratosphere (161,162). Instruments using multiple wavelength transmitters and receivers (55,145,154,163–168) and polarization techniques (55,56,58,169–173) have been used to help quantify aerosol properties. A review of aerosol lidar studies is given by Reagan et al. (174). Lidars have been used to study polar stratospheric clouds (PSCs) (175–181) to help understand the role they play in ozone depletion (182–184).

In September 1994, NASA flew a space shuttle mission, STS-64, which included the LITE experiment (185–187). LITE was a technology development and validation exercise for future space lidar systems. The scientific potential of LITE was recognized early in its development, and a science steering committee was established to ensure that the scientific potential of the experiment was exploited. LITE used a Nd:YAG operating simultaneously at three frequencies, the fundamental 1,064 nm, the second harmonic 532 nm, and the third harmonic 355 nm. It also incorporated a system for automatically aligning the laser beam into the FOV of the detector system. The science objectives of LITE were to study the following atmospheric properties:

1. tropospheric aerosols, including scattering ratio and its wavelength dependence, planetary boundary layer height, structure and optical depth;
2. stratospheric aerosols, including scattering ratio and its wavelength dependence, averaged integrated backscatter, as well as stratospheric density and temperature;
3. the vertical distribution, multi layer structure, fractional cover, and optical depth of clouds;
4. the radiation budget via measurements of surface reflectance and albedo as a function of incidence angle.

Figure 11 shows a sample of the LITE measurements. This figure clearly shows regions of enhanced scatter from cloud and dust from the Saharan Desert in Northwest Africa. A worldwide correlative measurement program was undertaken for validation and intercomparison with LITE measurements. This correlative measurement program included more than 60 ground-based and several aircraft-based lidar systems (188–190).

Atmospheric aerosols have the same average velocity as atmospheric molecules; thus, the average Doppler shift of their distributions is the same, see section Doppler Effects earlier. The spectral broadening of the light backscattered from aerosols is much narrower than that backscattered from molecules because the mass of aerosols is much greater than that of air molecules. Light backscattered from aerosols can be separated from that backscattered from molecules using this difference in Doppler width (119,191); however, spectral separation is not necessary if only wind is to be measured because the average Doppler shift is the same for both molecular and aerosol scattering. Wind lidar using incoherent detection has been used in the troposphere (51,137); however, coherent detection techniques are more commonly used.

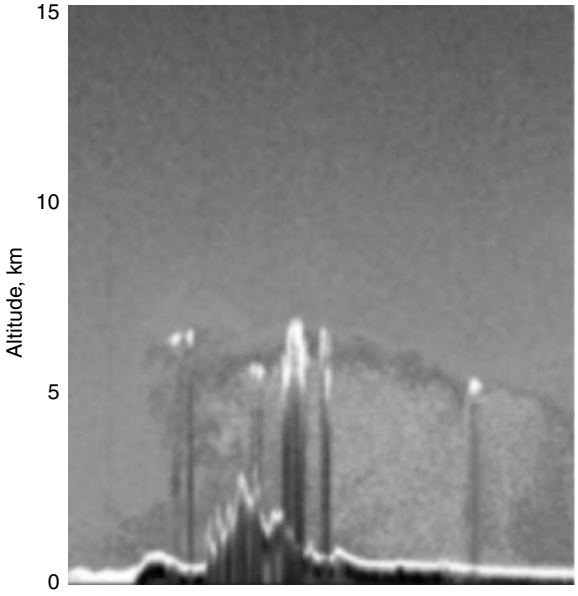


Figure 11. LITE Observations of Saharan dust, 12 September, 1994. Elevated dust layers exceeding 5 km above the Saharan Desert in Northwest Africa were observed by the Lidar In-Space Technology Experiment (LITE). The intensity plot for the 532-nm wavelength shows an aerosol layer associated with wind-blown dust from the Saharan Desert. This image is composed of individual lidar profiles sampled at 10 Hz and extends 1,000 km along the Space Shuttle Discovery orbit track during nighttime conditions. Weaker signals due to molecular backscatter are in blue, moderate backscatter signals from the dust layer are in yellow and red, and the strongest backscatter signals from clouds and the surface are in white. Opaque clouds, shown in white, prevent LITE from making observations at lower altitudes and create a shadowing effect beneath the cloud layer. The Atlas Mountain range is seen near 31°N, 6°W (David M. Winker, NASA Langley Research Center, and Kathleen A. Powell, SAIC). See color insert.

Coherent Doppler Lidar

Because of stronger the signal levels in the lower atmosphere, the measurement of the Doppler shift via coherent detection techniques becomes viable. Coherent Doppler lidar is used extensively in wind field mapping from the ground (192,193), from the air (194–196), and has been suggested as a possible method for global wind measurement from space platforms (194,197).

Differential Absorption Lidar (Dial)

In 1964, Schotland (198) suggested using a lidar technique now known as differential absorption lidar (DIAL). DIAL is useful for measuring the concentration of trace species in the atmosphere. The method relies on the sharp variation in optical transmission near an absorption line of the species to be detected. A DIAL transmits two closely spaced wavelengths. One of these wavelengths coincides with an absorption line of the constituent of interest, and the other is in the wing of this absorption line. During the transmission of these two wavelengths through the atmosphere, the emission that is tuned to the absorption line is attenuated more than the emission in the wing of the absorption line. The intensity of the two wavelengths that are backscattered to the DIAL instrument can then be used to determine the optical attenuation due to the species and thus, the concentration of the species. The first use of a DIAL system was for measuring atmospheric water vapor concentration (199). The DIAL technique has been extensively used for pollution monitoring (200–206). This technique is also used very successfully in the lower atmosphere for high spatiotemporal measurements of species such as NO (207), H_2O (208–210), O_3 (211–213), SO_2 (214,215), and CH_4 (216–218). Atmospheric temperature measurement is possible by the DIAL technique if the absorption line selected is temperature-dependent (219–221).

Use of the DIAL technique in the middle atmosphere has been restricted mainly to measuring ozone profiles (211,222–227). DIAL ozone measurements have extended as high as 50 km with integration times of at least a few hours required. These same lidar systems can obtain profiles up to 20 km in approximately 15 min due to the much higher ozone densities and available scatterers at the lower levels. Typically, a stratospheric ozone DIAL uses a XeCl laser that operates at 308 nm for the "on-line" or absorbed wavelength and a frequency-tripled YAG at 355 nm for the "off-line" or reference wavelength. The spectral separation between the wavelengths means that when large stratospheric aerosol loading events occurs (such as after a large volcanic eruption), the measurements become difficult to interpret due to the optical effects of the aerosols. These shortcomings have been addressed by recording the Raman-shifted backscatter from N_2 at both of the transmitted wavelengths (228).

The DIAL technique has also been used with hard targets (229,230) and is called differential optical absorption spectroscopy (DOAS). DOAS measurements are an average across the entire path from the instrument to the target, so a DOAS system is not strictly a lidar because it does not perform any ranging. DOAS has been used to monitor large areas from aircraft using the ground as the target or reflector and has been used for monitoring chemical (6–8) and biological (9–12) weapons agents.

RAMAN LIDAR

When monochromatic light, or light of sufficiently narrow spectral width, is scattered by a molecular gas or liquid, the spectrum of the scattered light, it can be observed, contains lines at wavelengths different from those of the incident radiation (231). Raman first observed this effect (232), that is due to the interaction of radiation with the quantized vibrational and rotational energy levels of the molecule. Raman scattering involves a transfer of energy between scattered light and a molecule and is therefore, an inelastic process. The cross sections due to Raman scattering are included in the Rayleigh scattering theory (106), although Raman spectroscopists use the term Rayleigh line to indicate only the unshifted central component of the scattered light.

Each type of molecule has unique vibrational and rotational quantum energy levels and therefore, Raman scattering from each type of molecule has a unique spectral signature. This allows the identification of molecules by their scattered light spectra. Scattered radiation that loses energy during interaction with a molecule, and so decreases in frequency, is said to have a Stokes shift, whereas radiation that gains energy and increases in frequency is said to have an anti-Stokes shift. In general, Stokes radiation is more intense than anti-Stokes because the Stokes can always occur, subject to selection rules, whereas anti-Stokes also requires that the molecule is initially in an excited state.

The quantum numbers v and J describe the vibrational and rotational states of a molecule, respectively. The Q-branch, $\Delta J = 0$, contains a number of degenerate lines leading to higher intensity for light scattered in this branch. The $\Delta v = +1$ frequency shifts and backscatter cross sections for a number of atmospheric molecules are given in Fig. 12. Measures (22) gives a comprehensive list of atmospheric molecules.

The pure rotational Raman spectrum (PRRS), which occurs when there is no vibrational transition, that is, $\Delta v = 0$, is more difficult to measure because the spectral shift of the lines is quite small. This small shift leads to technical difficulties in blocking the nearby elastic scatter from entering the detector. The PRRS of an N_2 molecule is shown in Fig. 13. The intensities of the individual lines and thus the shape of the envelope of the lines are temperature-dependent.

The term Raman lidar is generally used to refer to a lidar system that uses the Raman-shifted component where $\Delta v = \pm 1$, that is, a transition that involves a change in vibrational energy level. In practice the $\Delta v = +1$ transition is commonly used because it has higher intensity. The spectral selection of the $\Delta v = +1$ line in the receiver system of a lidar can be achieved by using a high-quality narrowband interference filter. It is necessary to ensure that blocking of the filter at the laser wavelength is sufficiently high that the detected elastic backscatter from molecules and aerosols is insignificant compared to Raman scattering. Generally, special order filters are required to meet this specification.

In the mid-1960s, Cooney (233) and Leonard (234) demonstrated the measurement of the Raman-shifted component of N_2 in the troposphere by lidar. The Raman lidar technique has been used most often for measuring atmospheric water vapor (34,235–240). Clouds (241–243) and aerosols (148,156,244,245) have also been studied by this technique. The use of Raman lidar is restricted to the more abundant species in the atmosphere due to the small backscatter cross section involved. The measurement of

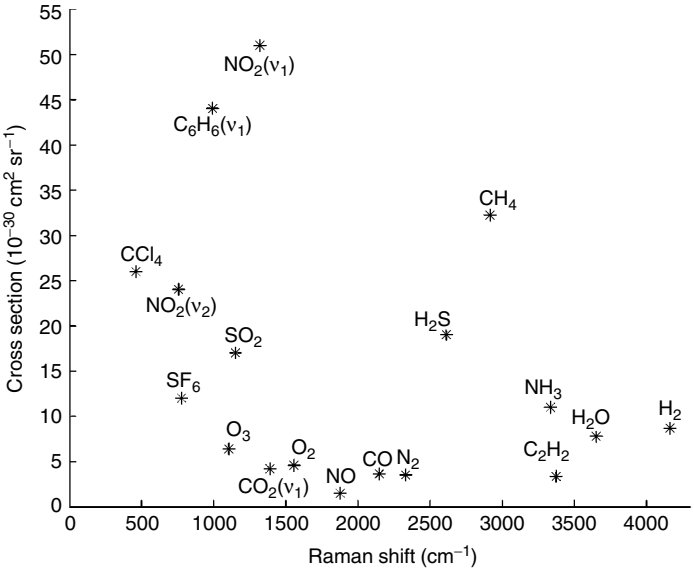


Figure 12. Vibrational Raman frequency shifts and cross sections for a number of molecules found in the atmosphere.

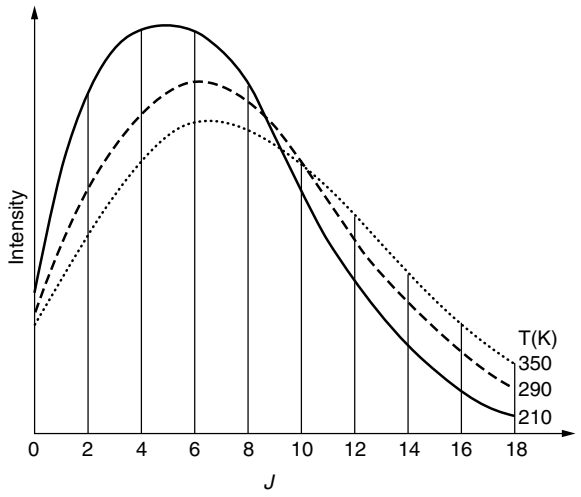


Figure 13. Intensity distribution of PRRS for N_2 at three temperatures.

atmospheric water vapor concentration by Raman lidar requires measuring the Raman backscatter from both water vapor and molecular nitrogen. The nitrogen signal is used as a reference to determine the water vapor mixing ratio from the lidar's Raman water vapor signal.

There are two methods by which Raman lidar can be used to determine atmospheric temperature. In the upper troposphere and throughout the stratosphere, the Rayleigh lidar temperature retrieval algorithm can be applied to appropriately corrected Raman N_2 measurements. Due to its spectral shift, the Raman component of the scatter from N_2 is free from the contamination of scattering from aerosols. However, aerosols affect the optical transmission of the atmosphere, an effect for which the Raman N_2 signal must be corrected before it is used for temperature calculations (246–248). Unlike Rayleigh temperature retrieval, here, the transmission is not constant with altitude. The characteristics of the background stratospheric aerosol layer are known well enough that the correction for atmospheric transmission leads to an acceptable uncertainty in calculated temperatures. However, this correction cannot be made with sufficient accuracy lower in the atmosphere and during increased loading of the stratospheric aerosol layer.

Cooney (249) was the first to propose temperature measurement based on the shape of the PRRS for molecular nitrogen. This method uses the variation in the population of the rotational levels of a molecule with temperature; at higher temperature, the probability that a higher level is populated is greater. Figure 13 shows the envelope of the PRRS lines of a nitrogen molecule at three temperatures. Thus, temperature measurements can be made by measuring the intensity of some or all of the PRRS lines. This differential technique determines the temperature from the intensity of the Raman backscatter across a very narrow wavelength range. Changes in atmospheric transmission due to changes in aerosol properties and loading are insignificant across such a small wavelength range, making the technique almost independent of aerosols.

Separation of the central Rayleigh line from the PRRS has proved to be very difficult, even though the backscatter cross section for PRRS is much greater than that for vibrational-rotational Raman scattering. For example, for the N_2 molecule, the backscatter cross sections for vibrational, pure-rotational and elastic scattering are 3.5×10^{-30}, 1.1×10^{-28} and 3.9×10^{-27}, respectively. The spectral separation of the PRRS and the central unshifted line is quite small, and this leads to technical difficulties when trying to separate these two signals. Nevertheless, a number of Raman lidar systems have been constructed that infer temperature from rotational Raman spectra (250–255).

Resonance Lidar

Resonant scattering occurs when the energy of an incident photon is equal to the energy of an allowed transition within an atom. This is an elastic process; the atom absorbs the photon and instantly emits another photon at the same frequency. As each type of atom and molecule

has a unique absorption and hence, fluorescent spectrum, these measurements may be used to identify and measure the concentration of a particular species. A description of the theory of fluorescence and resonance can be found in both Chamberlain (256) and Measures (22).

The constant ablation of meteors in the earth's upper atmosphere leads to the existence of extended layers of alkali metals in the 80 to 115 km region (257). These metals have low abundances but very high resonant-scattering cross sections. Because resonant scattering involves an atomic transition between allowed energy levels, the probability that this process occurs is much greater than that for Rayleigh scattering. For instance, at 589 nm, the resonance-fluorescence cross section for sodium is about 10^{15} times larger than the cross section for Rayleigh scattering from air. This means that the lidar signal from 85 km measured by a sodium resonance-fluorescence lidar is about the same as the Rayleigh scatter signal measured by the same lidar at about 30 km.

Sodium. Atmospheric sodium is the most widely used of the alkali metal layers in the atmosphere because it is relatively abundant and the transmitter frequency is easy to generate. Several research groups have measured the climatology of sodium abundance, parameters related to gravity wave dynamics, temperatures, and winds (83,258–265). The sodium layer exists in the earth's atmosphere between about 80 and 105 km in altitude, a region that covers the upper part of the mesosphere and the lower part of the thermosphere. This sodium layer is sometimes referred to as the mesospheric sodium layer, although it extends well above the top of the mesosphere. The first reported use of a resonance lidar to study sodium was in 1969 (266). The existence of the mesospheric sodium layer had been known many years previous to these first lidar measurements, due to the bright, natural airglow emission that was extensively studied using passive spectroscopy (267). These passive instruments could resolve the height structure of the region only during sunrise and sunset.

The spectral shape of the sodium line at 589 nm, the D_{2a} line, is temperature-dependent and the scattering cross section is proportional to the line shape. Using this information allows the measurement of the temperature of the sodium atoms and the atmosphere surrounding them from the spectral shape of the backscattered intensity. Figure 14 shows the shape of the sodium D_{2a} line for three temperatures that are within the range of temperatures expected around the mesopause region. The sodium D_{2a} shape has been measured by lidar in a number of ways (268,269). Usually, this measurement is achieved by transmitting narrow bandwidth laser pulses at two or three well-known frequencies within the sodium D_{2a} line and recording the backscatter intensity at each of the transmitted frequencies separately. By knowing the frequency of the transmitted laser pulses and the intensity of the backscatter at each of the transmitted frequencies, the atmospheric temperature can be determined.

A technique known as Doppler-free saturation spectroscopy is used to set the frequency of the laser transmitted into the atmosphere very precisely.

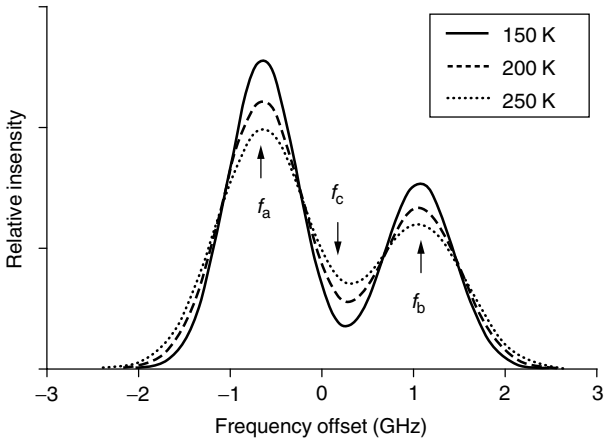


Figure 14. Shape of the sodium D_{2a} line at three temperatures.

Counterpropagating a sample of the laser output through a laboratory cell that contains sodium vapor generates the Doppler-free saturation spectrum. Under the right conditions, the fluorescence from the cell contains sharp spectral features (270) (Fig. 15). Measurements of these Doppler-free features are used in a feedback loop to control the output frequency of the laser and to lock the laser's output frequency to the frequency of the spectral feature (83,271). The Doppler-free spectrum of sodium provides three features that offer the possibility of locking the laser; f_a, f_b, and f_c. The atmospheric temperature can be determined from the ratio of the backscattered intensity at any two of three available frequencies. The pair of frequencies, which has the largest change in ratio with temperature, is f_a and f_c and so these two frequencies are commonly used. This method of temperature measurement is a direct spectral measurement and has associated errors several orders of magnitude lower than those associated with Rayleigh temperature measurements in this altitude range.

A slight drawback of this method is that it typically takes 5 to 10 seconds to switch the laser from one frequency to the other, f_a to f_c, or back again. To obtain a reasonable duty cycle, it is therefore necessary to operate the laser at each frequency for typically 30 to 60 seconds. The temperature is then determined from the ratio of measurements taken at slightly different times. The variability of the sodium and the atmosphere over this short timescale leads to some uncertainty in the temperatures measured using this technique (270).

Improvements in transmitter technology during the last decade have allowed winds as well as temperatures to be measured using narrowband sodium lidar systems (270,272,273) incorporating an acousto-optic (AO) modulator. The AO modulators are used to switch the transmitted frequency several hundred MHz to either side of a selected Doppler-free feature. This tuning enables measuring the Doppler shift and the width of the backscattered light simultaneously. Acousto-optic modulators can be turned on and off very quickly; this feature allows frequency switching between transmitted laser pulses. Typically a sodium temperature-wind lidar operates at three frequencies f_a and f_a plus and minus the AO offset. Today, such systems have been extended to a large scale, for example, the sodium lidar operated at the Starfire

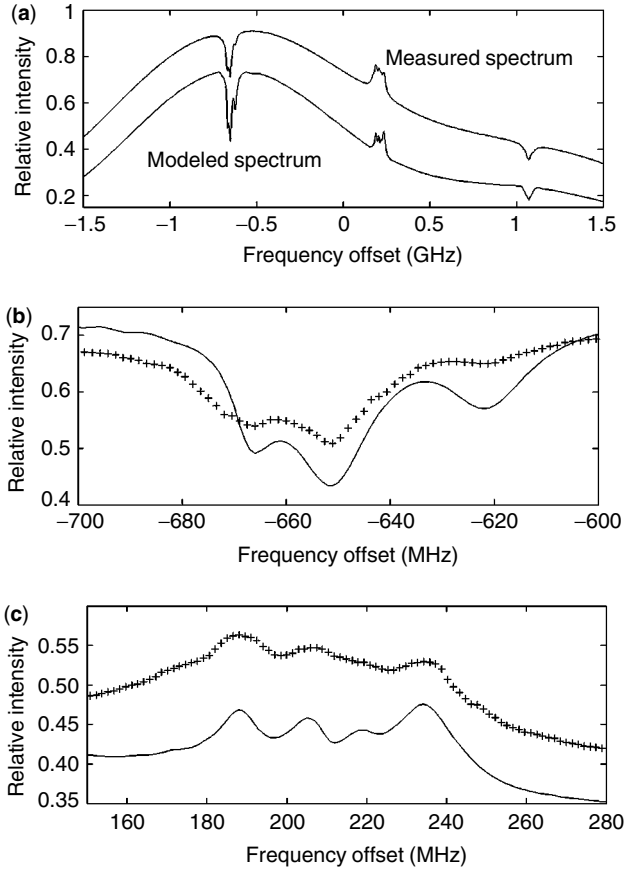


Figure 15. The Doppler-free-saturation spectra for the sodium D_{2a} line showing the locations of the spectral features f_a, f_b, and f_c. (**a**) D_{2a} line. (**b**) closeup of f_a, solid line is modeled '+'s are measured. (**c**) closeup of f_c.

Optical Range (SOR). Figure 16 shows an example of temperature measurements made at SOR. By simultaneously measuring temperature and vertical wind velocity, measurements at SOR have been used for the first determinations of the vertical flux of heat due to gravity waves in the mesopause region (40).

Other Metallic Species. Other alkali metals, including calcium (Ca and Ca^+) (274,275), potassium (276,277), lithium (278,279), and iron (280,281), that have resonance lines in the blue region of the visible spectrum, have also been used to study the mesopause region of the Earth's atmosphere. Thomas (282) reviews the early work in this field. Resonance lidar requires laser transmissions at the precise frequency of an absorption line of the species being studied. Traditionally, dye lasers have been used successfully to probe many of these species, though working with these dyes is difficult in the field environment. Recently, solid-state lasers have been applied to resonance lidar systems (283).

SUMMARY

Lidar has established itself as one of the most important measurements techniques for atmospheric composition

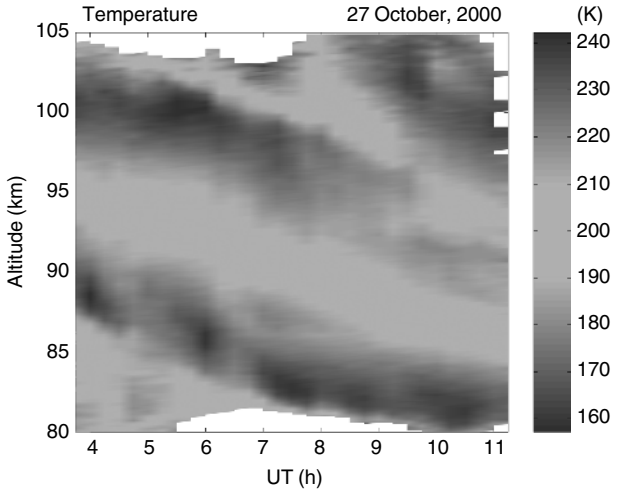


Figure 16. Temperature in the mesopause region of the atmosphere measured by the University of Illinois Sodium Wind and Temperature Lidar over the Starfire Optical Range (35.0N,106.5W), near Albuquerque, New Mexico, USA, on 27 October 2000. The local time is UT (Universal Time) 7 hours. Measurements shown in this image have been smoothed by about 0.5 hour in time and 0.5 km in altitude. The downward phase progression of the atmospheric tidal structure is clearly shown as the temperature structure move downward with time (courtesy of the University of Illinois lidar group). See color insert.

and dynamics from the surface to the upper atmosphere. It also has important uses in mapping, bathymetry, defense, oceanography and natural resource management. Lidar solutions offer themselves for a wide range of environmental monitoring problems. Except for the LITE experiment (184,185), present lidars systems are primarily located on the surface or, for campaign use, on aircraft. The next decade promises the launch of several significant space-based lidar systems to study the Earth's atmosphere. These systems include experiments to measure clouds on a global scale, for example, the GLAS (284,285), ATLID (286), and ESSP3–CENA (287) instruments, as well as ORACLE, (288) a proposed instrument to measure global ozone distribution. These space-based missions will complement existing ground-based systems by increasing global coverage. A new, ground-based, multitechnique lidar called ALOMAR (261) promises to provide measurements of air density, temperature, 3-D wind vector, momentum fluxes, aerosols, cloud particles, and selected trace gases at high vertical and temporal resolution.

The new millennium will bring synergistic combinations of space and ground-based radar and lidar facilities that will greatly enhance our ability to predict weather and climatic changes by making available measurements of wind, temperature, composition, and cloud properties.

ABBREVIATIONS AND ACRONYMS

ATLID	atmospheric lidar
ALOMAR	arctic lidar observatory for middle atmosphere research
AO	acousto-optic
CCD	charge coupled device
CNRS	centre natural de la recherche scientifique

cw	continuous wave
DIAL	differential absorption lidar
DOAS	differential optical absorption spectroscopy
ESSP3	earth system science pathfinder 3
FOV	field-of-view
GLAS	geoscience laser altimeter system
Lidar	light detection and ranging
LITE	lidar in space technology experiment
LMT	liquid mirror telescope
MCP	micro channel plate
MCS	multichannel scaler
NASA	national aeronautics and space administration
Nd:YAG	neodymium:yttrium-aluminum garnet
ORACLE	ozone research with advanced cooperative lidar experiment
PCL	purple crow lidar
PMT	photomultiplier tube
PPRS	pure rotational raman lidar
PRF	pulse repetition frequency
RF	radio frequency
SIN	signal induced noise
SOR	starfire optical range
STS	space transport system
UT	Universal time

BIBLIOGRAPHY

1. D. A. Leonard, B. Caputo, and F. E. Hoge, *Appl. Opt.* **18**, 1,732–1,745 (1979).
2. J. L. Irish and T. E. White, *Coast. Eng.* **35**, 47–71 (1998).
3. R. Barbini et al., *ICES J. Mar. Sci.* **55**, 793–802 (1998).
4. I. M. Levin and K. S. Shifrin, *Remote Sensing Environ.* **65**, 105–111 (1998).
5. J. H. Churnside, V. V. Tatarskii, and J. J. Wilson, *Appl. Opt.* **37**, 3,105–3,112 (1998).
6. N. S. Higdon et al., *Proc. Nineteenth ILRC*, NASA, Hampton, Va., 1998, p. 651.
7. D. R. Alexander, M. L. Rohlfs, and J. C. Stauffer, *Proc. SPIE* **3,082**, 22–29 (1997).
8. G. Klauber, C. Sini, P. M. Brinegar II, and M. M. Williams, *Proc. SPIE* **3,082**, 92–103 (1997).
9. R. A. Mendonsa, *Photon Spectra* **31**, 20 (1997).
10. [ANON], *Laser Focus World* **32**, 13 (1996).
11. W. B. Scott, *Aviat. Week Space Technol.* **143**, 44 (1995).
12. B. T. N. Evans, E. Yee, G. Roy, and J. Ho, *J. Aerosol Sci.* **25**, 1,549–1,566 (1994).
13. A. V. Jelalian, W. H. Keene, and E. F. Pearson, in D. K. Killinger and A. Mooradian, eds., *Optical and Laser Remote Sensing*, Springer-Verlag, Berlin, 1983, pp. 341–349.
14. *www.bushnell.com*.
15. *www.leica-camera.com*.
16. U. N. Singh, in *Optical Measurement Techniques and Application*, P. K. Rastogi, ed., Artech House, Norwood, MA, 1997, pp. 369–396.
17. C. Weitkamp, in *Radiation and Water in the Climate System*, E. Raschke, ed., Springer-Verlig, Berlin, Germany, 1996, pp. 217–247.
18. D. K. Killinger and A. Mooradian, eds., *Optical and Laser Remote Sensing*, Springer-Verlag, Berlin, 1983.
19. L. Thomas, in *Spectroscopy in Environmental Science*, R. J. H. Clark and R. E. Hester, eds., Wiley, Chichester, England, 1995, pp. 1–47.
20. R. Frehlich, in *Trends in Optics: Research, Development and Applications*, A. Consortini, ed., Academic Press, London, England, 1996, pp. 351–370.
21. W. B. Grant, in *Tunable Laser Applications*, F. J. Duarte, ed., Marcel Dekker, NY, 1995, pp. 213–305.
22. R. M. Measures, *Laser Remote Sensing: Fundamentals and Applications*, John Wiley & Sons, Inc., New York, NY, 1984.
23. E. H. Synge, *Philos. Mag.* **52**, 1,014–1,020 (1930).
24. Duclaux, *J. Phys. Radiat.* **7**, 361 (1936).
25. E. O. Hulbert, *J. Opt. Soc. Am.* **27**, 377–382 (1937).
26. R. Bureau, *Meteorologie* **3**, 292 (1946).
27. L. Elterman, *J. Geophys. Res.* **58**, 519–530 (1953).
28. S. S. Friedland, J. Katzenstein, and M. R. Zatzick, *J. Geophys. Res.* **61**, 415–434 (1956).
29. T. H. Maiman, *Nature* **187**, 493 (1960).
30. F. J. McClung and R. W. Hellworth, *J. Appl. Phys.* **33**, 828–829 (1962).
31. L. D. Smullins and G. Fiocco, *Nature* **194**, 1,267 (1962).
32. G. Fiocco and L. D. Smullins, *Nature* **199**, 1,275–1,276 (1963).
33. H. Chen et al., *Opt. Lett.* **21**, 1,093–1,095 (1997).
34. S. E. Bisson, J. E. M. Goldsmith, and M. G. Mitchell, *Appl. Opt.* **38**, 1,841–1,849 (1999).
35. D. Rees, U. von Zahn et al., *Adv. Space Res.* **26**, 893–902 (2000).
36. J. D. Spinhirne, *IEEE Trans. Geosci. Remote* **31**, 48 (1993)
37. C. Nagasawa et al., *Appl. Opt.* **29**, 1,466–1,470 (1990).
38. Y. Emery and C. Flesia, *Appl. Opt.* **37**, 2,238–2,241 (1998).
39. R. J. Sica et al., *Appl. Opt.* **43**, 6,925–6,936 (1995).
40. C. S. Gardner and W. M. Yang, *J. Geophys. Res.* **103**, 8,699–8,713 (1998).
41. E. F. Borra and S. Thibault, *Photon Spectra* **32**, 142–146 (1998).
42. G. V. Guerra et al., *J. Geophys. Res.* **104**, 22,287–22,292 (1999).
43. D. M. Chambers and G. P. Nordin, *J. Opt. Soc. Am.* **16**, 1,184–1,193 (1999).
44. *http://control.cass.usu.edu/lidar/index.htm*.
45. S. Ishii et al., *Rev. Sci. Instrum.* **67**, 3,270–3,273 (1996).
46. J. L. Baray et al., *Appl. Opt.* **38**, 6,808–6,817 (1999).
47. K. W. Fischer et al., *Opt. Eng.* **34**, 499–511 (1995).
48. Z. L. Hu et al., *Opt. Commun.* **156**, 289–293 (1998).
49. J. A. McKay, *Appl. Opt.* **38**, 5,851–5,858 (1999).
50. G. Beneditti-Michelangeli, F. Congeduti, and G. Fiocco, *JAS* **29**, 906–910 (1972).
51. V. J. Abreu, J. E. Barnes, and P. B. Hays, *Appl. Opt.* **31**, 4,509–4,514 (1992).
52. S. T. Shipley et al., *Appl. Opt.* **22**, 3,716–3,724 (1983).
53. G. Fiocco and J. B. DeWolf, *JAS* **25**, 488–496 (1968).
54. M. L. Chanin et al., *J. Geophy. Res.* **16**, 1,273–1,276 (1989).
55. A. I. Carswell, in D. K. Killinger and A. Mooradian, eds., *Optical and Laser Remote Sensing*, Springer-Verlag, Berlin, 1983, pp. 318–326.
56. K. Sassen, R. P. Benson, and J. D. Spinhirne, *Geophys. Res. Lett.* **27**, 673–676 (2000).
57. G. P. Gobbi, *Appl. Opt.* **37**, 5,505–5,508 (1998).
58. F. Cairo et al., *Appl. Opt.* **38**, 4,425–4,432 (1999).
59. F. Cairo et al., *Rev. Sci. Instrum.* **67**, 3,274–3,280 (1996).
60. J. P. Thayer et al., *Opt. Eng.* **36**, 2,045–2,061 (1997).

61. P. S. Argall and F. Jacka, *Appl. Opt.* **35**, 2,619–2,629 (1996).
62. F. L. Pedrotti and L. S. Pedrotti, *Introduction to Optics*, 2nd ed., Prentice-Hall, Englewood Cliffs, NJ, 1993, pp. 24–25.
63. E. L. Dereniak and D. G. Crowe, *Optical Radiation Detectors*, John Wiley & Sons, Inc., New York, NY, 1984, pp. 116–121.
64. M. J. McGill and W. R. Skinner, *Opt. Eng.* **36**, 139–145 (1997).
65. W. C. Priedhorsky, R. C. Smith, and C. Ho, *Appl. Opt.* **35**, 441–452 (1996).
66. T. Erikson et al., *Appl. Opt.* **38**, 2,605–2,613 (1999).
67. N. S. Higdon et al., *Appl. Opt.* **33**, 6,422–6,438 (1994).
68. M. Wu et al., *Appl. Spectrosc.* **54**, 800–806 (2000).
69. A. M. South, I. M. Povey, and R. L. Jones, *J. Geophys. Res.* **103**, 31,191–31,202 (1998).
70. R. W. Engstrom, *Photomultiplier Handbook*, RCA Corporation, USA, 1980.
71. J. Wilson and J. F. B. Hawkes, *Optoelectronics, An Introduction*, 2nd ed., Prentice-Hall, Cambridge, 1989, pp. 265–270.
72. D. P. Donovan, J. A. Whiteway, and A. I. Carswell, *Appl. Opt.* **32**, 6,742–6,753 (1993).
73. A. O. Langford, *Appl. Opt.* **34**, 8,330–8,340 (1995).
74. M. P. Bristow, D. H. Bundy, and A. G. Wright, *Appl. Opt.* **34**, 4,437–4,452 (1995).
75. Y. Z. Zhao, *Appl. Opt.* **38**, 4,639–4,648 (1999).
76. C. K. Williamson and R. J. De Young, *Appl. Opt.* **39**, 1,973–1,979 (2000).
77. J. M. Vaughan, *Phys. Scripta* **T78**, 73–81 (1998).
78. R. M. Huffaker and P. A. Reveley, *Pure Appl. Opt.* **7**, 863–873 (1998).
79. R. Targ et al., *Appl. Opt.* **35**, 7,117–7,127 (1996).
80. R. M. Huffaker and R. M. Hardesty, *Proc. IEEE* **84**, 181–204 (1996).
81. S. M. Hannon and J. A. Thomson, *J. Mod. Opt.* **41**, 2,175–2,196 (1994).
82. V. M. Gordienko et al., *Opt. Eng.* **33**, 3,206–3,213 (1994).
83. P. S. Argall et al., *Appl. Opt.* **39**, 2,393–2,400 (2000).
84. A. Ben-David, *Appl. Opt.* **38**, 2,616–2,624 (1999).
85. Y. J. Park, S. W. Dho, and H. J. Kong, *Appl. Opt.* **36**, 5,158–5,161 (1997).
86. K. L. Coulson, *Solar and Terrestrial Radiation*, Academic Press, NY, 1975.
87. E. J. McCartney, *Optics of the Atmosphere*, John Wiley & Sons, Inc., New York, NY, 1976.
88. P. N. Slater, *Remote Sensing, Optics and Optical Systems*, Addison-Wesley, Toronto, 1980.
89. V. V. Sobolev, *Light Scattering in Planetary Atmospheres*, Pergamon Press, Oxford, 1975.
90. G. Mie, *Ann. Physik* **25**, 377–445 (1908).
91. D. Muller et al., *Appl. Opt.* **39**, 1,879–1,892 (2000).
92. J. P. Diaz et al., *J. Geophys. Res.* **105**, 4,979–4,991 (2000).
93. F. Masci, *Ann. Geofis.* **42**, 71–83 (1999).
94. D. Muller, U. Wandinger, and A. Ansmann, *Appl. Opt.* **38**, 2,358–2,368 (1999).
95. C. Erlick and J. E. Frederick, *J. Geophys. Res.* **103**, 23,275–23,285 (1998).
96. I. N. Sokolik, O. B. Toon, and R. W. Bergstrom, *J. Geophys. Res.* **103**, 8,813–8,826 (1998).
97. A. A. Kokhanovsky, *J. Atmos. Sci.* **55**, 314–320 (1998).
98. W. C. Conant, *J. Geophys. Res.* **105**, 15,347–15,360 (2000).
99. G. M. McFarquhar et al., *J. Atmos. Sci.* **57**, 1,841–1,853 (2000).
100. R. M. Hoff et al., *J. Geophys. Res.* **101**, 19,199–19,209 (1996).
101. J. L. Brenguier et al., *Tellus B* **52**, 815–827 (2000).
102. J. Redemann et al., *J. Geophys. Res.* **105**, 9,949–9,970 (2000).
103. M. Minomura et al., *Adv. Space Res.* **25**, 1,033–1,036 (2000).
104. A. T. Young, *Appl. Opt.* **19**, 3,427–3,428 (1980).
105. A. T. Young, *J. Appl. Meteorol.* **20**, 328–330 (1981).
106. A. T. Young, *Phys. Today* **35**, 42–48 (1982).
107. Rayleigh (J. W. Strutt), *Philos. Mag.* **41**, 274–279 (1871).
108. Rayleigh (J. W. Strutt), *Philos. Mag.* **41**, 447–454 (1871).
109. Rayleigh (J. W. Strutt), *Philos. Mag.* **12**, 81 (1881).
110. Rayleigh (J. W. Strutt), *Philos. Mag.* **47**, 375–384 (1899).
111. J. A. Stratton, *Electromagnetic Theory*, McGraw-Hill, NY, 1941.
112. M. Kerker, *The Scattering of Light and Electromagnetic Radiation*, Academic Press, NY, 1969.
113. R. Penndorf, *J. Opt. Soc. Am.* **52**, 402–408 (1962).
114. W. E. K. Middleton, *Vision Through the Atmosphere*, University of Toronto Press, Toronto, 1952.
115. M. Born and E. Wolf, *Principles of Optics*, Pergamon Press, Great Britain, Oxford, 1970.
116. G. S. Kent and R. W. H. Wright, *J. Atmos. Terrestrial Phys.* **32**, 917–943 (1970).
117. R. T. H. Collis and P. B. Russell, in E. D. Hinkley, ed., *Laser Monitoring of the Atmosphere*, Springer-Verlag, Berlin, 1976.
118. G. Fiocco, in R. A. Vincent, ed., *Handbook for MAP*, vol. 13, ICSU, SCOSTEP, Urbana, IL, 1984.
119. G. Fiocco et al., *Nature* **229**, 79–80 (1971).
120. A. Hauchecorne and M. L. Chanin, *Geophys. Res. Lett.* **7**, 565–568 (1980).
121. T. Shibata, M. Kobuchi, and M. Maeda, *Appl. Opt.* **25**, 685–688 (1986).
122. C. O. Hines, *Can. J. Phys.* **38**, 1,441–1,481 (1960).
123. R. J. Sica and M. D. Thorsley, *Geophys. Res. Lett.* **23**, 2,797–2,800 (1996).
124. T. J. Duck, J. A. Whiteway, and I. A. Carswell, *J. Geophys. Res.* **105**, 22,909–22,918 (2000).
125. T. Leblanc et al., *J. Geophys. Res.* **103**, 17,191–17,204 (1998).
126. M. L. Chanin and A. Hauchecorne, *J. Geophys. Res.* **86**, 9,715–9,721 (1981).
127. M. L. Chanin and A. Hauchecorne, in R. A. Vincent, ed., *Handbook for MAP*, vol. 13, ICSU, SCOSTEP, Urbana, IL, 1984, pp. 87–98.
128. M. L. Chanin, N. Smires, and A. Hauchecorne, *J. Geophys. Res.* **92**, 10,933–10,941 (1987).
129. C. S. Gardner, M. S. Miller, and C. H. Liu, *J. Atmos. Sci.* **46**, 1,838–1,854 (1989).
130. T. J. Beatty, C. A. Hostetler, and C. S. Gardner, *J. Atmos. Sci.* **49**, 477–496 (1992).
131. A. I. Carswell et al., *Can. J. Phys.* **69**, 1,076 (1991).
132. M. M. Mwangi, R. J. Sica, and P. S. Argall, *J. Geophys. Res.* **106**, 10,313 (2001).
133. R. J. States and C. S. Gardner, *J. Geophys. Res.* **104**, 11,783–11,798 (1999).
134. E. A. Hyllerass, *Mathematical and Theoretical Physics*, John Wiley & Sons, Inc., New York, NY, 1970.

135. E. H. Kennard, *Kinetic Theory of Gases*, McGraw-Hill, NY, 1938.
136. C. A. Tepley, S. I. Sargoytchev, and R. Rojas, *IEEE Trans. Geosci. Remote Sensing* **31**, 36–47 (1993).
137. C. Souprayen et al., *Appl. Opt.* **38**, 2,410–2,421 (1999).
138. H. C. Van de Hulst, *Light Scattering by Small Particles*, John Wiley & Sons, Inc., New York, NY, 1951.
139. C. E. Bohren and D. R. Huffman, *Absorption and Scattering of Light by Small Particles*, John Wiley & Sons, Inc., New York, NY, 1983.
140. L. P. Bayvel and A. R. Jones, *Electromagnetic Scattering and its Applications*, Applied Science, England, London, 1981.
141. C. N. Davies, *J. Aerosol. Sci.* **18**, 469–477 (1987).
142. L. G. Yaskovich, *Izvestiya, Atmos. Oceanic Phys.* **22**, 640–645 (1986).
143. Y. S. Georgiyevskiy et al., *Izvestiya, Atmos. Oceanic Phys.* **22**, 646–651 (1986).
144. J. Rosen et al., *J. Geophys. Res.* **105**, 17,833–17,842 (2000).
145. A. Ansmann et al., *Geophys. Res. Lett.* **27**, 964–966 (2000).
146. T. Sakai et al., *Atmos. Environ.* **34**, 431–442 (2000).
147. M. A. Fenn et al., *J. Geophys. Res.* **104**, 16,197–16,212 (1999).
148. R. A. Ferrare et al., *J. Geophys. Res.* **103**, 19,673–19,689 (1998).
149. C. E. Junge, *Air Chemistry and Radioactivity*, Academic Press, NY, 1963.
150. G. Di Donfrancesco et al., *J. Atmos. Sol-Terrestrial Phys.* **62**, 713–723 (2000).
151. D. Guzzi et al., *Geophys. Res. Lett.* **26**, 2,199–2,202 (1999).
152. V. V. Zuev, V. D. Burlakov, and A. V. El'nikov, *J. Aersol Sci.* **103**, 13,873–13,891 (1998).
153. A. di Sarra et al., *J. Geophys. Res.* **103**, 13,873–13,891 (1998).
154. G. S. Kent and G. M. Hansen, *Appl. Opt.* **37**, 3,861–3,872 (1998).
155. M. H. Bergin et al., *J. Geophys. Res.* **105**, 6,807–6,816 (2000).
156. R. A. Ferrare et al., *J. Geophys. Res.* **103**, 19,663–19,672 (1998).
157. C. M. R. Platt et al., *J. Atmos. Sci.* **55**, 1,977–1,996 (1998).
158. A. Robock, *Rev. Geophys.* **38**, 191–219 (2000).
159. W. T. Hyde and T. J. Crowley, *J. Climate* **13**, 1,445–1,450 (2000).
160. H. Kuhnert et al., *Int. J. Earth Sci.* **88**, 725–732 (2000).
161. C. J. Grund and E. W. Eloranta, *Opt. Eng.* **30**, 6–12 (1991).
162. C. Y. She et al., *Appl. Opt.* **31**, 2,095–2,106 (1992).
163. D. P. Donovan et al., *Geophys. Res. Lett.* **25**, 3,139–3,142 (1998).
164. Y. Sasano and E. V. Browell, *Appl. Opt.* **28**, 1,670–1,679 (1989).
165. D. Muller et al., *Geophys. Res. Lett.* **27**, 1,403–1,406 (2000).
166. G. Beyerle et al., *Geophys. Res. Lett.* **25**, 919–922 (1998).
167. M. J. Post et al., *J. Geophys. Res.* **102**, 13,535–13,542 (1997).
168. J. D. Spinhirne et al., *Appl. Opt.* **36**, 3,475–3,490 (1997).
169. T. Murayama et al., *J. Geophys. Res.* **104**, 3,1781–3,1792 (1999).
170. G. Roy et al., *Appl. Opt.* **38**, 5,202–5,211 (1999).
171. K. Sassen K and C. Y. Hsueh, *Geophys. Res. Lett.* **25**, 1,165–1,168 (1998).
172. T. Murayama et al., *J. Meteorol. Soc. Jpn.* **74**, 571–578 (1996).
173. K. Sassen, *Bull. Am. Meteorol. Soc.* **72**, 1,848–1,866 (1991).
174. G. A. Reagan, J. D. Spinhirne, and M. P. McCormick, *Proc. IEEE* **77**, 433–448 (1989).
175. D. P. Donovan et al., *Geophys Res. Lett.* **24**, 2,709–2,712 (1997).
176. G. P. Gobbi, G. Di Donfrancesco, and A. Adriani, *J. Geophys. Res.* **103**, 10,859–10,873 (1998).
177. M. Pantani et al., *J. Aerosol Sci.* **30**, 559–567 (1999).
178. H. Mehrtens et al., *Geophys. Res. Lett.* **26**, 603–606 (1999).
179. T. Shibata et al., *J. Geophys Res.* **104**, 21,603–21,611 (1999).
180. A. Tsias et al., *J. Geophys. Res.* **104**, 23,961–23,969 (1999).
181. F. Stefanutti et al., *Appl. Phy.* **B55**, 13–17 (1992).
182. K. S. Carslaw et al., *Nature* **391**, 675–678 (1998).
183. B. M. Knudsen et al., *Geophys. Res. Lett.* **25**, 627–630 (1998).
184. I. A. MacKenzie and R. S. Harwood, *J. Geophys. Res.* **105**, 9,033–9,051 (2000).
185. *http://oea.larc.nasa.gov/PAIS/LITE.html*.
186. D. M. Winker, R. H. Couch, and M. P. McCormick, *Proc. IEEE* **84**, 164–180 (1996).
187. L. O'Connor, *Mech. Eng.* **117**, 77–79 (1995).
188. K. B. Strawbridge and R. M. Hoff, *Geophys. Res. Lett.* **23**, 73–76 (1996).
189. Y. Y. Y. Gu et al., *Appl. Opt.* **36**, 5,148–5,157 (1997).
190. V. Cuomo et al., *J. Geophys. Res.* **103**, 11,455–11,464 (1998).
191. H. Shimizu, S. A. Lee, and C. Y. She, *Appl. Opt.* **22**, 1,373–1,382 (1983).
192. R. M. Hardesty, in D. K. Killinger and A. Mooradian, eds., *Optical and Laser Remote Sensing*, Springer-Verlag, Berlin, 1983.
193. S. D. Mayor et al., *J. Atmos. Ocean. Tech.* **14**, 1,110–1,126 (1997).
194. J. Bilbro, in D. K. Killinger and A. Mooradian, eds., *Optical and Laser Remote Sensing*, Springer-Verlag, Berlin, 1983.
195. J. Rothermel et al., *Opt. Express* **2**, 40–50 (1998).
196. J. Rothermel et al., *Bull. Am. Meteorol. Soc.* **79**, 581–599 (1998).
197. R. Frehlich, *J. Appl. Meteorol* **39**, 245–262 (2000).
198. R. M. Schotland, *Proc. 3rd Symp. Remote Sensing Environ.*, 1964, pp. 215–224.
199. R. M. Schotland, *Proc. 4th Symp. Remote Sensing Environ.*, 1966, pp. 273–283.
200. D. K. Killinger and N. Menyuk, *Science* **235**, 37–45 (1987).
201. K. W. Rothe, U. Brinkmann, and H. Walther, *Appl. Phys.* **3**, 115 (1974).
202. N. Menyuk, D. K. Killinger, and W. E. DeFeo, in D. K. Killinger and A. Mooradian, eds., *Optical and Laser Remote Sensing*, Springer-Verlag, Berlin, 1983.
203. E. E. Uthe, *Appl. Opt.* **25**, 2,492–2,498 (1986).
204. E. Zanzottera, *Crit. Rev. Anal. Chem.* **21**, 279 (1990).
205. M. Pinandito et al., *Opt. Rev.* **5**, 252–256 (1998).
206. R. Toriumi et al., *Jpn. J. Appl. Phys.* **38**, 6,372–6,378 (1999).
207. R. Toriumi, H. Tai, and N. Takeuchi, *Opt. Eng.* **35**, 2,371–2,375 (1996).
208. D. Kim et al., *J. Korean Phys. Soc.* **30**, 458–462 (1997).
209. V. Wulfmeyer, *J. Atmos. Sci.* **56**, 1,055–1,076 (1999).
210. A. Fix, V. Weiss, and G. Ehret, *Pure Appl. Opt.* **7**, 837–852 (1998).

211. E. V. Browell, *Proc. IEEE* **77**, 419–432 (1989).
212. R. M. Banta et al., *J. Geophys. Res.* **103**, 22,519–22,544 (1998).
213. E. Durieux et al., *Atmos. Environ.* **32**, 2,141–2,150 (1998).
214. P. Weibring et al., *Appl. Phys. B* **67**, 419–426 (1998).
215. T. Fukuchi et al., *Opt. Eng.* **38**, 141–145 (1999).
216. N. S. Prasad and A. R. Geiger, *Opt Eng.* **35**, 1,105–1,111 (1996).
217. M. J. T. Milton et al., *Opt. Commun.* **142**, 153–160 (1997).
218. K. Ikuta et al., *Jpn. J. Appl. Phys.* **38**, 110–114 (1999).
219. J. E. Kalshoven et al., *Appl. Opt.* **20**, 1,967–1,971 (1981).
220. G. K. Schwemmer et al., *Rev. Sci. Instrum.* **58**, 2,226–2,237 (1987).
221. V. Wulfmeyer, *Appl. Opt.* **37**, 3,804–3,824 (1998).
222. J. Pelon, S. Godin, and G. Megie, *J. Geophys. Res.* **91**, 8,667–8,671 (1986).
223. J. Werner, K. W. Rothe, and H. Walther, *Appl. Phys., Ser. B.* **32**, 113 1983.
224. I. S. McDermid, S. M. Godin, and L. O. Lindqvist, *Appl. Opt.* **29**, 3,603–3,612 (1990).
225. T. J. McGee et al., *Opt. Eng.* **30**, 31–39 (1991).
226. T. Leblanc and I. S. McDermid, *J. Geophys. Res.* **105**, 14,613–14,623 (2000).
227. W. B. Grant et al., *Geophys. Res. Lett.* **25**, 623–626 (1998).
228. T. J. McGee et al., *Opt. Eng.* **34**, 1,421–1,430 (1995).
229. J. R. Quagliano et al., *Appl. Opt.* **36**, 1,915–1,927 (1997).
230. C. Bellecci and F. De Donato, *Appl. Opt.* **38**, 5,212–5,217 (1999).
231. G. Herzberg, *Molecular Spectra and Molecular Structure I. Spectra of Diatomic Molecules*, 2nd ed., Van Nostrand Reinhold Company, NY, 1950.
232. C. V. Raman, *Indian J. Phys.* **2**, 387 (1928).
233. J. A. Cooney, *Appl. Phys. Lett.* **12**, 40–42 (1968).
234. D. A. Leonard, *Nature* **216**, 142–143 (1967).
235. J. A. Cooney, *J. Appl. Meteorol.* **9**, 182 (1970).
236. J. A. Cooney, *J. Geophys. Res.* **77**, 1,078 (1972).
237. J. A. Cooney, K. Petri, and A. Salik, *Appl. Opt.* **24**, 104–108 (1985).
238. S. H. Melfi, *Appl. Opt.* **11**, 1,605 (1972).
239. V. Sherlock et al., *Appl. Opt.* **38**, 5,838–5,850 (1999).
240. W. E. Eichinger et al., *J. Atmos. Oceanic Technol.* **16**, 1,753–1,766 (1999).
241. S. H. Melfi et al., *Appl. Opt.* **36**, 3,551–3,559 (1997).
242. D. N. Whiteman and S. H. Melfi, *J. Geophys. Res.* **104**, 31,411–31,419 (1999).
243. B. Demoz et al., *Geophys Res. Lett.* **27**, 1,899–1,902 (2000).
244. A. Ansmann et al., *J. Atmos. Sci.* **54**, 2,630–2,641 (1997).
245. R. Ferrare et al., *J. Geophys. Res.* **105**, 9,935–9,947 (2000).
246. P. Keckhut, M. L. Chanin, and A. Hauchecorne, *Appl. Opt.* **29**, 5,182–5,186 (1990).
247. K. D. Evans et al., *Appl. Opt.* **36**, 2,594–2,602 (1997).
248. M. R. Gross et al., *Appl. Opt.* **36**, 5,987–5,995 (1997).
249. J. A. Cooney, *J. Appl. Meteorol.* **11**, 108–112 (1972).
250. A. Cohen, J. A. Cooney, and K. N. Geller, *Appl. Opt.* **15**, 2,896 (1976).
251. J. A. Cooney and M. Pina, *Appl. Opt.* **15**, 602 (1976).
252. R. Gill et al., *Izvestiya, Atmos. Oceanic Phys.* **22**, 646–651 (1979).
253. Yu. F. Arshinov et al., *Appl. Opt.* **22**, 2,984–2,990 (1983).
254. Yu. F. Arshinov and S. M. Bobrovnikov, *Appl. Opt.* **38**, 4,635–4,638 (1999).
255. A. Behrendt and J. Reichardt, *Appl. Opt.* **39**, 1,372–1,378 (2000).
256. J. W. Chamberlain, *Physics of Aurora and Airglow*, Academic Press, NY, 1961.
257. J. M. C. Plane, R. M. Cox, and R. J. Rollason, *Adv. Space Res.* **24**, 1,559–1,570 (1999).
258. V. W. J. H. Kirchhoff et al., *J. Geophys. Res.* **91**, 13,303–13,307 (1986).
259. B. R. Clemesha et al., *Geophys. Res. Lett.* **26**, 1,681–1,684 (1999).
260. K. H. Fricke and U. von Zahn, *J. Atmos. Terrestrial Phys.* **47**, 499–512 (1985).
261. U. von Zahn et al., *Ann. Geophys.-Atmos. Hydr.* **18**, 815–833 (2000).
262. C. S. Gardner et al., *J. Geophys. Res.* **91**, 13,659–13,673 (1986).
263. X. Z. Chu et al., *Geophys. Res. Lett.* **27**, 1,815–1,818 (2000).
264. A. Nomura et al., *Geophys. Res. Lett.* **14**, 700–703 (1987).
265. C. Y. She et al., *Geophys. Res. Lett.* **22**, 377–380 (1995).
266. M. R. Bowman, A. J. Gibson, and M. C. W. Sandford, *Nature* **221**, 456–457 (1969).
267. D. M. Hunten, *Space Sci. Rev.* **6**, 493 (1967).
268. A. Gibson, L. Thomas, and S. Bhattachacharyya, *Nature* **281**, 131–132 (1979).
269. K. H. Frick and U. von Zahn, *J. Atmos. Terrestrial Phys.* **47**, 499–512 (1985).
270. G. C. Papen, W. M. Pfenninger, and D. M. Simonich, *Appl. Opt.* **34**, 480–498 (1995).
271. C. Y. She et al., *Geophys. Res. Lett.* **17**, 929–932 (1990).
272. C. Y. She and J. R. Yu, *Geophys. Res. Lett.* **21**, 1,771–1,774 (1994).
273. R. E. Bills, C. S. Gardner, and C. Y. She, *Opt. Eng.* **30**, 13–21 (1991).
274. C. Granier, J. P. Jegou, and G. Megie, *Proc. 12th Int. Laser Radar Conf.*, Aix en Provence, France, 1984, pp. 229–232.
275. M. Alpers, J. Hoffner, and U. von Zahn, *Geophys. Res. Lett.* **23**, 567–570 (1996).
276. G. C. Papen, C. S. Gardner, and W. M. Pfenninger, *Appl. Opt.* **34**, 6,950–6,958 (1995).
277. V. Eska, U. von Zahn, and J. M. C. Plane, *J. Geophys. Res.* **104**, 17,173–17,186 (1999).
278. J. P. Jegou et al., *Geophys. Res. Lett.* **7**, 995–998 (1980).
279. B. R. Clemesha, *MAP Handbook* **13**, 99–112 (1984).
280. J. A. Gelbwachs, *Appl. Opt.* **33**, 7,151–7,156 (1994).
281. X. Z. Chu et al., *Geophys. Res. Lett.* **27**, 1,807–1,810 (2000).
282. L. Thomas, *Phil. Trans. R. Soc. Lond. Ser. A* **323**, 597–609 (1987).
283. U. von Zahn and J. Hoffner, *Geophys. Res. Lett.* **23**, 141–144 (1996).
284. J. D. Spinhirne and S. P. Palm, *18th Int. Laser Radar Conf. (ILRC)*, Springer-Verlag, 1996, pp. 213–216.
285. *http://virl.gsfc.nasa.gov/glas/index.html*.
286. A. E. Marini, *ESA Bull-Eur. Space* **95**, 113–118 (1998).
287. *http://essp.gsfc.nasa.gov/cena/index.html*.
288. *http://aesd.larc.nasa.gov/oracle/index.htm*.
289. S. Fleming et al., *Tech. Memoir*, NASA TM-100697, 1988.

LIGHTNING LOCATORS

HAMPTON N. SHIRER
Penn State University
University Park, PA

WILLIAM P. ROEDER
MS 7302, 45 WS/SYR
Patrick AFB, FL

HAMPTON W. SHIRER
University of Kansas
Lawrence, KS

DAVID L. D'ARCANGELO
Delta Airlines
Hartford International Airport
Atlanta, GA

JOBY HILLIKER
Penn State University
University Park, PA

JOE KOVAL
The Weather Channel
Atlanta, GA

NATHAN MAGEE
Penn State University
University Park, PA

INTRODUCTION

Locating lightning in real time is an old problem (1). Radio techniques developed in the early to mid-twentieth century used crossed-loop *cathode-ray direction finders* (CRDF) that provide the bearing but not the range to the lightning source (2–4). *Direction-finding* (DF) systems typically sense the radio signal, known as *atmospherics*, *spherics*, or *'sferics*, that is emitted by lightning and that most listeners of AM radios interpret as interference, *static*, or *radio noise* (5, p. 351). Quite generally, lightning radiates electromagnetic pulses that span an enormous range of frequencies. In this article, the *radio signal* refers to the portion of the electromagnetic spectrum that covers frequencies less than 3×10^8 kilohertz (denoted kHz), or 300 GHz, and the optical signal refers to frequencies greater than 3×10^8 kHz. Also, when not qualified here, electromagnetic radiation refers to the radio signal.

Modern real-time lightning locating systems have their origins in the work of Krider, Noggle, Uman, and Weiman, who published several important papers between the mid-1970s and early 1980s describing the unique characteristics of the electromagnetic waveforms radiated by both cloud–ground and intracloud lightning and their components (6–10). The initial application of their locating method was to identify where cloud–ground strokes might have initiated forest fires in the western United States and Alaska (11). Today, their method provides the basis for the North American Lightning Detection Network (NALDN) (12) operated by Global Atmospherics, Inc. (GAI) of Tucson, Arizona, the combination of the National Lightning Detection Network™ (NLDN) (13–15) in the United States and the Canadian Lightning Detection Network (CLDN) (12). Similar networks, noted in Table 1, are installed in Europe, South America, and Asia. A smaller scale network, the Cloud to Ground Lightning Surveillance System (CGLSS), is operated by the 45th Weather Squadron of the United States Air Force (USAF) at the Cape Canaveral Air Force Station (CCAFS) and by the John F. Kennedy Space Center (KSC) at Cape Canaveral, Florida (16).

These lightning location networks are by no means the only ground-based systems operating in either real time or for research. As listed in Table 1, there are numerous networks, including the long-range Arrival Time Difference (ATD) network operated by the British Meteorological Office at Bracknell in the United Kingdom (17–19); the Long-Range Lightning Detection Network (LRLDN) operated in North America by GAI (20); the Global Position and Tracking Systems® (GPATS) network operated by Global Position and Tracking Systems Pty. Ltd. in Ultimo, New South Wales, Australia (21), that uses an electric field (*E*-field) sensor similar to that in the Lightning Position and Tracking System (LPATS) (22, pp. 160–162), which was incorporated into the NLDN in the mid-1990s (13); the *E*-field Change Sensor Array (EDOT) operated by the Los Alamos National Laboratory (LANL) in Los Alamos, New Mexico (23); the Surveillance et Alerte Foudre par Interférométrie Radioélectrique (SAFIR), a direction-finding system that is marketed by Vaisala Dimensions SA of Meyreuil, France, and is used in several locations in Europe (24,25), Japan (26), and Singapore; the research version of SAFIR, the ONERA three-dimensional interferometric mapper (27), operated by the French Office National d'Etudes et de Recherches Aérospatiales (ONERA); the Lightning Detection and Ranging (LDAR) system operated by the USAF and the United States National Aeronautics and Space Administration (NASA) at CCAFS/KSC (28–31); the deployable Lightning Mapping Array (LMA) or Lightning Mapping System (LMS) operated by the New Mexico Institute of Mining and Technology in Socorro, New Mexico (32–34); networks of electric field mills, among them the Launch Pad Lightning Warning System (LPLWS) operating at the CCAFS/KSC (35–38) and the Electric Field Measurement System (EFMS) operating at the Wallops Flight Facility at Wallops Island, Virginia; and networks of flash counters such as the Cloud–Ground Ratio 3 (CGR3) (39–41) and the Conference Internationale des Grands Reseaux Electriques (CIGRE) (42).

Other systems listed in Table 1 include past and current satellite-mounted sensors such as the Defense Meteorological Satellite Program (DMSP) Operational Linescan System (OLS), which provided data from 1973–1996 (43,44); NASA's Optical Transient Detector (OTD) on the Microlab-1 satellite, which provided data from 1995–2000 (45–47); NASA's Lightning Imaging Sensor (LIS) on the Tropical Rainfall Measuring Mission (TRMM) satellite, which has been providing data since 1997 (31,34,48); the instruments on the Fast On-Orbit Recording of Transient Events (FORTÉ) satellite that has been providing data since 1997 (49–51) and is operated by

Table 1. Many of the Current Lightning Sensing Systems, Organized by Class. Details of Each System are Provided in Table 2. The Radio-Frequency Bands Under Lightning Property are ULF (ultralow frequency, 300 Hz to 3 kHz), VLF (very low frequency, 3 to 30 kHz), LF (low frequency, 30 to 300 kHz), and VHF (very high frequency, 30 to 300 MHz).

	Instrument/Network and Class of System *NL: Network Locator; NM: Network Mapper;* *SSC: Single-Station Counter; SSL: Single-Station Locator;* *WD: Warning Device* **Manufacturer's literature for commercial product*	Acronym	Lightning Property: Electric Field Only (*E*-field)	Lightning Property: Electromagnetic Radiation	Lightning Property: Other	Reference
NL	Arrival Time Difference Network [http:/www.torro.org.uk/sfinfo.htm]	ATD		VLF		(17,18)
NL	Cloud to Ground Lightning Surveillance System	CGLSS		VLF/LF		(16)
NL	Long-Range Lightning Detection Network [http:/ghrc.msfc.nasa.gov/uso/readme/gailong.html]	LRLDN		VLF/LF		(20)
NL	Global Position and Tracking Systems [http:/gpats.com.au/]*	GPATS		VLF/LF		(21)
NL	Los Alamos *E*-field Change Sensor Array [http:/edot.lanl.gov/edot_home.htm]	EDOT	ULF/ VLF/LF			(23)
NL	North American Lightning Detection Network	NALDN		VLF/LF		(12)
NL	National Lightning Detection Network™ [http:/www.glatmos.com/products/data/nldnproducts.html]* [http:/ghrc.msfc.nasa.gov/uso/readme/gai.html]	NLDN		VLF/LF		(13–15)
NL	Canadian Lightning Detection Network [http:/www.weatheroffice.com/lightningnews/]	CLDN		VLF/LF		(12)
	NLDN is combination of Lightning Location and Protection, Inc. (LLP) MDF network with the Atmospheric Research Systems, Inc. (ARSI) TOA network, Lightning Position and Tracking System	LPATS		VLF/LF		(22, pp. 160–162)
	Other Lightning Detection Networks using same instruments as NALDN:					
NL	Bureau of Land Management Alaska Lightning Network [http:/www.nwstc.noaa.gov/d.HMD/Lightning/Alaska.htm]			VLF/LF		
NL	Austrian Lightning Detection & Information System [http:/www.aldis.at/english/index.html]	ALDIS		VLF/LF		
NL	Brazilian LDN [http:/physics.uwstout.edu/staff/jtr/msfc_final_report_99.html]			VLF/LF		(114)
NL	Central European (sensors in Switzerland, Germany, France, Netherlands, Czech Republic, Slovakia, Slovenia, Italy, Hungary, Poland, and Austria) [http:/www.aldis.at/english/index.html]			VLF/LF		

(*Continued*)

Table 1. (*Continued*)

Instrument/Network and Class of System *NL: Network Locator; NM: Network Mapper;* *SSC: Single-Station Counter; SSL: Single-Station Locator;* *WD: Warning Device* **Manufacturer's literature for commercial product*	Acronym	Lightning Property: Electric Field Only (*E*-field)	Lightning Property: Electromagnetic Radiation	Lightning Property: Other	Reference
Other Lightning Detection Networks using same instruments as NALDN (continued):					
NL French LDN [http:/www.meteorage.com/ang/Accueil.htm]			VLF/LF		(218)
NL LPATS network in Israel			VLF/LF		(219)
NL Italian LDN [http:/www.cesi.it/services/sirf/index.cfm]			VLF/LF		
NL Japanese LDN [http:/www.sankosha.co.jp/topics_n/frank.html]			VLF/LF		
NL Nevada Automated Lightning Detection System [http:/www.wrcc.sage.dri.edu/fire/ALDS.html]	ALDS		VLF/LF		
NL Nordic Center for Lightning Research (Denmark, Finland, Iceland, Norway, Sweden) http:/thunder.hvi.uu.se/NCLR/NCLR.html]	NCLR		VLF/LF		(216,217)
NL Slovenian LDN [http:/observer.eimv.si/]			VLF/LF		
NM Lightning Detection and Ranging [http:/ghrc.msfc.nasa.gov/uso/readme/ldar.html] [http:/www.nasatech.com/Briefs/Apr98/KSC11785.html]	LDAR		VHF		(29–31)
NM New Mexico Tech Lightning Mapping Array *or* Lightning Mapping System [http:/ibis.nmt.edu/nmt_lms/]	LMA/ LMS		VHF		(32–34)
NM/NL French National Agency for Aerospace Research (ONERA) Interferometric Three-Dimensional Mapper [http:/www.onera.fr/]*	ONERA-3D		VHF/LF		(27)
NM/ NL Surveillance et Alerte Foudre par Interférométrie Radioélectrique [http:/www.eurostorm.com/]* [http:/www.vaisala.com/]*	SAFIR		VHF/LF		(24,25)
Other Lightning Detection Networks using same instruments as SAFIR					
SAFIRKEPCO Network in Kansai, and Tokyo, Japan SAFIRJWA Network in Tokyo, Japan SAFIR 3,000 JMA National Network for Meteorological Service, Japan SAFIR 3,000 IRM National Network for Meteorological Service, Belgium SAFIR KNMI National Network for Meteorological Service, Netherlands					
NM/ NL SAFIR 3,000 HMS National Network for Meteorological Service, Hungary SAFIR 3,000 SHMI National Network for Meteorological Service, Slovakia SAFIR MSS National Network for Meteorological Service, Singapore SAFIR 3,000 IMGW National Network for Meteorological Service, Poland SAFIR DIRIC Network in Paris area for Meteo France, France SAFIR 3,000 Sud-Est, Network in southeast France SAFIR CSG Network European Space Center (CSG), Kourou, French Guiana SAFIR DGA/CEL Network Defense Test Center (CEL), France SAFIR 3,000 Network for Institute of Meteorology, Hanover University, Germany	SAFIR		VHF/LF		

(*Continued*)

Table 1. (*Continued*)

Instrument/Network and Class of System *NL: Network Locator; NM: Network Mapper;* *SSC: Single-Station Counter; SSL: Single-Station Locator;* *WD: Warning Device* **Manufacturer's literature for commercial product*		Acronym	Lightning Property			Reference
			Electric Field Only (*E*-field)	Electromagnetic Radiation	Other	
NM	Thunder Recording Employed in Mapping Branched Lightning Events [http:/www.cs.unc.edu/~stotts/145/OLD99/homes/tremble/]*	TREMBLE		Event Trigger	Sound	
SSC	Cloud–Ground Ratio network of Flash Counters	CGR3	Electrostatic			(39–41)
SSC	Conference Internationale des Grands Reseaux Electriques Lightning Flash Counters [http:/retd.edf.fr/gb/futur/mois/foudre/detect/a.en.htm]	CIGRE	VLF			(42)
SSC	Electrical Storm Identification Device/ASOS Lightning Sensor (TSS 924) [http:/www.glatmos.com/products/local/localarea.html]*	ESID/ALS	VLF/LF	Light		(54), (155)
SSC	Guardian Angel (also Sky Scan™) Lightning Detector [http:/lightningdetector.com/angel/index.html]*			VLF		
SSC	Lightning Alert® Lightning Detector [http:/www.stormwise.com/]*			VLF		
SSC	StrikeAlert Personal Lightning Detector [http:/www.strikealert.com/]*					
SSC	Thunder Bolt [http:/www.spectrumthunderbolt.com/]*			VLF		
SSL	Aircraft Total Lightning Advisory System [http:/ae.atmos.uah.edu/AE/ams_1999b.html]	ATLAS		VHF		(55)
SSL	BLACKBEARD VHF Sensor on ALEXIS satellite [http:/nis-www.lanl.gov/nis-projects/blackbeard/]	BLACK-BEARD		VHF		(52,53)
SSL	Defense Meteorological Satellite Program (DMSP) Operational Linescan System (OLS) [http:/thunder.msfc.nasa.gov/ols/]	DMSP/ OLS		Light		(43,44)
SSL	Fast On-orbit Recording of Transient Events satellite [http:/forte.lanl.gov/]	FORTÉ		VHF/ Light		(49–51)
SSL	Great Plains-1 [http:/bub2.met.psu.edu/default.htm]	GP-1		VLF/LF		(56,57)

(*Continued*)

Table 1. (*Continued*)

Instrument/Network and Class of System *NL: Network Locator; NM: Network Mapper;* *SSC: Single-Station Counter; SSL: Single-Station Locator;* *WD: Warning Device* **Manufacturer's literature for commercial product*		Acronym	Lightning Property			Reference
			Electric Field Only (*E*-field)	Electromagnetic Radiation	Other	
SSL	Lightning Detection and Ranging System [http:/www.chinatech.com/lightning.html]*	LDARS		VLF/LF		
SSL	Lightning Imaging Sensor [http:/thunder.msfc.nasa.gov/lis/] [http:/ghrc.msfc.nasa.gov/uso/readme/lis.html]	LIS		Light		(31,34,48,163)
SSL	Optical Transient Detector [http:/thunder.msfc.nasa.gov/otd/] [http:/ghrc.msfc.nasa.gov/uso/readme/otd.html]	OTD		Light		(45–47)
SSL	Long-Range Shipborne Detector			ELF/VLF		(58)
SSL	SOnic Lightning LOcation system [http:/www.nasatech.com/Briefs/July00/KSC11992.html]	SOLLO	Electrostatic		Sound	(59)
SSL	Storm Scope® [http:/www.gis.net/~aviaelec/stormscope.htm]*			VLF/LF		
SSL	StormTracker Lightning Detection System [http:/www.boltek.com/]*			VLF/LF		
SSL	Strike Finder® [http:/www.strikefinder.com/stk.html]*			VLF/LF		
SSL	Weather radars				Channel Ionization	(60)
	Field Mill Networks					
WD	The Launch Pad Lightning Warning System at CCAFS/KSC [http:/www.tstorm.com/lplws.html] [http:/ghrc.msfc.nasa.gov/uso/readme/kscmill.html]	LPLWS	ELF			(35,37,38)
WD	Electric Field Measurement System, Wallops Island [http:/www.tstorm.com/wffefms.html]	EFMS	ELF/ULF			
WD	Lightning Warning System (SAFe) [http:/www.eurostorm.com/]*	SAFe	Electrostatic			
WD	SAFe Lightning Warning Network *E*-field network at European Space Center (CSG), Kourou, French Guiana	SAFe	Electrostatic			
WD	Thor Guard [http:/www.thorguard.com/default1.asp]*		Electrostatic			

Table 2. Approach, Application, Detected Stroke Type, Maximum Range, Accuracy, and Detection Efficiency for the Lightning Sensing Systems and Classes Listed in Table 1. The Data Presented are From Peer-Reviewed Work Where Possible. As in Table 1, If Marked with an Asterisk (*), then the Information for a Commercial Product is Uncritically Provided Based on Documentation Given by the Manufacturer

Lightning System (class)	Approach ITF: interferometry MDF: magnetic direction finding TOA: time of arrival	Maximum Range	Location Accuracy	Detection Efficiency	Applications	Stroke Type CG: cloud–ground IC: intracloud
ALDIS (NL)	Locations given by MDF triangulation of VLF/LF peak signals from a network of 8 receivers	Network coverage over Austria; mean range 400 km for each DF sensor	<1 km owing to sensor spacing	>90% of flashes, but appreciably <90% for individual CG strokes	Operational weather forecasting support; locations of potential lightning damage	CG and IC distinguished; only CG reported, with individual stroke polarity within a flash identified
ALDS (NL)	Locations given by MDF triangulation of VLF/LF peak signals from a network of receivers	Network coverage over Nevada; mean range 400 km for each DF sensor	Varies from 1.5% of the range at 100 km to 3% at 400 km	~85% of CG strokes	Possible forest fire initiation	CG and IC distinguished; only CG reported, with individual stroke polarity within a flash identified
ATD (NL)	Locations given by TOA differences from 5 receivers in UK, 1 in Gibraltar, and 1 in Cyprus. Analysis of complete 2 to 23 kHz (VLF) waveforms performed	Several 1,000 km	From <2 km in UK to 15 km at service area limit; error increases with range	>90% of storm areas; maximum of 400 flashes/h	Locate thunderstorms over ranges of several 1,000 km	CG, primarily
ATLAS (SSL)	Aircraft-mounted VHF MDF receiver. Senses radiation given by initial breakdown of leader step into virgin air. Radiated power assumed to be same in each leader, so range given by attenuation of radiated power	Not provided	Not provided	Not provided	Identification of the origin of lightning flashes in cloud, a hazard to aviation	Initial leader step
BLACK-BEARD (SSL)	ALEXIS-satellite-mounted VHF (28 to 95 or 108 to 166 MHz) detector. Operational 1993–1997	Instrument on polar-orbiting satellite	Continental-sized footprint	Not provided	Study of transionospheric pulse pairs (TIPPs)	Paired pulses in IC
CGLSS (NL)	Locations given by MDF triangulation using VLF/LF peak signals from network of 6 IMPACT sensors. Time-domain waveform analysis used to separate CG from IC strokes	~60 km	~250 m	~98% of flashes; up to 74 strokes per minute	Rocket launch support and personnel safety at Cape Canaveral Air Force Station/ Kennedy Space Center (CCAFS/KSC)	CG and IC distinguished; only CG reported

(*Continued*)

Table 2. (*Continued*)

Lightning System (class)	Approach ITF: interferometry MDF: magnetic direction finding TOA: time of arrival	Maximum Range	Location Accuracy	Detection Efficiency	Applications	Stroke Type CG: cloud–ground IC: intracloud
CGR3 (SSC)	Detects a lightning discharge if electric field changes sufficiently; omnidirectional antenna used	Effective area (in which all flashes that occurred are detected) depends on height of cloud top, latitude, and stroke polarity; varies from 315 to 442 km^2	Locations not given	Essentially 100% within effective area	Determination of lightning climatology, as given by density of lightning strikes	CG and IC, combined
CIGRE (SSC)	Detects a lightning discharge if electric field changes sufficiently; omnidirectional 500-Hz or 10-kHz antenna used	Effective range (in which all flashes that occurred are detected) is 20 to 30 km; maximum range up to 3 times greater	Locations not given	For 500-Hz counter, ~80% within 40 km; for 10-kHz counter, 95% within 20 km	Determination of lightning climatology, as given by density of lightning strikes	CG and IC, combined
CLDN (NL)	Locations given by sensors using MDF triangulation and TOA differences of VLF/LF peak signals. Time-domain waveform analysis used to separate CG from IC strokes. Network has 55 LPATS-IV TOA *E*-field sensors and 26 IMPACT ESP hybrid MDF/TOA sensors	Coverage over 9 million km^2 area of Canada; each IMPACT ESP sensor detects in 90 to 1,000 km range	Median position accuracy ~500 m over network. Bearing error for individual IMPACT ESP sensors <1°	Flash detection efficiency 80% to 90% for peak currents >5 kA. Subsequent stroke detection efficiency 50%. Up to 15 strokes per flash reported. More than 98% of strokes correctly characterized as CG or IC. Stroke polarity assigned OK for 99.5% of strokes out to 600 km. Some indication that positive return strokes < 10 kA are misidentified IC strokes	Operational weather forecasting support; locations of potential lightning damage; lightning warning for outdoor events. Added to NLDN to form LRLDN and NALDN	CG and IC distinguished; individual CG stroke polarity within a flash identified. LPATS-IV sensor can distinguish inter- and intracloud strokes

(*Continued*)

Table 2. (*Continued*)

Lightning System (class)	Approach ITF: interferometry MDF: magnetic direction finding TOA: time of arrival	Maximum Range	Location Accuracy	Detection Efficiency	Applications	Stroke Type CG: cloud–ground IC: intracloud
DMSP/OLS (SSL)	Optical instrument mounted on DMSP polar-orbiting satellites that identifies a lightning flash from optical signature. Data available from 1973, 1977, 1978 and 1985–1996	Footprint on the earth of 700 to 1,360 km, with global coverage	Up to 100 km error due to manual digitization, navigation, and width of lightning streak; imagery of 2.7 km resolution	Often < 2% (dawn, dusk and local midnight only)	Research studies of NO_x budgets and transport, as lightning is a significant natural source of NO_x, and of the natural variability of thunderstorms	All types, combined
EDOT (NL)	Measures electric field change in 300-Hz to 300-KHz (ULF through LF) band. When above trigger level, locations given by TOA differences. Networks of 5 sensors in New Mexico/Texas, 5 in Florida, 1 in Nebraska	Few hundred km, within network coverage area	Within a few km for storms within the array	Near 100% for storms in network coverage area	Ground truth for lightning detections by FORTÉ satellite	CG and IC, combined
EFMS (WD)	Network of 10 to 16 field mills sampling from dc to 10 Hz and electric field change meters sampling from 10 Hz to 2 kHz. Their time-varying values over the region alert forecaster to possible lightning hazard	Wallops Island, Virginia, launch pad area	Similar to LPLWS	Similar to LPLWS	Provides detection of flashes or warning of possible triggering of lightning during rocket launches	All types; not separated
ESID/ALS (SSC)*	Optical and VLF/LF radiation received within milliseconds by single omnidirectional instrument triggers lightning detection; range within 1 of 3 intervals given by comparison of optical and radio signals	Nominally, up to 50 km; up to 80 km for intense CG or IC strokes	Range resolved into 0 to 8 km, 8 to 16 km, and 16 to 50 km intervals	Reports more than local human observers. Near 100% flashes reported within 16 km when at least 3 discharges; near 90% for CG within 16 km. Thunderstorm reported by ALS if 2 strokes in 15-minute period; thunderstorm detection has < 0.5% false alarm rate	ESID used for personnel safety and equipment protection; ALS part of Automated Surface Observing System (ASOS) in United States	CG and IC, distinguished
FORTÉ (SSL)	Satellite-mounted VHF and optical sensors; simultaneous observations by these sensors best identifies a lightning flash occurrence. Operational 1997 to date	Instrument on polar-orbiting satellite for nuclear test ban treaty verification	Detection within field of view of diameter 1,200 km; samples an area for ~15 min	Not yet determined, as correlating optical and VHF signals is made difficult by light scattering in cloud	Development of lightning climatologies and study of lightning processes in clouds	Stepped leaders, dart leaders, as well as CG and IC, distinguished

(*Continued*)

Table 2. (*Continued*)

Lightning System (class)	Approach ITF: interferometry MDF: magnetic direction finding TOA: time of arrival	Maximum Range	Location Accuracy	Detection Efficiency	Applications	Stroke Type CG: cloud–ground IC: intracloud
GP-1 (SSL)	Single VLF/LF (1 to 400 kHz) receiver. Bearing via MDF and range via application of signal propagation model relating peak of received magnetic wave intensity to distance. Time-domain waveform analysis used to separate CG from IC strokes	Location up to 400 km, depending on bearing; bearing up to 500 km	Best ranging between 100 and 300 km from receiver. Mean absolute location error of ~50 km	Not provided	Detection of coherent lines and regions of thunderstorms and electrically active convective cells	CG and IC distinguished; only CG used
GPATS (NL)*	Network of 6 VLF/LF (2 kHz-500 kHz) LPATS-III *E*-field sensors in Australia that locate lightning via TOA techniques	Covers the 1,000,000 km^2 area of Australia	Mean location error better than 200 m	Similar to other LPATS TOA systems: >95% of CG strokes detected within a specified range	Operational weather forecasting support; locations of potential lightning damage; lightning warning for outdoor events	CG and IC distinguished
Guardian Angel/ Sky Scan™ (SSC)*	Single portable omnidirectional receiver that estimates range based on received signal at pairs of VLF/LF frequencies	Up to 70 km	Range resolved into 0 to 5 km, 5 to 15 km, 15 to 35 km and 35 to 70 km intervals	~97% in the 0 to 5 km bin	Warning of approaching thunderstorms for outdoor events	CG and IC, combined
LDAR (NM)	Locations given by TOA differences of discrete broadband VHF (63 to 69 MHz) peak signals from a network of 7 receivers. Channel mapped using tips of negative leader steps above 2 km; ground strike point given by CGLSS	Up to 300 km in horizontal, but from 2 km to 16 km in vertical	Median error of 3-D locations 50 to 100 m within the sensor baseline of 10 km when stepped leader originates >3 km, increasing steadily to 900 m at 40 km. Standard deviations of source locations increase from 4 km at 60 km range to 30 km at 260 km range, and of flash locations are ~1 km at 50 km range and remain within ±7 km out to 250 km range. Maximum of 10^4 samples per second	3-D locations of up to 100 pulses in each flash. At < 25 km range, >99%. In a separate study, estimates of flash detection efficiency >90% in 90–100 km range, but < 25% at 200 km range	Rocket launch support and personnel safety at Cape Canaveral Air Force Station/Kennedy Space Center (CCAFS/KSC) in Florida. Some warning of CG strokes possible, as detects precursor IC strokes	Negative leader steps produced in impulsive VHF radiation; cannot detect positive leaders. CG and IC stroke locations inferred from channel orientation

(*Continued*)

Table 2. (*Continued*)

Lightning System (class)	Approach ITF: interferometry MDF: magnetic direction finding TOA: time of arrival	Maximum Range	Location Accuracy	Detection Efficiency	Applications	Stroke Type CG: cloud–ground IC: intracloud
LDARS (SSL)*	Single VLF/LF receiver that locates radiation source using phase and time differences of the received electromagnetic waves	Up to 250 km for CG, 60 km for IC	Bearing within 1° and range within 1 km	Not provided	Lightning warning for aviation and other regional transportation authorities	CG and IC distinguished, with individual stroke polarity within a flash identified
Lightning Alert® Detector (SSC)*	Single omnidirectional VLF receiver triggers on signal above a threshold; both portable and fixed models available	Up to 600 km; range increases with antenna height above ground	Range estimates reflected in rate at which warning light flashes	Not provided; will vary with antenna siting	Warning of approaching thunderstorms	CG and IC, combined
LIS (SSL)	Satellite-mounted optical instrument on Tropical Rainfall Measuring Mission (TRMM) satellite that identifies a lightning flash from optical signature. Operational from 1997 to date	Satellite covers only half the earth centered on equator	Detection within field of view that spans a 600 × 600 km area; samples one storm ~90 s. Spatial resolution 4 to 7 km. Location accuracy better than 10 km.	~90% owing to excellent filtering of sun glint and reflected visible light	Lightning climatology; comparison of cloud lightning signature with that given by ground-based sensors	All types, combined
LMA/LMS (NM)	Locations given by TOA differences of VHF (60 to 66 MHz) peak signals from a deployable network of 10 sensors over a 60 km diameter area	Up to 100 km in horizontal, but above 2 km in vertical	50 to 100 m rms errors over the 60 km diameter area of network	Near 100%, as with LDAR	Field programs studying lightning characteristics	Leader steps; CG and IC strokes inferred from channel orientation
Long-Range Ship-borne Locator (SSL)	Single-station system that finds range using spectral components of ELF (< 1 kHz) and VLF (~1 to 10 kHz) signals and finds bearing using Poynting vector based on covariances between temporally varying horizontal magnetic and vertical electric fields	1,000 to 3,000 km	5% uncertainty for range, few degrees for bearing	Not provided	Lightning detection from a ship for thunderstorm climatological studies	CG and IC; not separated

(*Continued*)

Table 2. (*Continued*)

Lightning System (class)	Approach ITF: interferometry MDF: magnetic direction finding TOA: time of arrival	Maximum Range	Location Accuracy	Detection Efficiency	Applications	Stroke Type CG: cloud–ground IC: intracloud
LPATS (NL)	Network of VLF/LF (2 kHz to 500 kHz) *E*-field sensors that locate lightning using TOA techniques. Part of NALDN; portion in Canada (CLDN) uses series IV sensors that distinguish intercloud and intracloud strokes. Series III sensors used in NLDN and in GPATS	Coverage over network; each sensor detects up to several hundred km	Median accuracy better than 500 m	>95% CG detection within a specified range	Operational weather forecasting support; locations of potential lightning damage; lightning warning for outdoor events	CG and IC distinguished. LPATS-IV sensor can distinguish inter- and intracloud strokes
LPLWS (WD)	Network of 31 field mills over a 20 × 20 km area sampling electric field potential gradient at 50 Hz. Contours of gradient in coverage area alerts forecaster to possible lightning hazard. Flash inferred if sufficient *E*-field gradient	~20 km outside network	Median charge location errors of 2 to 20 km	Flash detection ~90%	Provides detection of flashes or warning of possible triggering of lightning during rocket launches at CCAFS/KSC	All types; not separated
LRLDN (NL)	Long-range network operated by GAI. Uses data from the combined MDF/TOA NALDN sensors, reprocessed to use sky-wave signals for lightning detection over the North Atlantic and Pacific Oceans	Up to 2,000 to 4,000 km. Location accuracy and detection efficiency studied over 1,200 to 1,600 km range	Median location accuracy ~5 km when lightning is located between subsets of sensors and propagation paths between 1,200 and 1,600 km. Limited to 16 to 32 km over 2,000 to 4,000 km range	Only lightning flashes with peak currents >30 kA detected. Overall peak detection efficiency in a test region over United States 10–25% during the night and a few percent during the day. For peak currents >45 kA, detection efficiency in a 24-hour period approaches 50%	Lightning detection over oceans in support of aviation	CG and IC distinguished; only CG reported, with individual stroke polarity within a flash identified
NALDN/ NLDN (NL)*	Locations given by sensors using MDF triangulation and TOA differences of VLF/LF peak signals. Time-domain waveform analysis is used to separate CG from IC strokes. NLDN has 59 LPATS-III TOA *E*-field sensors and 47 IMPACT hybrid MDF/TOA sensors. CLDN has 55 LPATS-IV TOA *E*-field sensors and 26 IMPACT ESP hybrid MDF/TOA sensors	Coverage over United States (NLDN) & Canada (CLDN); each IMPACT ESP sensor detects in 90 to 1,000 km range	Median position accuracy ~500 m over network. Bearing error for individual IMPACT ESP sensors < 1 °	Flash detection efficiency 80% to 90% for peak currents >5 kA. Subsequent stroke detection efficiency 50%. Up to 15 strokes per flash reported. More than 98% of strokes correctly characterized as CG or IC. Stroke polarity assigned OK for 99.5% of strokes out to 600 km. Some indication that positive return strokes < 10 kA are misidentified IC strokes	Operational weather forecasting support; locations of potential lightning damage; lightning warning for outdoor events	CG and IC distinguished; only CG reported over network, with individual stroke polarity within a flash identified. CLDN can distinguish intercloud and intracloud strokes using LPATS-IV sensors

(*Continued*)

Table 2. (*Continued*)

Lightning System (class)	Approach ITF: interferometry MDF: magnetic direction finding TOA: time of arrival	Maximum Range	Location Accuracy	Detection Efficiency	Applications	Stroke Type CG: cloud–ground IC: intracloud
ONERA-3D (NM/NL)*	ITF triangulation of narrowband VHF (at a selectable center frequency in the 110 to 118 MHz band in 1 MHz bandwidths) signals from pairs of receivers provides phase differences and so direction to source; triangulation using these radials allows 3-D mapping of channel. Stroke type and ground strike points given by LF radiation techniques	Nominal maximum range ~100 km. Plan view depiction at large ranges, 3-D mapping at closer ranges	Varies with receiver spacing: from 500 m rms for 30-km spacing to 750 m rms for 120 km spacing. Temporal resolution 10 μs. Maximum of 4,000 samples per second. Bearing error $< 0.1^\circ$ for each interferometer in network	Detection rate ~95% within 200 km range for low-elevation sources	Originally, launch support and personnel safety for European space program, followed by many commercial and weather service users. Some warning of CG strokes possible, as detects precursor IC strokes	All stages of breakdown in lightning process, especially negative leader steps. Best detects continuous VHF noise associated with *K*-processes or dart leaders. CG and IC stroke type and polarity inferred from channel orientation and LF wave analysis
OTD (SSL)	Satellite-mounted optical instrument on MicroLab-1 satellite that identifies lightning flash from optical signature. Operational 1995–2000	Low-earth-orbiting satellite in a 70°-inclination orbit covers any spot on earth twice per day	Detection within field of view that spans a $1,300 \times 1,300$ km area; 8 km nadir resolution. Ground range errors typically 20 to 40 km, with median of 50 km	Cloud–ground detection efficiency 40% to 69% based on comparison with NLDN observations	Lightning climatology; global flash rate found to be 40 strokes per sec; lightning mostly over land	CG and IC, combined
Radar (SSL)	Scattering of radar beam off ionized lightning channel; from nonscanning or dual polarization radar	Nearby radar up to a few 100 km	Resolution depends on sampling strategy	Varies with beam width and amount of precipitation	Case studies relating occurrence of flash to rain gush	All types, combined

(*Continued*)

Table 2. (*Continued*)

Lightning System (class)	Approach ITF: interferometry MDF: magnetic direction finding TOA: time of arrival	Maximum Range	Location Accuracy	Detection Efficiency	Applications	Stroke Type CG: cloud–ground IC: intracloud
SAFe (WD)*	Analysis of cumulonimbus electrostatic field evolutions for estimation of lightning hazard	15 km for cumulonimbus without lightning; 30 km for cumulonimbus with lightning	No location provided	Close to 100% up to 10 km	Lightning hazard warning	Static field; both IC and CG
SAFIR (NM/NL)*	ITF triangulation of narrowband VHF (at a selectable center frequency in the 110 to 118 MHz band in 1 MHz bandwidths) signals from pairs of receivers provide phase differences and so direction to source; triangulation using these radials allows 2-D or 3-D mapping of channel. Stroke type and ground strike points given by LF radiation techniques	Nominal maximum range ~250 km, but up to ~400 km, depending on sensor baseline and elevation. Plan view depiction at large ranges; 3-D mapping at closer ranges	Varies with receiver spacing: from 500 m rms for 30-km spacing to 750 m rms for 120-km spacing. Temporal resolution 10 μs. Maximum of 4,000 samples per second. Bearing error $< 0.1^\circ$ for each interferometer in network	Detection rate >90% within ~250 km range for all lightning types, or within 400 km for high-elevation sources	Originally, launch support and personnel safety for European space program in French Guiana, followed by many commercial and weather service users (France, Belgium, Netherlands, Germany, Hungary, Slovakia, Poland, Italy, Singapore, Japan). Some warning of CG strokes possible, as detects precursor IC strokes	All stages of breakdown in lightning process, especially negative leader steps. Best detects continuous VHF noise associated with *K*-processes or dart leaders. CG and IC stroke type and polarity inferred from channel orientation and LF wave analysis
SOLLO (SSL)	ITF using five acoustic sensors provides 3-D direction to source. TOA differences between electric and acoustic waves provide range to source; one electric field sensor.	Up to 1.6 km	Median position accuracy <5 m for strokes within 1 km	Not provided	Provides strike point on launch pad at CCAFS/KSC to determine if inspections for damage needed	CG and IC strokes producing thunder detected; CG prime interest

(*Continued*)

Table 2. (*Continued*)

Lightning System (class)	Approach ITF: interferometry MDF: magnetic direction finding TOA: time of arrival	Maximum Range	Location Accuracy	Detection Efficiency	Applications	Stroke Type CG: cloud–ground IC: intracloud
Storm Scope® (SSL)*	Aircraft-mounted sensor that finds bearing via MDF and range using a function of received electromagnetic signal in a narrow frequency band and the correlations between the electric and magnetic field signals.	Up to 400 km	Not provided	Not provided	Identification of lightning flashes in cloud, a hazard to aviation	CG and IC, combined
Storm Tracker (SSL)*	Single VLF/LF receiver. Bearing via MDF and range via application of signal propagation model relating received signal intensity to distance	Up to 500 km	Not provided	Not provided	Warning of approaching thunderstorms	CG
StrikeAlert Lightning Detector (SSC)*	Single portable omnidirectional receiver triggers on VLF/LF radiation above a threshold; estimates whether storm is approaching or receding	Up to 70 km	Range resolved into 0 to 10 km, 10 to 20 km, 20 to 40 km and 35 to 70 km intervals	Not provided	Warning of approaching thunderstorms for outdoor events	CG and IC, combined
Strike Finder® (SSL)*	Aircraft-mounted sensor that finds bearing via MDF and range using a function of received electromagnetic signal intensity	Up to 400 km	Not provided	Not provided	Identification of lightning flashes in cloud, a hazard to aviation	CG and IC, combined

(*Continued*)

Table 2. (*Continued*)

Lightning System (class)	Approach ITF: interferometry MDF: magnetic direction finding TOA: time of arrival	Maximum Range	Location Accuracy	Detection Efficiency	Applications	Stroke Type CG: cloud–ground IC: intracloud
Thor Guard (WD)*	Electric field change meter that uses the character of these changes to estimate lightning hazard probability	Up to 8 to 20 minutes warning; proximity sensor	Not provided	Not provided	Warning of lightning risk for outdoor events	CG and IC, combined
Thunder Bolt Lightning Detector (SSC)*	Single portable omnidirectional receiver triggers on VLF/LF radiation above a threshold; estimates whether storm is approaching or receding	Up to 100 km	Within 5% of range at 50 km and within 10% of range at 100 km	~100% flash detection if storm occurs within 5 minutes; unit displays strike probability	Warning of approaching thunderstorms for outdoor events	CG and IC, combined
TREMBLE (NM)*	Locations given by TOA differences from a network of 3 to 6 microphones; event timing triggered by radio detection of lightning signal. 3-D mapping of lightning channel provided	Within 15 km radius of each recording station	Not provided; will depend on microphone spacing	Not provided; will depend on number of simultaneous detections by at least four microphones	Visualization of lightning channel for thunderstorm research	Leader steps; CG and IC strokes inferred from channel orientation

the United States Department of Energy (DOE); and the BLACKBEARD instrument on the Array of Low Energy X-ray Imaging Sensors (ALEXIS) satellite (52,53) that provided data from 1993 to 1997 and is operated by LANL. Limited ranging may be accomplished by using functions of the electromagnetic wave radiated by a lightning stroke, as done by a host of single-station detectors (54–58); by combining measurements of the electromagnetic signal with those of the acoustic signal (thunder), as done, for example, in the SOLLO system currently under development at CCAFS/KSC (59); or by using the fact that the heated and highly ionized lightning channel reflects a radar signal (60). As summarized in Table 2, estimates of the range to the lightning that are given by a particular system may vary from local to global according to both the particular aspect of the radio or optical signal that is detected and the intended application of the information.

Lightning sensing systems have important applications in meteorology. Certainly, these systems can help to detect and forecast the arrival of thunderstorms by monitoring their positions, motions, and intensities. Such systems are also important for lightning safety and for protecting natural resources in a diverse set of real-time applications, including fire and personnel safety, aircraft routing, missile launch decisions, power grid operation, and weather forecasting (1,11,13). Lightning is a significant threat to people (Fig. 1) but far too often is underrated. Lightning is ranked the #2 source of direct weather deaths in the United States and kills an average of about 100 people each year — more than hurricanes and tornadoes combined. Lightning inflicts lifelong severe injuries on many more, about 750 people each year in the United States. Lightning also causes about $5 billion in damage in the United States each year (61–63). More information on lightning safety is available at the Lightning Safety web page of the 45th Weather Squadron, USAF (64). Lightning sensing systems, especially those that detect intracloud lightning, are being investigated as tools for predicting severe thunderstorms, including tornadoes, microbursts, and hail (47,65–67). Lightning data are also being investigated as possible input for small-scale numerical weather forecast models to supplement the usual observations of pressure, temperature, humidity, and winds.



Figure 1. Cloud–ground lightning stroke near Lane Stadium in Blacksburg, Virginia, just before the start of the Black Coaches Association Bowl on 27 August 2000. More than 56,000 fans were in the stadium at the time. They were evacuated and the game was postponed to a later date (Photograph by Eric Brady reproduced with the permission of *The Roanoke Times* in Roanoke, Virginia). See color insert.

This article is not an exhaustive survey of all real-time and research lightning location systems, but it provides a comprehensive overview of the topic. It is split into two major parts; the first reviews the many important concepts and defines the terminology used to describe them, and the second covers the different major types of operational lightning location systems by describing an example of each in detail. These sections refer to one another but can be read separately, as many readers may wish to learn about certain aspects of the lightning location problem quickly and so may not want to read the entire article. This approach to presenting the material has led to redundancy in some cases, but the repeated information is retained to benefit the reader.

There are many excellent books that describe lightning characteristics and lightning location methods in greater detail than presented here (22,68–77). Many of the lightning facts herein are documented thoroughly in these books and in the articles they reference. Other important sources cited here include papers from conference preprint volumes as supplements to the numerous papers published in the reviewed literature. Finally, the methods and detection systems reviewed in this article are also described on various World Wide Web (www) sites. Table 1 lists the addresses for many of these sites, enclosed in brackets, recognizing that www addresses change frequently or that some web sites are not kept as current as users would like.

CLASSES OF LIGHTNING SENSING SYSTEMS

Real-time lightning sensing systems fall into four broad classes: locators, direction finders, flash counters, and warning systems (Tables 1 and 2). Lightning locating systems may themselves be separated into two major types: *network locators* (NL) that provide a two-dimensional *plan view* that gives the coordinates at the earth's surface at which cloud–ground strokes are occurring and *network mappers* (NM) that provide a *mapping view* that gives the spatial coordinates within the atmosphere for the entire lightning channel in three dimensions. Depending on the selected radio receiver *bandwidth*, or breadth of accepted frequency content, plan view systems can provide coverage varying from local to global (Table 2). In contrast, ground-based mapping systems are restricted to local coverage, because they depend on line-of-sight reception of very high frequency (VHF, 30 MHz to 300 MHz) emitted electromagnetic radiation.

Lightning Direction Finders

Lightning direction finders (DF), the oldest type of electronic lightning sensing systems (1), give only the bearing from which the lightning signals originate. DF systems use magnetic direction finding (MDF), time of arrival (TOA), or interferometric (ITF) techniques. When used in concert, DF systems form networks of lightning locators or lightning mappers (see later).

Magnetic Direction Finding (MDF). As illustrated in Fig. 2, the *magnetic direction finding* technique finds the tangent of the bearing angle θ that is given by the ratio of the peaks *ew* and *ns* of the two horizontal components **EW** and **NS** of the magnetic field (*B-field*) signal. These components are received by two vertical and orthogonal, *crossed-loop antennas*, or *B-field antennas* (1;5, p. 370; 6,11), that are typically aligned in the east–west (EW) and north–south (NS) planes. Because the value of θ determined this way from the *B*-field has an inherent bearing ambiguity of π radians (180°), an electric field (*E-field*) sense antenna, often called simply an *E-field antenna*, is included with the crossed-loop antennas to determine stroke polarity (see later); by knowing the polarity given by the *E*-field measurement, one can easily determine whether the signal arrived from quadrants 1 or 4 or from quadrants 2 or 3 (11).

Generally, MDF systems have ranges up to a few hundred kilometers, and, depending on the passband of the receiver chosen, can produce bearing errors ranging from 1 to more than 10°. Large bearing errors are found predominantly in narrowband receiving systems (78–80;81, pp. 184–185), which were the type first used for radio direction finding in the 1920s (2). Because these systems detect the signal from all parts of the lightning channel, which can branch considerably above ground (74, p. 357), the bearing errors are related to multipath propagation caused by nonvertical channel orientation (5, p. 370) that lead to sky wave reflections off the ionosphere (81, pp. 184–185) (see later). In contrast, wideband receiving systems that resolve the initial peak *b* of the return-stroke magnetic field and that use time-domain waveform analysis to separate cloud–ground from intracloud discharges (11) (see later), have accuracies of the order of 1° because they use only the single ground-wave signal radiated from the vertically oriented cloud–ground stroke channel near where it contacts the ground (7,82).

Time of Arrival (TOA). The *time of arrival* (TOA) method for locating lightning is based on the fact that, if the source is on or near the earth's surface, then the difference in times at which a signal is detected by a pair of stations, when multiplied by the speed of light, is described spatially

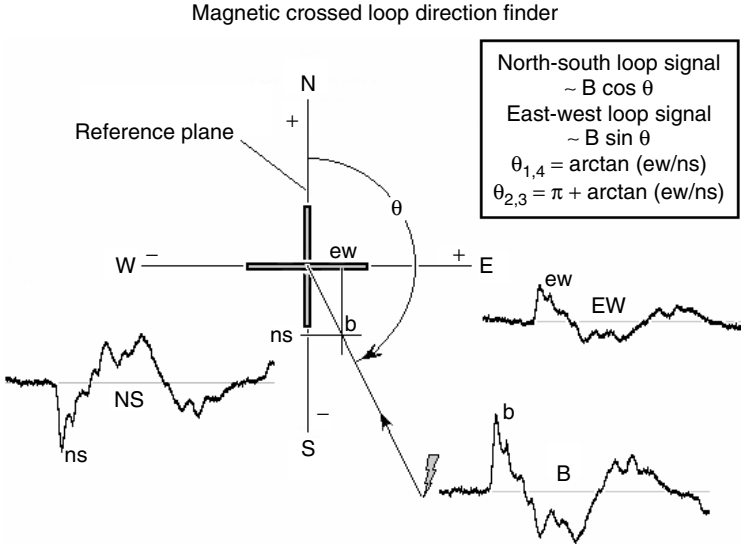


Figure 2. Plan view of the crossed-loop antenna array used for magnetic direction finding (MDF). The radiation pattern of each loop provides a signal level proportional to the cosine of the bearing θ to the source indicated by the lightning bolt. If the reference plane is assigned to one loop, such as the north–south loop that is appropriate for navigational bearings used by meteorologists, then the east–west loop signal level is proportional to the sine of the bearing relative to the reference plane. Because the loop signal voltage is proportional to the rate of change of the intercepted magnetic field, integrators in each loop-processing channel provide the level of the loop signal. **NS** and **EW** provide the corresponding loop signal levels that are used to calculate the incoming time-domain magnetic field **B** using standard trigonometry. Unless the electromagnetic field polarity is known, the signal bearing is ambiguous. In networked systems, however, only the bearings given by intersecting radials are the correct ones (Fig. 3). In single-station locators (see later), the field polarity can be found from the polarity of the initial peak of the voltage field by using a separate voltage field (*E*-field) sense antenna. The initial peak levels, *ns* and *ew*, of each vector are used for bearing estimation. The navigational bearing to the source is found as the arctangent of the ratio of the pair of peak signal levels, *ew*/*ns*. The stroke waveforms from Fig. 6, south south-east (SSE) of Lawrence, Kansas are used in this example. With the value of θ increasing clockwise from due north, the NE quadrant is 1, the SE quadrant is 2, the SW quadrant is 3 and the NW quadrant is 4. As the SE quadrant is quadrant 2 because $ew > 0$ and $ns < 0$, π radians (180°) must be added to the calculated arctangent to give a bearing of 152°. Finally, the peak level *b* of **B** that can be used for range estimation via a signal propagation model is given by $b = ns\cos(\theta) + ew\sin(\theta)$, once the value of θ is calculated.

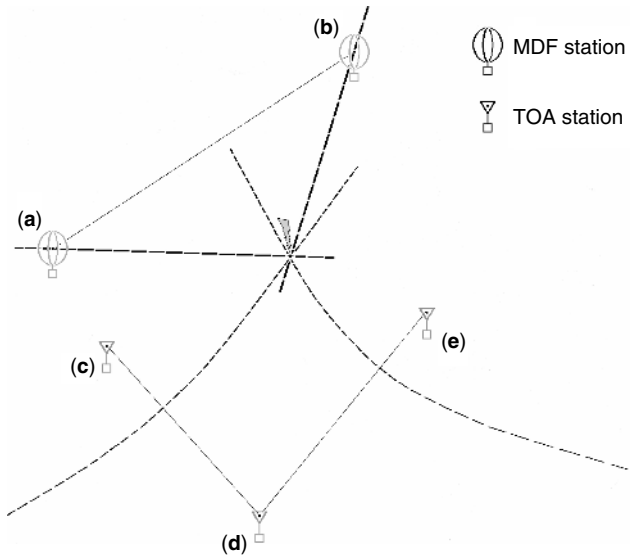


Figure 3. Illustration of magnetic direction finding (MDF) triangulation and time-of-arrival (TOA) methods for locating where a cloud–ground lightning stroke has occurred, as given by the lightning bolt in the center. Stations A and B to the west and north of the strike location use the MDF technique to define the radials, given by the dashed lines, along which the electromagnetic signal from the lightning travels (Fig. 2). The TOA method applied at each station pair (C, D) and (D, E) to the south of the strike location is based on the fact that a hyperbola defines the locus of points from which a ground-based signal could have originated when the difference in the arrival times at a station pair is the same; this fact relies on the good assumption that radio waves emanating from lightning propagate at the speed of light. Each station is at a focus of a hyperbola. Because the pair of foci defines two possible hyperbolas, the station that receives the signal first gives the particular hyperbola used. Here, stations C and E detect the lightning signal before station D, giving hyperbolas closer to C and E. In both methods, the intersection of the radials or the hyperbolas provides an estimate of the strike location; combining the two methods via a least-squares optimization approach, as done in a hybrid lightning location network (see later), provides the best estimate.

by a two-dimensional hyperbola, with the stations at its foci (1,83). A hyperbola describes the locus of points whose difference in distances from the foci is constant, as for the station pairs (C, D) and (D, E) in Fig. 3 south of the lightning strike point. The sensors in the original TOA network in the United States, operated by Atmospheric Research Systems, Inc. (ARSI) (13,14), as well as those in other networks currently operated throughout the globe by Global Position and Tracking Systems® Pty. Ltd. (GPATS) (21), are commonly called Lightning Position and Tracking Systems (LPATS) sensors (Table 2).

Interferometry (ITF). An *interferometer* is an example of a DF system employing antennas that combine both the magnetic and voltage component of the incoming field (Fig. 4) (84). *Interferometry* (ITF) measures the direction angle θ of an incoming signal using the difference in phase ϕ between the radio signals received by closely spaced pairs of receivers. When a pair of receivers is spaced at a distance D that is equal to half the wavelength λ of the sampled radiation, ϕ is given simply by $\pi \cos\theta$ (Fig. 4). A direction angle θ perpendicular to the baseline of the interferometer, $\theta = \pi/2$, is indicated by a phase $\phi = 0$. A phase of $\phi = \pi$ occurs when $\theta = 0$ and a phase of $\phi = -\pi$ occurs when $\theta = \pi$. Because of the unavoidable consequence of having to calculate a bearing from the phase using the arccosine function, a signal arriving from a bearing in the upper half plane of the figure $(+\theta)$ produces the same phase as a signal arriving from the lower half plane $(-\theta)$. Additional bearing ambiguities occur when the sensor spacing $D > \lambda/2$ (see later) (22, p. 155). Because interferometers are used with triangulation in network systems, however, this uncertainty does not pose a problem when the interferometers are distributed wisely or use different frequencies. Depending on the separation between the receiver pairs, systems such as SAFIR (see later) that determine bearing using interferometry can operate in either plan view NL or three-dimensional NM modes (22, p. 155; 24).

Single-Station Systems

Lightning warning systems (WD) measure the thunderstorm electric field or lightning precursors within a few tens of kilometers of the instrument. An example of such a system is the Thor Guard instrument that warns of imminent lightning when the magnitudes of the temporal variations in the electric field near the earth's surface exceed their fair-weather values. Vaisala Dimensions SA also manufactures an early lightning-warning system called SAFe that uses electric field measurement; a network of SAFe units is operated at the European Space Center in Kourou, French Guiana, to aid launch decisions.

Single-station flash counters or *proximity sensors* (SSC) indicate that lightning is occurring within a few tens or hundreds of kilometers of the receiver but are *omnidirectional* because they provide no bearing to the source. Some of these instruments may provide limited distance information, usually divided into three or four range-bins (Table 2). For example, both the solar-powered Electrical Storm Identification Device (ESID) (see later) and the similar Thunderstorm Sensor Series (TSS924) device, which is also called the ASOS Lightning Sensor (ALS) because it is the instrument used on some Automated Surface Observing System (ASOS) weather stations operated by the United States National Weather Service (NWS), are proximity sensors that report the occurrence of lightning in three distance ranges from the station (54). These instruments use both the received radio and optical signals from a possible flash to identify it unambiguously as lightning. An early single-station proximity sensor marketed by the A.D. Little corporation is used at CCAFS/KSC, but it has low detection efficiency, and so has been supplanted by more advanced systems such as CGLSS (see later) and LDAR (see later). Many inexpensive detectors, some of which are listed in Table 2, are aimed at the consumer market to warn of nearby or approaching lightning. Higher precision flash counters such as CGR3 and CIGRE are often used in thunderstorm research projects, and such instruments have provided much of what is known about lightning phenomenology (39–42;74, Chap. 2).

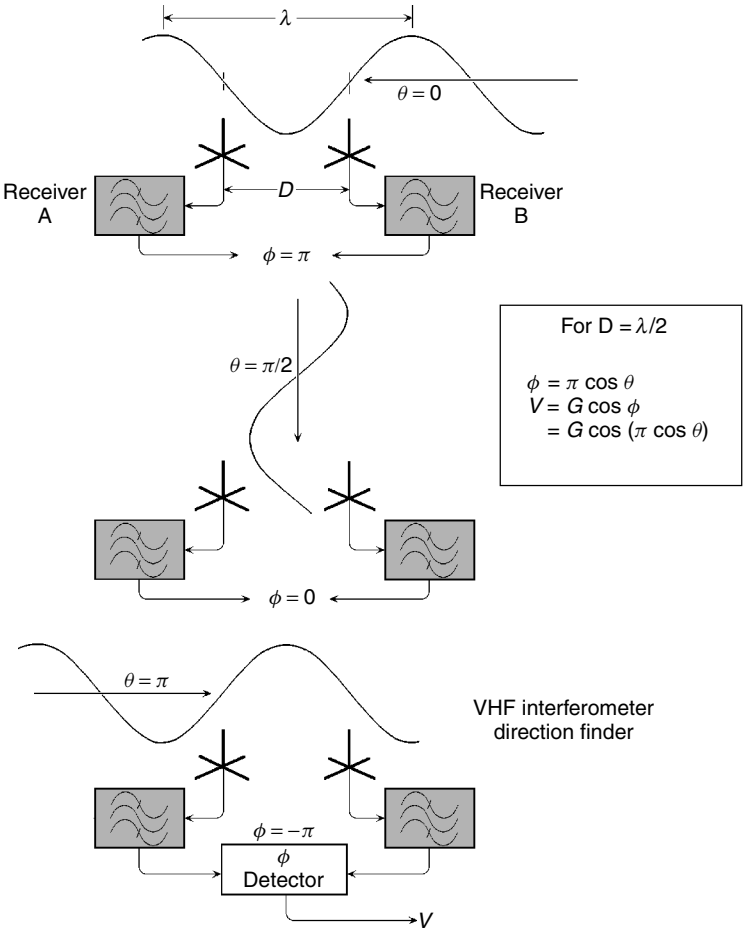


Figure 4. Simplified very high frequency (VHF) interferometer (ITF) direction finder. In this special case, the pair of receiving antennas is spaced apart a distance $D = \lambda/2$, where λ is the wavelength of the incoming signal. Here, λ is the wavelength of the center frequency chosen by the pair of selective receivers, usually in the VHF range of 30 to 300 MHz; this passband is indicated schematically by the middle of the three sine waves at each receiver. The pair uses a common local oscillator to maintain phase coherency. The receiver pair output phase difference ϕ is at a maximum of π radians (180°) when the signal bearing is from the right, parallel to the vertical antenna pair plane, falls progressively to zero as the bearing reaches the normal to the plane, then decreases to $-\pi$ when the bearing is again parallel to the plane but with signals arriving from the left, the opposite direction. Like magnetic direction finders, interferometers experience directional ambiguity because the phase difference does not discriminate between bearings from either side of the antenna plane. The phase detector driven by the receiver pair provides a signal voltage proportional to ϕ. This voltage is then a function of the direction angle θ. The magnitude V of the output voltage for a given direction angle depends on antenna gain, receiver sensitivity, and phase detector characteristics, grouped together as a signal transfer constant G. Because the wavelengths λ are quite short at VHF, a fraction of a meter to a few meters, it is practical to construct and orient the antenna arrays for three-dimensional direction finding and signal source mapping (see later).

Plan view, ground-based *single-station location* (SSL) systems generally obtain the bearing to the lightning source via the MDF technique. Such systems, for example, the Great Plains-1 (GP-1) at Penn State (see later), estimate range using *signal propagation models* that relate the statistical properties of the received magnetic signal level to the distance to the lightning source (56,57). Using this means for estimating range relies on the assumption that there is only one region of lightning in the bearing of interest. If there are multiple regions that are widely separated, then such an algorithm based on peak radiated field magnitude, or peak *signal level b* (Fig. 2), can provide only a single intermediate range that may not apply to any of the actual regions.

Satellite-based SSL platforms are also used. They identify the occurrence of lightning from above by using optical signatures (34,43,46,47) (see later) and/or VHF fields radiated by lightning (49–53).

Finally, a new hybrid single-station locating system, SOnic Lightning LOcation (SOLLO), is being tested at KSC (59). It combines electric field and acoustic signal processing to give extremely accurate lightning locations over short ranges (see later).

Lightning Detection Networks

Networks of electric field mills measure the electric field strength near the ground, and so can provide an indication of impending lightning. Networks of field mills, such as the Electric Field Mill System (EFMS) at the Wallops Flight Facility at Wallops Island, Virginia, and the 31 field mills in the Launch Pad Lightning Warning System (LPLWS) at CCAFS/KSC in Florida (see later), measure the electric field potential gradient to identify the locations of charge centers (38); these centers are related to the point of origin of the lightning channel and so may indicate the threat of lightning triggered by a rocket launch (85). The LPLWS can also detect all types of lightning, albeit over a short range ~20 km (35) and with only moderate detection efficiency and poor location accuracy (Table 2).

Plan view multistation systems or *network locators* (NL) use spatially distributed arrays to detect near-simultaneous lightning signals at multiple receivers; the information from all receivers is then combined by a central processor to deduce the lightning location. Lightning locations can be obtained via *triangulation* to the point of intersection of radials along which the radiated signal traveled toward one or more receivers. The MDF technique illustrated in Fig. 2 provides a radial for each station, as indicated by the dashed lines that pass through stations A and B in Fig. 3 that are west and north of a lightning strike point in the center. In such networks, only crossed-loop B-field antennas are required because E-field measurements are not needed to determine

stroke polarity to eliminate any π-radian (180°) ambiguity. Before an upgrade completed in the mid-1990s, the NLDN was operated by the company Lightning Location and Protection, Inc. (LLP), whose wideband sensors used only the MDF technique to locate lightning (11,13,14).

As indicated in Fig. 3, the lightning location can be obtained from either the intersection of the radials given by the MDF triangulation technique or by the intersection of the hyperbolas given by the TOA technique (see earlier). Normally, systems obtain a least-squares optimum location of a stroke by combining the outputs of two or more MDF stations or three or more TOA stations. Both methods can be applied in the same system to reduce the location error further, as in the current upgraded NLDN that was created by Sankosha Corporation of Tokyo, Japan, when the LLP and ARSI networks were combined (13). This error is reduced by adjusting the initial and subsequent locations iteratively via an objective procedure that accounts for the uncertainties in the estimates given by the separate methods (see later) (22, p. 162).

Three-dimensional *lightning mapping systems* or *network mappers* (NM) use spatially dispersed arrays of instruments to measure either the time of arrival of or the direction to bursts of *radio-frequency* (RF) signals at a rate rapid enough to resolve the entire lightning channel, thereby depicting its development in three dimensions. The locations of the RF sources can be obtained from either intersecting hyperboloids of two sheets given by the TOA technique (86,87) using at least four, but typically at least five, receivers (22, pp. 152–154) (see later), or can be determined from intersecting radials given by the interferometric technique.

Other Methods

Lightning forecasting — predicting the initial occurrence of lightning many minutes in advance — often must be done without the aid of direct precursors such as electric field strength because it must rely on instruments that are routinely available to the weather forecaster. An example of this approach relates the probable occurrence of imminent cloud–ground lightning to the temporal and vertical changes of reflectivity given by precipitation radars, as done, for example, by operational forecasters at CCAFS/KSC (16,88,89). Rapid increases in reflectivity in certain layers within the cloud can be associated with large vertical velocities in the regions where charge separation and lightning initiation primarily occur.

The large current in the lightning channel heats it to temperatures near 30,000 K and generates acoustic shock waves that produce thunder. An old technique for locating lightning is to use networks of microphones to detect where thunder originates (74, pp. 306–307; 90). Such acoustic mapping techniques require an accompanying electric field detector to sense when lightning has occurred (22, pp. 149–151), as used by the TREMBLE system (Table 2). By knowing the time of the lightning stroke, one can obtain the range to the lightning channel by using the time difference between the electromagnetic and acoustic signals received. The location of the lightning channel is deduced from the intersection of each range circle given by the signal time from each microphone (22, p. 149). Ranges are limited to 20 km or so, and errors are of the order of 10 to 15% of the estimated range. Because the propagation of acoustic waves depends strongly on wind speed and direction, the strong wind shear common to thunderstorm environments is a significant source of this error.

Acoustic mapping is used rarely any longer because in many cases, better mapping techniques using a VHF electromagnetic signal are available (see later). One exception occurs when special applications require high accuracy lightning location over short ranges. For example, knowing exactly where lightning attaches to launch pads at CCAFS/KSC is important to determine if the payload, launch vehicle, or support electronics need to be inspected for electromagnetic pulse (EMP) damage. KSC is testing a new hybrid system, the SOnic Lightning LOcation system (SOLLO), that combines both electric field and acoustic detectors (59). SOLLO is a single-station lightning locator (SSL) system that uses one electric field detector and five acoustic sensors. Four of the acoustic sensors are placed in a horizontal circle of 2 or 4 m radius with acoustic sensors 90° apart; the electric field sensor is in the middle of the circle, and the fifth acoustic sensor is 2 or 4 m above the electric field sensor. A sufficiently strong and rapid change in the electric field indicates that a nearby lightning strike has occurred. Waveform analysis (see later) filters out lightning signals that originate more than one mile away. The time difference between the arrival of the electric and acoustic pulses gives the range, and interferometry (see earlier) using the five acoustic sensors gives the three-dimensional bearing to the lightning strike; these two values uniquely define the lightning strike location. SOLLO provides lightning location to within about 5 m for a range of up to 1 km (Table 2).

Finally, precipitation radars also have been used for the direct study of lightning because the highly ionized, overdense lightning channel effectively reflects the radiation emitted by the radar (60). The channel can be detected only for short periods of time, hundreds of milliseconds, because the ions in the channel diffuse quickly (22, p. 147). Moreover, operational radars use rotating antennas, meaning that the channel would be detected for only a brief interval during each scan. Radar is thus used more in research to learn more about lightning channel properties than to locate lightning operationally (60) (Table 2).

RADIATED FIELDS DUE TO LIGHTNING

Lightning is a brief, half-second or less (74, p. 21), electrical discharge that carries a large current and whose pathlength is typically many kilometers (74, p. 8). Thunderstorm clouds, often called *cumulonimbus clouds*, are the most common sources of lightning, and lightning strokes produced by them are the focus of locating systems (see earlier). To understand how these different systems detect various aspects of lightning, the lightning process and its signal characteristics are reviewed here and in the next section.

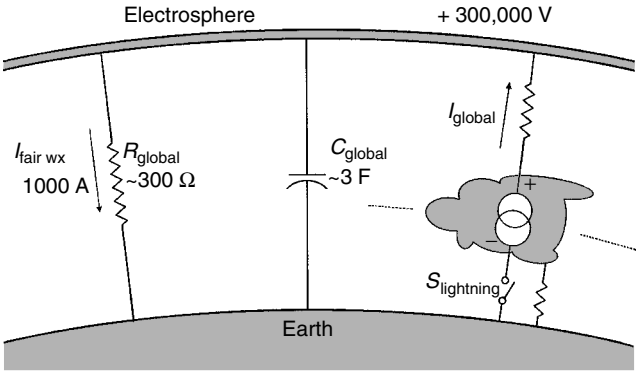


Figure 5. The global electric circuit. The electrosphere is a highly conductive layer of the atmosphere at an altitude of 50 to 75 km. The values of this circuit are estimated by Uman (74, p. 30) as follows: The potential of the electrosphere is about +300,000 volts (denoted V) relative to the earth's surface and establishes the fair-weather electrostatic field of 100 volts per meter (denoted V/m) at the surface. The electrosphere and the earth's surface form the global capacitor C_{global} that is charged to about 10^6 coulombs (denoted C). The electrosphere–surface capacitance is then about 3 C/V or 3 farads (denoted F). The global fair weather leakage current $I_{\text{fair wx}}$ is estimated at $\sim 1,000$ amperes (denoted A), and the capacitor is shunted by a leakage resistance R_{global} of approximately 300 ohms (denoted Ω). The time constant of the global resistance–capacitance combination is thus about 900 seconds (denoted s) and would discharge C_{global} in something more than an hour. In contrast, Volland (94) estimates a slightly smaller electrosphere potential of 240,000 V and a charge of about 7×10^5 C but the same leakage current $I_{\text{fair wx}}$ of $\sim 1,000$ A. Thus a slightly smaller leakage resistance R_{global} of 240 Ω and a capacitance C_{global} of 2.9 F produces a somewhat shorter discharge time constant of 700 s. On a daily basis, thousands of thunderstorms scattered over the earth's surface — up to 1,500 to 2,000 simultaneously (96, p. 211) — maintain a rather constant charging current I_{global} equal to the leakage current. The thundercloud current generators are connected into the global circuit by cloud–ground negative return strokes that act as intermittent switches, $S_{\text{lightning}}$.

The Global Circuit Model

A vertically varying electric field that has a magnitude of about 100 volts per meter (denoted V/m) = 0.1 kV/m at the earth's surface (22, p. 29) is present in the atmosphere at all times and it is called the *fair-weather*, *fine-weather*, or *clear-air* field. The fair-weather field also varies horizontally; it is influenced in part by natural ionizing radiation from different rock types and soil permeability. Globally, the charge of the earth's surface is negative, whereas the charge of the *electrosphere*, whose base is 50 to 75 km above the surface, is positive (74, p. 30; 91, p. 236). The earth–electrosphere combination can be modeled as a capacitor (22, p. 31), whose plates are at the earth's surface and in the electrosphere (Fig. 5). The annual average voltage difference between these two plates is approximately 300 kV, and they maintain a charge of about 10^6 coulombs (denoted C) (74, p. 30; 92,93). The capacitance C_{global} of this electrosphere–surface component is then about 3 C/V or 3 farads (denoted F). The *global leakage current* $I_{\text{fair wx}}$, also called the *fair-weather current*, and the *air-earth conduction current*, is of the order of 1,000 amperes (denoted A) (22, p. 29; 73, p. 30; 91, p. 236), which results from a *global leakage resistance* R_{global} of about 300 ohms (denoted Ω). The *global circuit time constant*, $R_{\text{global}}C_{\text{global}}$, is then 900 seconds (denoted s) and would discharge the capacitor almost completely in a bit more than an hour. Volland (94, pp. 15–20) estimates slightly different values for the electrosphere charge (6.8×10^5 C) and potential difference (240 kV) but still finds about the same capacitance (2.9 F), although with a slightly lower resistance R_{global} (240 Ω) and a shorter time constant $R_{\text{global}}C_{\text{global}}$ (700 s) that would imply a discharge time of about an hour. Estimates of R_{global} closer to 200 Ω are also found in the literature [e.g., (91, p. 252)]. As proposed by Wilson (95), numerous simultaneous thunderstorms across the earth — as many as 1,500 to 2,000 (96, p. 211) — provide a mechanism for maintaining the charge in this capacitor by supplying an average *thunderstorm current* or *supply current* I_{global} equal to the global leakage current $I_{\text{fair wx}}$ (97). Together, the two currents form the *global circuit* (22, p. 29; 96). The thunderstorm generators are connected effectively to the global circuit via switches $S_{\text{lightning}}$ that are closed intermittently by cloud–ground negative return strokes.

Uman (74, pp. 9–10) cautions that the net charge redistribution by a lightning stroke is well known but is expressed only as a difference in charge between centers in the cloud and at the ground. The observed difference can be equally well described by deposits of positive charge in the cloud or by deposits of negative charge at the ground, a process labeled by Uman as *effective lowering of charge*. The positive current resulting from the effective lowering of negative charge to the ground is I_{global}, and schematically it extends upward through the thunderstorms to the electrosphere (Fig. 5).

For a lightning discharge from a thunderstorm to effectively lower negative charge to the earth's surface, separate regions of net positive and net negative charge must occur within the cloud itself. Various charge separation mechanisms have been proposed, as recently reviewed in MacGorman and Rust's book (22, Chap. 3). Essentially, the descending heavier hydrometeors in the cloud tend to accumulate negative charge, and the smaller and lighter convectively rising hydrometeors accumulate positive charge. This leads to the *dipole/tripole model* of thunderstorm electrification, a somewhat oversimplified description of the typical charge distributions within a cumulonimbus cloud (22, pp. 49–53; 98). In this model, generally a lower positive charge center occurs near cloud base, a main negative charge center in the middle portion, and a main positive charge center in the upper portion of a cumulonimbus cloud [see Fig. 8.1 in (92), Fig. 1 in (97), or Fig. 3 in (99)]. This distribution results in a net positive charge within the thunderstorm anvil that occurs downwind of the main cell and is near the tropopause at altitudes of 10 kilometers (denoted km) or more above the ground. Depending on the storm type (99), the main negative charge centers occur at altitudes from 5 to 9 km or so where the temperature is between −10 and −25 °C [Fig. 8.5 in (92);(100,101)]. The main positive centers occur 1 to 5 km above the negative centers where the temperatures are in the range

−30 to −60 °C (22, pp. 193–194; 38,92,97). These altitudes are well above a typical cloud base, which is 1 to 4 km above the ground where temperatures are usually higher than 0 °C. Potential gradients near the main negative charge centers are of the order of 100 kV/m and at cloud base are of the order of 50 kV/m (22, p. 51; 97). Pointed objects at the ground respond to the electric field produced by the cloud and so emit charge. As a result, the potential gradient near the surface is reduced to values between 1 and 10 kV/m (92). Imminent lightning is thus expected when the magnitude of the potential gradient measured near the earth's surface exceeds a threshold of about 1 kV/m, which is about 10 times greater than the typical fair-weather field. The field mills in the LPLWS and EFMS networks (Tables 1 and 2) are designed to measure electric fields from ±15 kV/m (35,38) up to ±32 kV/m (102), and so they can easily capture the surface field of an intense thunderstorm overhead that could be as great as 20 kV/m (22, p. 121).

Lightning discharges may effectively transfer charge from cloud to ground, via *cloud–ground strokes*, and may also redistribute charge within the atmosphere, via *cloud discharges*. Such discharges may occur entirely within the cloud itself (*intracloud* lightning), between two separate clouds (*cloud-to-cloud* lightning), or between the cloud and the surrounding clear air (*cloud-to-air* lightning). Commonly, and throughout this article, the term *intracloud stroke* is used synonymously with the term *cloud discharge*. Globally, estimates given by flash counters of the average ratio of the number of cloud discharges to the number of cloud–ground discharges vary from 3.5 (40) to 5.7 (103). Estimates from the OTD satellite-based sensor (see later) range from 2.5 to 5.7 (46). Moreover, this ratio is not constant throughout the globe; it varies in the Northern Hemisphere from about 2 at latitude 60° N to between 6 and 9 at the equator (103,104). Results obtained from efficient total lightning detection systems, however, now show that ratios are typically much greater, between 10 and 30 (105) (Fig. 35). Although not as frequent as cloud discharges, cloud–ground lightning has been a greater focus of lightning studies and real-time location systems, owing to its critical roles in maintaining the global circuit, compromising human safety (63), and producing economic impacts via, for example, disruption of electrical power, initiation of forest fires, and damage to residences (61;74, p. 8). Interest has focused recently on using the frequency of intracloud lightning to forecast the onset of cloud–ground lightning or of other thunderstorm hazards (24,65) (see later).

Radiation and Electrostatic Components

For most applications, the *far-field* or *radiation* components of the electric (*E*-) and magnetic (*B*-) fields propagate from their source at the speed of light (73, p. 62) and are of importance to lightning location methods that use radio techniques (5, p. 353). To detect the rates of change in these components on the timescale of a microsecond (75, pp. 194–197), fast *E*-field sense antenna systems, called *fast antennas*, are typically used (11;22, pp. 106–107). In contrast, *slow antennas* measure the total change in the field, or the *near-field* or *electrostatic* component, produced by an entire flash over a timescale of a second (75, pp. 191–193). (Both "fast-antennas" and "slow-antennas" may use the same physical antenna; the difference is simply based on the following receiver bandwidth; see Lightning Signal Discrimination.) A sufficiently large temporal rate of change of the current I in a flash generates a radiated field about a certain frequency f. This frequency is proportional to the inverse of the quarter wavelength $\lambda/4$ of the radiated field, or equivalently to the inverse of the length $l \sim \lambda/4$ of the lightning channel. The far-field components dominate at ranges r from the radiator greater than a few wavelengths λ, or equivalently at ranges r that are much greater than l; this latter relationship is usually expressed as the inequality $r > 10l$. Thus, for a typical cloud–ground lightning channel of length $l = 5$ km, the far-field components would dominate at ranges r greater than 50 km. Pierce (5, p. 353) notes that the far-field components dominate at ranges $r > c/2\pi f$, where c is the speed of light equal to 3×10^8 m/s. At distances greater than 50 km, broadband receivers using frequencies f at or above 1 kHz, as is typical in systems using the magnetic direction finding (MDF) technique (11; Fig. 2; see later), would detect primarily the far-field components. Equivalently, receivers that measure signal magnitudes or signal levels at frequencies of 10 kHz or greater need be concerned only with these components at ranges greater than 5 km (106, p. 545). Restriction to reception of the far-field component of the lightning signal is important in SSL systems, because theoretically the received amplitude of this field varies at a rate proportional to $1/r$ (5, p. 353; 106, p. 545). Ranging algorithms used by many single-station locator (SSL) systems are based on this relationship between peak B-field signal level b and range r, although the power to which b must be raised may differ from -1 (56) (see later). Finally, single-station flash counter (SSC) systems designed to detect lightning at ranges less than 5 km from the receiver must contend with both the far-field and near-field components, making the determination of accurate range estimates challenging (see later).

The earliest single-station locating (SSL) and plan view network (NL) systems for locating lightning concentrated on cloud–ground strokes (1) because the direction to these strokes tends to be easier to locate unambiguously from a distant receiver than the direction to cloud discharges. The cloud–ground stroke channel tends to be vertically oriented near the ground where the initial magnetic field peak is radiated (82), and so at the far field — at distances much greater than the channel length (14) — the channel appears as a vertical radiating antenna at a point source (106, p. 544). In contrast, intracloud strokes have channels that typically span 5 to 10 km in the horizontal (74, p. 21), but these channels may extend much farther, up to many tens of kilometers (107–109). At relatively close range, definition of an unambiguous bearing using a radio receiver is made difficult by such long horizontal channels for intracloud strokes but is relatively simple for the near-surface vertical channels of cloud–ground strokes. For example, a 10-km long intracloud stroke oriented orthogonal to a radial extending

from the receiver spans 11° of azimuth at a 50-km range, about 5 1/2° at 100 km, but only about 1° at 500 km.

At ranges greater than 100 km or so, the bearing to an intracloud stroke can be hard to determine accurately for a different, perhaps somewhat subtle reason. The predominantly horizontal intracloud channel acts as a horizontally polarized radiation source (58;81, p. 184). Such sources propagate electromagnetic waves via two pathways, one via a ground wave and the other by a longer path sky wave (75, pp. 224–228). In contrast, vertically polarized sources, such as cloud–ground strokes near the earth's surface, propagate only via the shorter path ground wave. The apparent bearing of the combined sky wave and ground wave from an intracloud stroke is ambiguous owing to phase distortion and yields an elliptical pattern on an oscilloscope display (4,7;81, p. 185), as used in the original cathode ray direction finding (CRDF) technique (2). In contrast, the apparent bearing of a return stroke is quite discrete and appears as a spike on such a display (11), which allows MDF systems to produce bearing accuracies of a degree or so (7).

As described by ray or reflection theory, the *ground wave* or *ground pulse* is a signal that propagates to a receiver along the earth's surface, whereas the *sky wave* arrives at a receiver after one or more ionosphere–earth reflections. The *first-hop sky wave* after one reflection is normally the one of concern for lightning location (75, pp. 224–228), although some very long-range lightning detection systems have begun to employ *multiple-hop sky waves* (20). The difference in time between the arrival of the later first-hop sky wave and the earlier ground wave is $\Delta t_1 = (r_1 - r)/c$, where as before $c = 3 \times 10^8$ m/s is the speed of light and r is the range from the receiver to the source. Here r_1 is the phase path of the first-hop wave, and it is related to r and the virtual height h_1 of the first-hop sky wave by $r_1^2 = r^2 + 4h_1^2$. During the day, the virtual height $h_1 \sim 70$ km, and at night $h_1 \sim 90$ km because the ionosphere becomes less active after sunset. As a result, the height of the ionosphere effectively rises and so leads to the common occurrence of more distant nighttime AM radio receptions due to the sky wave. For unambiguous separation between the ground and sky waves, the arrival-time difference Δt_1 must be greater than the duration τ of the stroke, often called its *pulse length*. The condition here that $\Delta t_1 > \tau$ can be written as

$$r < \frac{4h_1^2 - \tau^2 c^2}{2\tau c}. \qquad (1)$$

For typical pulse lengths τ of 100 to 200 microseconds (denoted μs) for return strokes, separation of the ground and sky waves based on arrival-time considerations alone can be done for ranges only up to 300 km during daytime and up to 500 km at night. Because the bearing to a return stroke is determined from the ratio of the peak values of the two B-field components (*ew* and *ns* in Fig. 2), and because these peaks occur less than 10 μs after the signal level rise begins, bearing accuracies are maximized (82). Thus, many plan view network locator (NL) systems such as the NLDN focus on locating cloud–ground strokes based on the detected ground wave and in so doing can attain reasonably good bearing accuracies for ranges upto several hundred kilometers (13).

Time-Domain Waveform Analysis

Fortunately, the electromagnetic waveforms of intracloud and cloud–ground strokes graphed as functions of time are noticeably different (Figs. 6, 7, and 8), meaning that the stroke type can be determined by using suitable *time-domain waveform analysis* of measurements of radio signals received by very low frequency (VLF, 3 kHz to 30 kHz)/low frequency (LF, 30 kHz to 300 kHz) systems (11). Negative return strokes have an initial positive E-field polarity, seen in the graph as an increase of the E-field to a positive peak value (Fig. 6), and positive return strokes have an initial negative polarity, seen as a rapid decrease to a negative peak value. The waveforms for either the electric or magnetic waves can be used in this analysis, because the temporal values of the radiated electric and magnetic components in the far field follow exactly the same-shaped curve (6,110). In time-domain waveform analysis, the characteristics of the waves are distinguished by calculating various parameters obtainable when the radio wave is graphed as a function of time.

One useful wave-shape parameter is the ratio of the first peak value to the largest peak value of opposite sign — known as the *overshoot tolerance* (14) — that separates the distinct bipolar shape of the intracloud wave,

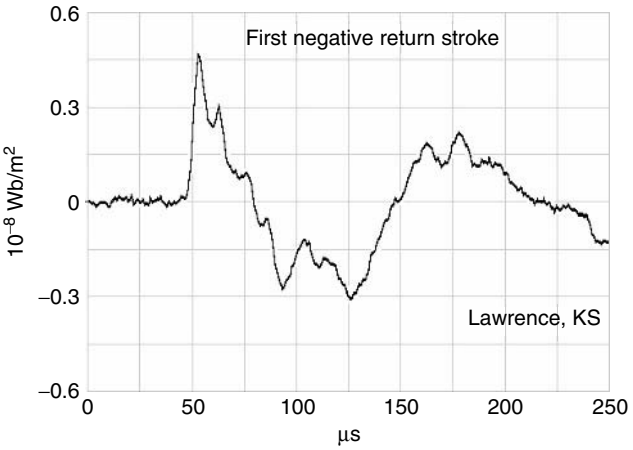


Figure 6. Magnetic field (B-field) waveform acquired by the VLF/LF GP-1 system at Lawrence, Kansas, at 23 : 14 UTC on 18 August 2000 from lightning in the region of Little Rock, Arkansas, about 500 km distant at a bearing of 150°. This first return stroke wave shows the characteristic multiple peaks separated by ~10 microseconds (denoted μs) on the descending limb of the initial excursion created by the stroke branches. The rise time, or the time necessary for the signal level to increase from 10% to 90% of its peak value, is ~4 μs and the waveform width, or the time over which the value remains positive after the peak is attained, is ~25 μs; both values are typical for return strokes. The negative-going undershoot between 80 and 150 μs and the following overshoot between 150 and 220 μs are somewhat exaggerated by high-pass filtering necessary to remove excessive interference from the local 117Vac power service. (NLDN data for stroke location are cited with permission of Global Atmospherics, Inc.)

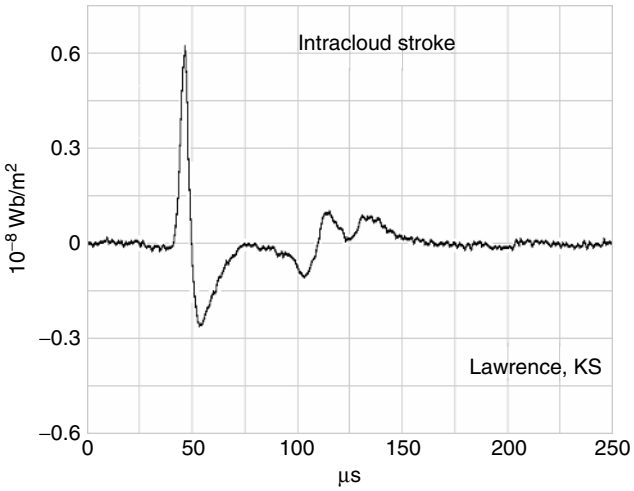


Figure 7. As in Fig. 6, but from an intracloud stroke at 22 : 48 UTC. Although such strokes can be many kilometers long, they are frequently far shorter. Such a short stroke generates a somewhat steeper rise to the initial peak and a far more rapid fall in the descending limb than a cloud–ground return stroke. The rise time here is ~1 μs, which is noticeably shorter than the value in Fig. 6 for the return stroke, and the waveform width is ~4 μs, much less than that for the return stroke. These values are typical of those for intracloud strokes. A sharp undershoot, whose peak occurs about 10 μs after the initial peak, is also a common feature. (NLDN data for stroke location are cited with permission of Global Atmospherics, Inc.)

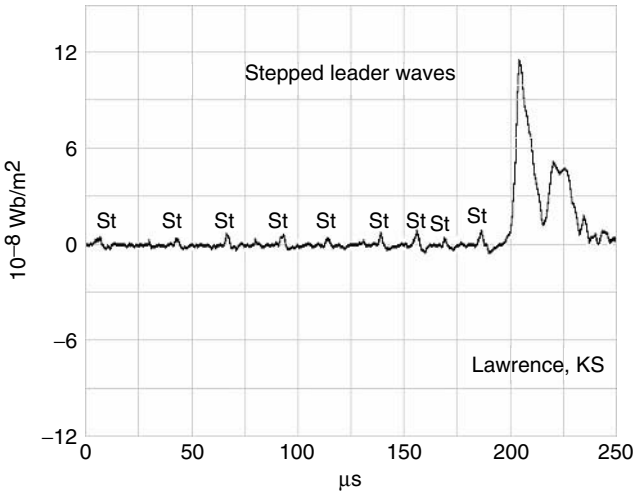


Figure 8. Magnetic field (*B*-field) waveform showing the last few stepped leader pulses (St) preceding a negative return stroke in a flash acquired by the VLF/LF GP-1 system at Lawrence, Kansas, at 2 : 11 UTC on 20 August 2000. This stroke, reported by the NLDN to be about 75 km northeast of Lawrence, had an extremely large peak current of 41 kA. This value was the largest reported between 2 : 00 UTC and 3 : 00 UTC, over which the median value was the much lower value of 9 kA. Large currents of many tens of kA are rare but can produce sizable signal levels even at ranges of many tens of kilometers. Small stepped leader waves are difficult to discern in recordings from strokes beyond 100 km, because on average, even the largest peak from any leader pulse has a radiated signal level that is less than 10% of that of the return stroke peak. For the return stroke that occurred after 200 μs, the rise time is ~2 μs, and the waveform width >25 μs; both are values typical of return strokes. (NLDN data are cited with permission of Global Atmospherics, Inc.)

in which the positive peak is often about equal in magnitude to that of the negative peak, from the monopolar shape of the negative return stroke, in which the positive peak level is significantly greater than the level for the negative peak. As noted by Weidman and Krider (9), the monopolar shape of the return stroke wave, with a rapid rise to a peak value, is caused by the current surge up the predominantly vertical channel that is formed by the stepped leader as it nears the earth's surface (see next section). In contrast, Weidman and Krider (10) suggest that the bipolar shape of the intracloud stroke wave arises from the superposition of two types of currents: A series of fast pulses associated with the formation of the lightning channel produces the positive peak; this series is followed by a slower surge of current giving the second peak that occurs either during the establishment of the channel or just after it is formed.

A characteristic of a lightning radio wave is that there is a short *E*- or *B*-field *rise time*, the time during which the absolute value of the signal level increases from 10 to 90% of its first peak (111). Apparent rise times vary, depending on whether the electromagnetic wave propagates over salt water or over land. For return strokes that propagate over salt water, which is a good conductor, rise times of the order of 0.2 μs or less are observed (9); but for waves traveling 200 km over land, rise times increase to 1 μs or so (112). Such fast rise times are used to separate the electromagnetic waveforms of lightning from other received signals, but to do so requires lightning signal processors that have at least a 100-nanosecond (denoted ns) resolution. In the sample waveforms from the GP-1 system at Lawrence, Kansas, rise times for first return strokes (see next section) over land increase from ~2 μs at a 75-km range (Fig. 8) to ~4 μs at a 500-km range (Fig. 6). To distinguish signals from lightning and other nonlightning sources such as arcing when motors start or light switches are thrown, the original gated magnetic direction finders used an upper limit of 8 μs (7); in contrast, the direction finders used in the first lightning location networks used rise times as long as 20 μs (11). Finally, the *B*-field rise time for the intracloud stroke, ~1 μs in Fig. 7, is substantially less than the rise time for the return stroke, ~4 μs in Fig. 6, a characteristic in addition to the overshoot tolerance that aids in distinguishing intracloud from return strokes.

Return strokes also have a sufficiently long, *waveform width* or *pulse width*, the time during which the *B*-field signal value has the same sign after the first peak is attained. Thus, differences in the waveform width can also be used to aid in distinguishing intracloud from return strokes. The positive *B*-field signal has a waveform width of about 25 μs for the negative return stroke, but it is only about 4 μs for the intracloud stroke. For wideband MDF networks used in the 1970s to detect return strokes, pulse widths greater than 15 μs (11) were proposed. The minimum value of the waveform width applied to the sensors used in the NLDN was decreased from 11 to 7 μs

as part of the system upgrade in 1995 (14,15,113,114), and the overshoot tolerance was increased from 0.85 to 1.2 (14,113). These changes may have led to an increase in the number of positive return strokes detected by the system (13), an increase apparently contaminated by misdiagnosis of some strokes that were actually a type of intracloud pulse (15) known as the narrow positive bipolar pulse (NPBP) (115) (see later). Finally, the characteristic values of parameters such as the rise time and waveform width for the return stroke waveforms vary seasonally and with the location of the thunderstorms on the globe (116).

The stepped leader that precedes the return stroke (see next section) appears as a series of monopolar pulses of significantly smaller amplitude than that of the return stroke (Fig. 8). Indeed, a nearby stepped leader can produce pulses whose peak amplitudes are as large as that for a more distant return stroke. Time-domain waveform analysis can be used, however, to distinguish the stepped leader pulses from the return stroke pulse.

Smaller values of these wave-shape parameters characterize waves sensed by wideband very high frequency (VHF) systems. For example, the waveform measured by one 10- to 500-MHz interferometric system has the same shape as that measured by wideband VLF systems, but has much shorter rise times (~5 ns), much smaller waveform widths (~10 ns), and much smaller amplitudes (27). Measurements using vertically polarized dipole antennas have produced similar results (117). Thus, ground-based VHF lightning-mapping systems such as LDAR, LMA/LMS, or SAFIR (see later) also use the ground wave radiated either by stepped leaders and *K*-processes or dart leaders (see later section) that have much shorter, 1 to 20 μs, pulse lengths (75, p. 228; 86), or by the streamer–leader transition that has a much shorter duration, less than 10 ns (27). For unambiguous discrimination of ground and sky waves, such short-duration, ~20 μs, pulse lengths lead via Eq. (1) to maximum ranges greater than 1,500 km during the day and 3,000 km at night (75, p. 228). Thus, the limited detection range ~250 km for VHF systems is not tied to distinguishing the ground and sky waves but instead is tied to the much lower signal level than in the VLF (see later sections) and to line-of-sight propagation of the VHF radiation produced by the discharge process (8).

Stroke Polarity Conventions

The conventions used to relate the positive and negative polarities of a lightning stroke to the deflection of the measured electric field (*E*-field) vary with author, and so the reader must remain wary. In general, strokes of negative polarity effectively lower negative charge to the earth's surface, and strokes of positive polarity raise negative charge away from the earth's surface (see next section). The literature on return stroke waveforms normally uses the convention used in this article, that strokes of negative polarity produce a positive deflection on an instrument measuring the *E*-field [e.g., Fig. 2 in (9), Figs. 6 and 8, and Fig. 10], and strokes of positive polarity produce a negative deflection.

Unfortunately, this convention is not followed uniformly in studies of intracloud strokes. Consistent with the convention used in studies of cloud–ground strokes, some authors state that intracloud strokes of negative polarity produce a positive deflection, and strokes of positive polarity produce a negative deflection, of the *E*-field intensity (10,118,119). The opposite, so-called *physics convention*, is used by others, however, whereby a stroke of positive polarity is said to produce a positive *E*-field deflection (115,120).

The physics convention is followed primarily in articles that discuss a particular short-duration bipolar pulse, upon which substantial high-frequency components are superimposed, that is labeled the *narrow positive bipolar pulse* (NPBP) or *narrow negative bipolar pulse* (NNBP). These pulses have a significantly shorter duration — only 10 to 20 μs (115,119,121) — than those of the intracloud strokes discussed in (10) or illustrated in Fig. 7, have peak amplitudes comparable to those of return strokes (120), tend to occur early in the flash (118), and produce the largest signal of any discharge process in the VHF (119). Generally, NNBPs occur much more rarely than NPBPs (53). These narrow bipolar pulses are confined to the main negative and upper positive charge regions of the thunderstorm dipole and are the initial event of otherwise normal intracloud strokes (32,118). Recently, these pulses have been associated with *transionospheric pulse pairs* (TIPPs), which are pairs of powerful high frequency (HF, 3 MHz to 30 MHz)/VHF signals that last ~10 μs, that occur ~50 μs apart, and that were first detected by the BLACKBEARD VHF radio instrument on the ALEXIS satellite (52,53,122).

LIGHTNING MORPHOLOGY

Understanding the design requirements of a lightning locating system requires thorough knowledge of the morphology of a lightning flash (Fig. 9). Intracloud flashes have many components similar to those in cloud–ground lightning. Many intracloud flashes have a bilevel, horizontally branching structure that extends within the main positive charge regions (32,120); in the decaying stage of a thunderstorm, these branches can extend horizontally many tens of kilometers in what is known as *spider lightning* (108,109). In its active phase, an intracloud flash typically discharges the main negative charge region of a thunderstorm by effectively transporting negative charge upward via nondiscrete, continuing currents in a channel that joins the main positive and negative centers. In contrast, a cloud–ground lightning flash, which is described in detail in the following paragraphs, is composed of one or more discrete strokes that effectively lower charge from either the main positive or negative charge centers in a thunderstorm to the earth's surface. Both types of lightning flashes produce the bright light that can be seen and the thunder that can be heard.

The Stepped Leader

When the potential difference between the main negative center in the middle portion of the cloud and the positive center below is large enough, *preliminary breakdown* is said to occur, and a small charge begins to be lowered

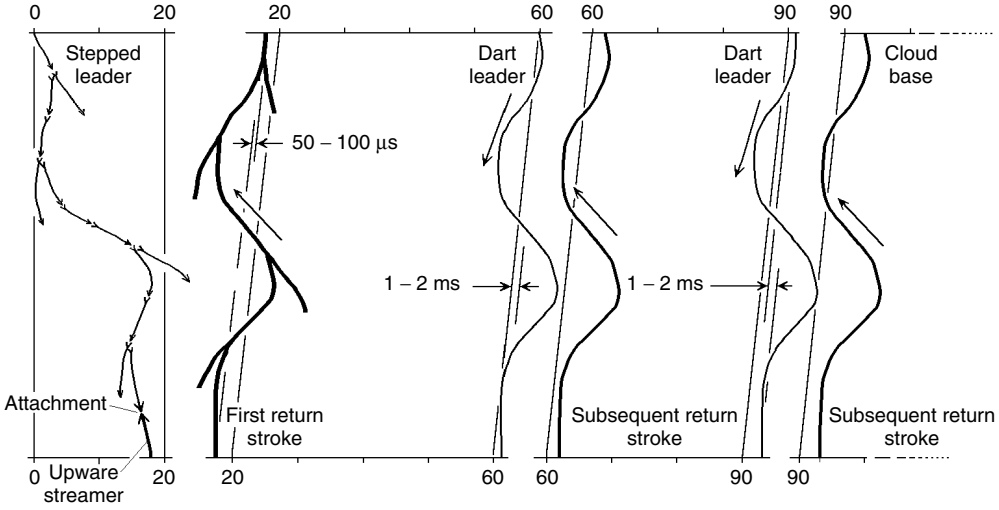


Figure 9. The sequence of events in a typical negative cloud–ground flash, where time increases from left to right. A stepped leader emerging from a region near cloud base initiates the flash. Each leader step is approximately 50 m long. The next step occurs about 50 μs after the first, and the process repeats until a leader step nears the surface in some 10 to 30 milliseconds (denoted ms). This near-ground step raises the local electric field gradient and initiates an upward streamer. When the upward streamer meets this last step of the stepped leader a few tens of meters above the surface, the return stroke begins. The return stroke travels up the ionized channel established by the stepped leader at about one-half the speed of light. The stepped leader generally displays many branches as it meanders toward the surface. The duration of this, and subsequent, return strokes, is 50 to 100 μs, depending on a stroke length of one to a few kilometers. If sufficient charge becomes available at the top of the lightning channel via *J*- and *K*-processes, then after a few tens of milliseconds, a dart leader descends continuously from the cloud, follows the first return stroke channel, and nears the surface in 1 to 2 ms. On reaching the surface, a subsequent return stroke is initiated that follows the channel of the dart leader. These subsequent return strokes are not branched. When they contain continuing currents, these subsequent strokes may contain impulsive brightenings known as M-components, which some have erroneously labeled as additional subsequent strokes. Each subsequent return stroke may be followed by another dart leader, in turn, initiating another return stroke. Thus a flash may consist of as many as 26 strokes (140) and has a flickering flash duration of up to two seconds [after Fig. 1.6 in (74)].

toward the ground. This lowering proceeds in several discrete intervals or steps in the process termed the *stepped leader*. Each step is tens of meters long, lasts about 1 μs, carries currents of the order of 1,000 A, and propagates at speeds that range from 1×10^5 m/s to 2.5×10^6 m/s, and at an average speed whose estimates range from 2×10^5 m/s to 3×10^5 m/s (22, pp. 89, 93; 74, pp. 11, 82; 68, p. 10; 123). Each leader step produces a small broadband electromagnetic pulse that has a distinct waveform (denoted St in Fig. 8) that precedes the much larger pulse produced by the return stroke. The average peak signal level, or magnitude, of a stepped leader pulse is ~10% of the peak signal level of the following return stroke field and has a significant frequency content ranging from 200 kHz to 1 GHz (5, p. 356; 8,28,117,124–126). The orientation of the stepped leader is not solely vertical and forms a lightning channel that has many branches or forks (Fig. 9). The horizontal extent of this channel is typically limited by the size of the thunderstorm (22, pp. 199–203), but it can extend a significant distance horizontally. If it remains aloft, then it will lead to a cloud discharge; but if it nears the ground, then it will lead to a cloud–ground discharge. In some cases, the horizontal channel can extend so far horizontally before becoming vertical that the resulting stroke can appear as the occasionally reported *bolt-out-of-the-blue* (32) (Fig. 33).

The First Return Stroke

After approximately 20 milliseconds (denoted ms), the lightning channel defined by downward propagation of the stepped leader is close enough to the ground that *upward streamers* or *connecting leaders* from the surface, which are initiated by electric field changes at sharp objects, attach to the channel at *striking distances* of several tens of meters above the ground (74, p. 107); this attachment process initiates the *return stroke* (Fig. 9) (74, p. 12). As many as 50% of the negative return strokes have multiple attachments to the ground (127,128), making the location of the contact point by networks of sensors challenging. This stroke, which lasts 70 to 100 μs, propagates up the leader channel at speeds that approach one-third to one-half of the speed of light and that decrease as elevation above the ground increases (74, p. 13; 123,129,130). Typically, several of the many tens of coulombs of available negative charge are effectively transferred to ground by

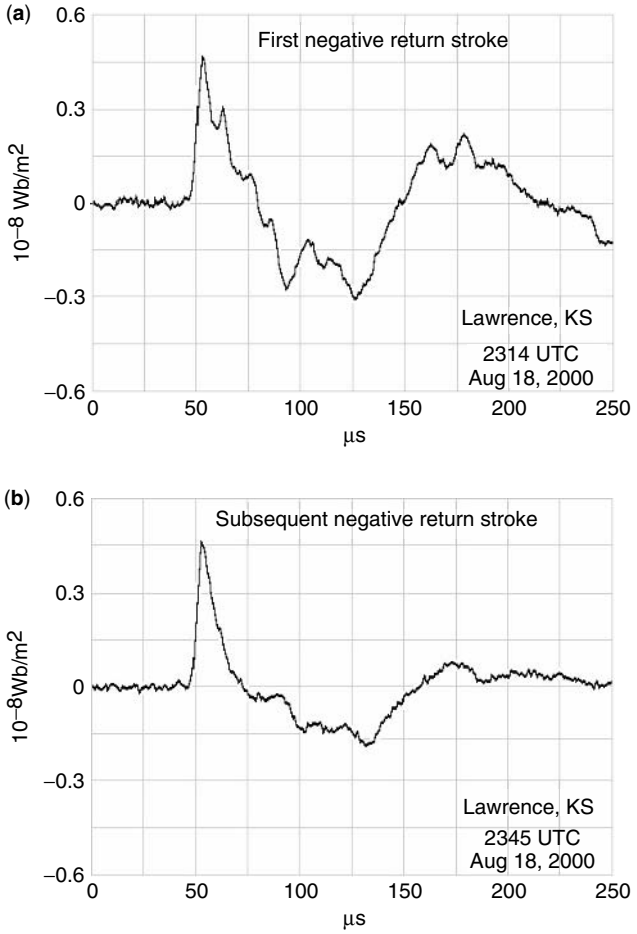


Figure 10. As in Fig. 6, but illustrating the difference in waveforms acquired for first return (**a**) and subsequent return strokes (**b**). The waveform in part (**a**) is the same one as in Fig. 6, but the waveform in part (**b**) was obtained from a different flash in the same region at 23 : 45 UTC. Subsequent return strokes are seldom branched and thus show a relatively smooth descending limb in the initial wave excursion. The rise time for the first stroke in (**a**) is ∼4 μs, and for the subsequent stroke in (**b**) is ∼5 μs. Owing to the lack of subsidiary peaks, the waveform width of the subsequent stroke (∼20 μs) is a bit narrower than that of the first stroke (∼25 μs). Although the peak levels shown are typical of those for return strokes, normally the peak level for a subsequent return stroke is less than that for the associated first return stroke. The two return stroke waveforms shown here are from two different flashes and so only coincidentally have similar peak levels.

the stroke. Table 3.1 in (22) provides a summary of several published estimates of the available charge in the main negative center of the thunderstorm dipole; these charges vary from 20 to 340 C, and the majority is in the range of 20 to 50 C. Table 7.2 in (74), adapted from the work of Berger et al. (131), reports that the median value of the charge transfer by a first return stroke is approximately 5 C. The typical peak current near the ground in this first return stroke is about 30 kA but can vary from 10 to 200 kA or more (131). Within 50 μs or so, the current in the lightning channel falls to 50% of its peak level, but currents of the order of 100 A may continue to flow in the channel for up to several hundred ms (74, p. 13) (see later). Owing to field attenuation with distance, the observed peak levels of the electromagnetic signals are often *range-normalized* to levels that would be observed if the lightning were at a fixed distance from the receiver. The standard distance used is 100 km (74, p. 111; 132,133) (see later), although 10 and 50 km are also used (106, p. 565; 125).

The first return stroke generally has a forked appearance because the stepped leader creates a branching pattern as it propagates toward the ground to define the return stroke channel (Fig. 9). These ionized branches in the channel cause the first return stroke to produce an electromagnetic waveform that has three or four local maxima, or *subsidiary peaks*, of substantial amplitude that occur every 10 to 30 μs after the initial largest peak (Fig. 10a) (9). Subsequent return strokes (see later) that might occur within the flash have much smaller subsidiary peaks, because these return strokes tend to be unbranched (Figs. 9 and 10b). This characteristic difference between the first and subsequent return stroke waveforms can be used by locating systems to identify the first return stroke that typically generates the largest signal level because it has the greatest current (11).

Generally speaking, the longer the radiating channel, the larger the peak current, and the lower the frequencies at which the peak amplitude of the generated electromagnetic signal occurs (106, p. 544). With a large-current discharge spanning a kilometer or more vertically, the return stroke thus has a peak amplitude at relatively low frequencies near 5 kHz (134). The fine spatial structure within the return stroke channel and the rapidity of the discharge ensure significant higher frequency components, creating a broadband spectrum whose power varies proportionally to $1/f$ through frequencies f of 1 MHz or so and to $1/f^2$ for larger frequencies (5, p. 369; 125,135). Design uncertainties in early narrowband receivers may have led to the suggestion of a slower, $1/f$, fall-off of amplitude at all frequencies tested above 1 MHz (136, pp. 196–197). These narrowband systems measure the net effect of several individual waveforms and so give misleading spectral results. Nevertheless, those using narrowband VHF systems to study the electromagnetic emissions from all stages of the lightning process (24) cite the $1/f$ decrease in amplitude with frequency. That the occurrence of radiated power at these higher frequencies is related to the forks in the lightning channel created by the stepped leader is suggested by the fact that the spectra for both the leader step waves and the first return strokes have similar shapes and radiated power distribution in the frequency range of 300 kHz to 20 MHz (125).

Subsequent Return Strokes

As noted earlier, the first return stroke effectively lowers several coulombs of the available charge (131). If an insufficient amount of charge remains to initiate a second stroke, then the lightning flash is a *single-stroke flash*. After the first return stroke has occurred, charge may become available at the top of the channel, moving within the cloud via relatively slow J- and fast K-processes, and about 40 ms after the return stroke ends may initiate a

continuous *dart leader* that, unlike the discrete steps of a stepped leader, propagates down the channel continuously (Fig. 9). Interferometric (ITF) mapping techniques, used, for example, by SAFIR/ONERA-3D, that detect extremely short duration (<10 ns), continuously pulsed events within the lightning discharge process (27,126) work well in detecting negative dart leaders and *K*-processes (86,137) (see later). In contrast, TOA techniques, used, for example, by LDAR (see later section), focus primarily on longer duration (a few μs), discretely pulsed negative stepped leader steps and junction processes in virgin air and so do not track dart leaders well (86). The dart leader propagates down the first-return stroke channel, generally without branching, at speeds ranging from about 0.2×10^7 to 4×10^7 m/s [Table 5.3 in (22); Table 8.1 in (74), which is adapted from (130); p. 10 of (68);(123)]. When the dart leader nearly reaches the ground, a second return stroke occurs as part of a *multistroke flash*. According to Table 7.2 in (74), which is adapted from the work of Berger et al. (131), the median value of the charge transfer by a subsequent return stroke is approximately one-quarter that of the first return stroke, a bit over 1 C. For two cells in Florida, Nisbet et al. (38) found that the charge transfer by the entire flash was in the range of 1 to 27 C; values up to 150 C have been measured [see Table 2 in (35)]. Because the human eye discerns only the light created by each return stroke, multistroke flashes appear to flicker. The second and any additional strokes in a flash follow the well-defined channel upward, are usually not forked (Fig. 9), and so produce a smoother electromagnetic waveform that has few or no subsidiary peaks (Fig. 10b), (9). As a result, only a relatively small amount of radiation above 1 MHz is generated (74, p. 119; 138). Because the current in a subsequent stroke is usually less than that in the first return stroke, typically about one-third [see Table 7.2 in (74)], the signal level produced by a subsequent stroke is generally much less than that of the first return stroke (131). Signals from first return strokes are the most reliable return stroke signals to describe the peak signal level of a lightning flash and to provide range estimation using signal propagation models; from their waveform shape, they can easily be distinguished from the subsequent stroke.

Additional dart leaders may occur to initiate subsequent return strokes. Although claims exist in the literature of more than 40 [e.g., (77, p. 65)], such a high number is likely to be given by erroneously counting the numerous impulsive increases in luminosity that may occur when the tip of the upwardly moving subsequent return stroke streamer reaches a branching point of the channel that was produced originally by the forked stepped leader (71, p. 18). Such impulsive increases in the brightness of the subsequent return stroke that occurs when the forks are illuminated are called *M-components* (71, pp. 24–26). More reliable maximum values for the total numbers of strokes in a flash are 17 (139), 18 (127), and 26 (140). Operationally, the NLDN limits the number of strokes per flash to 15, provided that they all occur within 1 s and 10 km of the first return stroke (13). The global mean value of strokes per ground flash is about 3.5 (141), which is about twice the number detected by the NLDN (12,13).

Not all flashes have only discrete episodes of current flow, however, because some channels may remain attached to an object such as a tree on the ground for hundreds of milliseconds. Such *continuing currents* can maintain flows of hundreds of amperes, comparable to those of an arc welder, and these may be enough to bring the object to a combustible temperature (22, pp. 106–107). Between one-quarter and one-half of all cloud–ground strokes contain continuing currents (74, p. 14) that generate a substantial amount of energy in the extremely low frequency (ELF, less than 300 Hz) range; in one study, 60% of single-stroke and 50% of multiple-stroke flashes had long continuing currents (140). Indeed, M-components are impulsive brightenings of the continuing current channel (22, p. 107; 74, pp. 14, 17). The fact that continuing currents can initiate combustion provided the original rationale for creating a lightning location network using VLF/LF techniques for the timely detection of forest fires (11).

Return Stroke Types

Cloud–ground strokes need not effectively carry negative charge to ground and have stepped leaders originating in the cloud. Berger (142) notes that there are four types of cloud–ground strokes, although Mazur and Ruhnke (143) propose a modification of Berger's types based on the argument that a return stroke can move only from the ground upward to the cloud (22, p. 97). Type 1 strokes originate in the cloud as negatively charged stepped leaders that effectively carry negative charge to ground and are called *negative return strokes*; these generate electric fields whose initial peaks have *positive polarity* (Figs. 6 and 8 earlier and Fig. 10) (144). Globally, more than 90% of all cloud–ground strokes are of this negative type (74, p. 8).

Positively charged stepped leaders originating in a cloud are also possible, leading to Type 3 *positive return strokes* that effectively lower positive charge to ground. Such strokes often come from the anvil that spreads downwind of the main part of the thunderstorm at heights of 10 km above the earth's surface (145,146). Positive strokes also appear with greater relative frequency in highly sheared and/or shallower thunderstorms, such as occur more frequently in winter. In the spring of 1998, a significantly greater relative number of positive return strokes were also observed in the Central Plains of the United States when numerous biomass fires burning in Central America produced aerosol particles that were advected over the United States (147). Although the vast majority of return strokes effectively carry negative charge to ground as part of negative flashes, an increase in the relative number of positive flashes is a signature of severe thunderstorms (22, p. 253). Although Berger et al. (131) found that the median value of peak current for positive and first-return negative strokes is about the same, 30 kA, positive return strokes have a greater number of peak current values that are substantially larger than the median value. As seen in Figs. 11 and 12 of (13), the distribution of peak currents measured by the NLDN

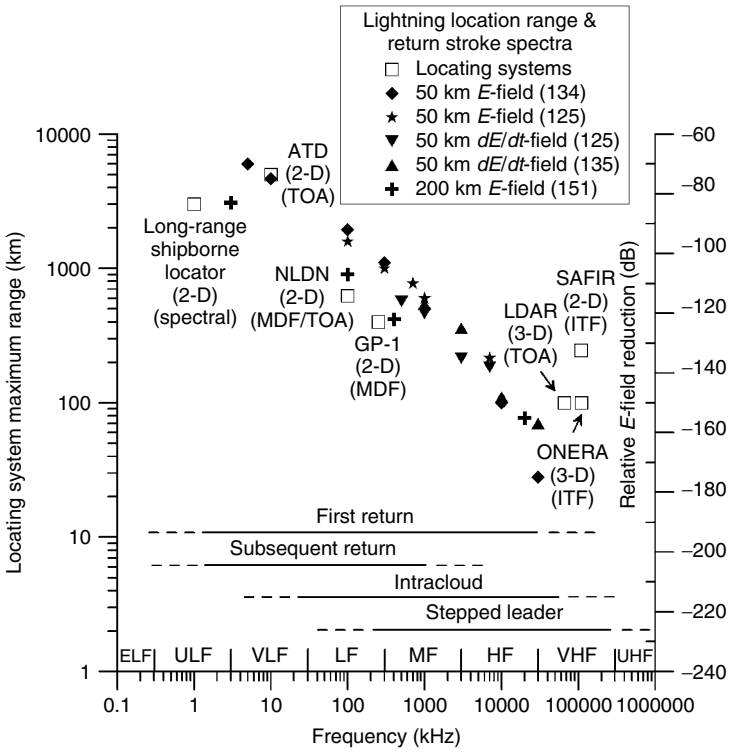


Figure 11. Maximum lightning location range for a few selected locating systems (Table 2) spanning ultralow to very high frequencies (ULF to VHF) (plan view locator, (2-d); channel mapper, (3-d); interferometry, (ITF); magnetic direction finding, (MDF); time-of-arrival, (TOA); ratio of spectral amplitudes, (spectral)), together with the frequency spectra for first-return strokes as given by measurements using wideband (MDF) systems (125, 134, 135, 151). The intermediate frequencies are chosen for the wideband observing systems. Note that systems using lower frequencies have a greater detection range. The return-stroke spectra are expressed in terms of decibels (dB), which logarithmically quantify the E-field reduction relative to the standard of 1 V/m or the dE/dt-field reduction relative to the standard of 1 V s/m. From a peak value near 5 kHz, the spectral amplitudes fall f^{-1} as through frequencies of 1 MHz, but fall more rapidly as f^{-2} for frequencies greater than 1 MHz. Note that the decrease in detection range for the lightning location systems roughly follows the decrease of the spectral amplitudes. Also shown via the horizontal bars in the lower part of the figure is a rough approximation of the frequency content of first and subsequent return strokes, intracloud (including cloud–cloud and cloud–air) strokes, and stepped leader pulses. Lower frequency systems sample the relatively long return strokes well but must discriminate them from intracloud strokes. In contrast, higher frequency systems sample the relatively short stepped leader pulses well but primarily infer the location of return and intracloud strokes from the channel orientation that is given by a coherent sequence of these short-duration pulses.

is more strongly skewed to larger current magnitudes (*right-skewed*) than the distribution of peak currents from negative return strokes. This more strongly skewed distribution is consistent with the origin of many positive return strokes in the anvil, whose lightning channels are much longer than those for negative return strokes that begin typically in the middle portion of the cloud (101).

The other two types of cloud–ground flashes defined in (142)—Types 2 and 4—are relatively rare. They involve leaders that begin at the earth's surface and propagate upward toward the cloud; generally, cloud–ground strokes of this type are initiated at mountaintops or tall buildings. According to Mazur and Ruhnke (143) and summarized by MacGorman and Rust (22, p. 97), Type 2 strokes are characterized initially by a positive leader that moves upward into the cloud. It is followed by a dart leader that propagates from cloud to ground and that initiates the return stroke which effectively lowers negative charge to ground. Type 4 strokes begin via a descending positive leader that initiates a negative leader which moves upward from the ground. When these two leaders meet, a return stroke that effectively lowers positive charge to ground occurs.

SAMPLING CONSIDERATIONS FOR LIGHTNING LOCATION SYSTEMS

The different types of lightning location systems introduced earlier exploit the different aspects of the lightning discharge process that are reviewed in the previous sections. These systems provide various information about

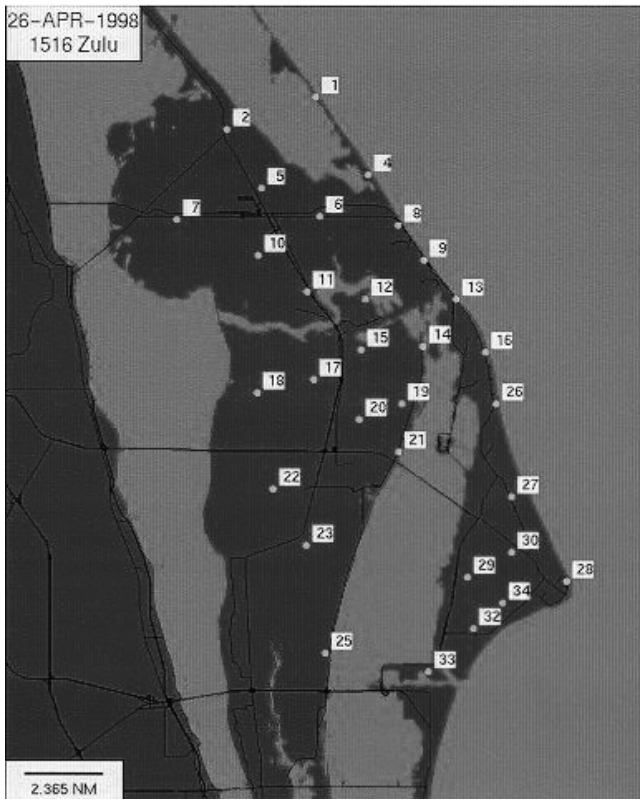


Figure 12. Map of the 31 field mill locations used in the Launch Pad Lightning Warning System (LPLWS) at the Kennedy Space Center. (Courtesy USAF, 45th Weather Squadron, used with permission.) See color insert.

the discharge, such as the location of the ground strike point or the orientation of the lightning channel, and provide vastly different stroke classification and ranging capabilities. Throughout the discussion here, emphasis is placed on application of the radio-frequency characteristics of lightning.

Stroke Classification

The electromagnetic field frequency spectrum of first cloud–ground strokes is quite broad and spans frequencies from at least 1 kHz to 10 GHz (136, p. 196), over which the signal level, or magnitude, normalized to the 10-km range, decreases as frequency increases, from a peak typically near 5 kHz (134) to a value seven orders of magnitude smaller. The peak in the spectrum is in the VLF portion of the radio-frequency spectrum; thus the origin of the common statement that "VLF sferics are produced predominantly by return strokes" (22, p. 249). One must be careful when applying this statement, however, because bipolar waveforms associated with intracloud strokes are commonly observed when using VLF systems (Fig. 7), and so, as discussed earlier, when detection of cloud–ground strokes is desired, time-domain waveform analysis must be performed to separate the cloud–ground and intracloud (including cloud–cloud and cloud–air) signals (11) (Figs. 6 and 7).

The wideband character of the electromagnetic radiation in first return strokes is caused by branching effects commonly seen in their channels (125,138). Thus network locating (NL) systems such as the NLDN and CGLSS that focus on cloud–ground strokes (see later) use frequencies predominantly in the VLF/LF range that detect cloud–ground lightning in the far field using the peak signal radiated near the surface where the channel is predominantly vertical. In contrast, network lightning mapping (NM) systems such as LDAR, LMA/LMS and SAFIR/ONERA-3D, which locate the lightning channel in three dimensions by tracking the position of negative-polarity streamer/leader transitions, stepped leaders, dart leaders, or other in-cloud breakdown processes (24,27,29,86,137), use VHF frequencies higher than 30 MHz (Table 2) (see later). Thus, VHF ground-based systems do not detect cloud–ground strokes or intracloud strokes directly, because there is little VHF radiation emitted by current flow in existing channels (137), but instead provide the stroke classification by tracking the formation of the channel given by a large number of detections of a coherent sequence of extremely short pulses (24,27,29,32,33,86). Recent work on data from the satellite-mounted FORTÉ VHF and optical sensors suggests, however, that cloud–ground and intracloud flashes can be distinguished from their optical waveforms and their properties in VHF spectrograms (50).

The electromagnetic fields produced by lightning propagate by line of sight in the VHF (see earlier). Consequently, ground-based VHF mapping systems cannot identify the strike point on the surface directly because intervening ground due to the earth's curvature or other obstructions blocks the signal, whereas VLF systems can. Thus, the VLF/LF CGLSS network supplements the LDAR network to provide such points (16), and VLF/LF receivers are used in the SAFIR network for the same purpose.

Stroke Ranging

The effective range of a locating system is related to the portion of the electromagnetic spectrum sampled. As noted before, the radiated power drops by several orders of magnitude between the VLF peak near 5 kHz and that measured at microwave frequencies near 10 GHz. Such drops in power are often described by using power ratios that are conveniently expressed in *decibels* (*dB*), where 1 dB is 10 times the base 10 logarithm of the ratio of the radiated powers. Because radiated power varies as the square of the electric field intensity, 1 dB is 20 times the base 10 logarithm of the ratio of the electric field E, or equivalently the magnetic field B, intensities. A common reference E-field intensity is 1 V m^{-1}, and for its temporal derivative dE/dt, the common reference is 1 V m^{-1} s^{-1} (125). Using decibels defined by these reference values allows comparing both the E and dE/dt spectra on one figure and so extends the frequency range over which the lightning properties can be characterized. For example, the E-field spectra are used by Weidman et al. (125) to characterize the frequency content of return strokes from frequencies of 100 kHz through 1 MHz, and the dE/dt spectra are used for frequencies in the range of 1 to 20 MHz (Fig. 11).

For lightning sources range-normalized to a 50-km distance, a typical spectral value of the E-field peak occurs at frequencies near 5 kHz and has dB values near −70 based on the reference intensity of 1 V m^{-1} (134). At a 50-km range, the broadband signal generated by a cloud–ground stroke varies by more than 80 dB between the VLF peak and the HF signal level at 10 MHz (125,134,135). Although a 20-dB decrease represents an order of magnitude drop in intensity, an 80-dB decrease represents a signal level only 0.01% of the magnitude of the peak level. Stepped leader waves that have average peak signal levels that are 10% of those of the following return strokes (8) thus have peak signal levels that are about 20 dB lower than those of the return strokes (125). The stepped leader waves in Fig. 8 have peak levels about 10% of the 11×10^{-8} weber per square meter Wb/m^2 (denoted) peak level of the return stroke; these steps are spaced at 20 to 25 μs, which is consistent with a medium frequency (MF, 300 kHz to 3 MHz) of 400 to 500 kHz and which is on the lower end of the frequency band for negative stepped leader waves given in Fig. 11. From this figure, it is apparent that the power at these frequencies is about 20 dB below that for return strokes whose frequencies are near 100 kHz, at which the NLDN samples. For frequencies greater than 50 MHz used by LDAR and SAFIR, much lower signal levels than those associated with the MF frequencies of the stepped leader step peaks would be expected. In the frequency range of 100 kHz to 1 MHz, the power at all frequencies for intracloud electromagnetic waves is typically about 10 dB lower than that for cloud–ground waves, indicating a significantly weaker signal level (125).

Subsequent return strokes in the flash generally do not have branches and, as noted in Fig. 11, also produce

a small signal level above 1 MHz (74, p. 119). Although VLF/LF systems such as the NLDN and CGLSS can detect these subsequent return strokes, VHF interferometric systems such as SAFIR can provide some information about them by detecting the preceding $K-$processes and dart leaders (86). In general, the mean E-fields produced by subsequent strokes are one-half to three-quarters of those produced by the first strokes in a flash [Table 7.1 in (74) and (144,148,149)], and so only the peak magnitude of the first strokes should be used for distance ranging. Finally, continuing currents in subsequent return strokes emit a substantial amount of ELF radiation at frequencies less than 300 Hz (5, pp. 362–363), and so can be detected at a great distance. For example, a shipborne system that uses the spectral components of received signals over ELF and VLF frequencies can provide ranging between 1,000 and 3,000 km (58). Additionally, measurements can be made on a truly global scale by applying magnetic direction finding techniques to received signal levels in the ULF/ELF transition range of 4 to 200 Hz (150).

The received signal level is affected by the attenuation rate over the distance propagated and also by the frequency-dependent propagation characteristics of the atmosphere itself (5, p. 352). At frequencies less than 300 kHz, which are primarily used by plan view SSC and SSL locating systems at ranges >100 km (Tables 1 and 2), the electromagnetic wave is a ground wave (see earlier). To a good approximation, the amplitude of this ground wave varies inversely with the distance propagated, and in the VLF this provides the basis for lightning detection and ranging at distances up to several hundred kilometers (133). For example, as seen by comparing the stars and pluses in Fig. 11, the signal is decreased by 12 dB, or four times, and also four times over a range that varies between 50 and 200 km (134,151). Using signals in frequencies between 300 kHz and 30 MHz can produce ambiguity because they may arrive at a ground-based receiver by multiple paths due to both ground and sky waves; for this reason, locating system receivers predominantly avoid this range. Very high frequency/ultrahigh frequency (VHF/UHF, 30 MHz to 3,000 MHz) signals penetrate the ionosphere, removing the multipath ambiguity, but the line-of-sight propagation limits ground-based receivers to ranging of only local storms (5, p. 367). Satellite-based systems, however, can use VHF signals to great advantage (49–53).

Taken together, the effects of radiated power and signal attenuation imply that the desired range to the lightning controls which radio-frequency band should be used by a ground-based system (Fig. 11). To detect lightning on the large scale of several thousand kilometers, one should choose receivers sampling predominantly in the VLF region near the 5-kHz peak in the spectrum, as is done by the British ATD system (17,18) (Tables 1 and 2) and by Füllekrug (150), who used simultaneous global observations of the earth–ionosphere cavity, or Schumann, resonances. The shipborne ELF/VLF system of Rafalsky et al. (58) combines signal levels whose frequencies are less than 1 kHz with those between 2 and 23 kHz to detect lightning in the range of 1,000 to 3,000 km. The broadband MDF and TOA receivers used by the NLDN, CGLSS, GPATS, and GP-1 systems measure signal levels predominantly in the VLF and LF ranges and so detect lightning up to distances of 400 to 600 km (13). The effectiveness of VLF/LF methods becomes degraded at close ranges, within 100 km or so, in part because stepped leaders in intense lightning may lead to premature triggering of the system, as may well happen in the example in Fig. 8, and in part because near-field components of the electromagnetic radiation are no longer negligible (5, p. 353). The LDAR 6-MHz bandwidth system uses signal levels centered on 66 MHz, and the LMA/LMS 6-MHz bandwidth system uses signal levels centered on 60 MHz, to map the lightning channel given by negative leader steps associated with discrete pulsed breakdown processes at ranges up to 100 to 150 km or so. In contrast, the SAFIR/ONERA-3D narrowband VHF ITF systems sample more continuous bursts of radio noise from most phases of a lightning discharge at frequencies between 100 and 118 MHz using 1-MHz bandwidths (24,27,86,137) up to ranges of 100 to 400 km (22, p. 157; 65), depending on the elevation and separation of the receiving stations in the network.

Note from Fig. 11 that the decrease in maximum range, as the ground-based receiver frequency increases, closely follows the decrease in lightning electric field spectral magnitude, as frequency increases. This decrease is in part a direct result of the relatively uniform attenuation rate of the signal with propagation range over land. The weaker the original signal, the shorter the distance it can propagate and still be discernible from sources other than lightning. In theory, the radiation or far-field component of the electric field of a thunderstorm decays directly with range r (5, p. 353; 106, p. 545).

In practice, however, the relationship between r and the received signal level is a bit more complicated and may be given by a signal propagation model of the form

$$y = \left(\frac{r}{D}\right)^{-p}, \tag{2}$$

in which y represents the received signal level for an individual stroke, as for the NLDN (13) (see later), and D is a site- or bearing-dependent parameter. For single-station SSL systems, y is a statistical parameter, such as the median, that describes an aspect of the distribution of strokes in a thunderstorm over a period of time, such as 30 minutes (56,57) (see later). Doing so assumes that the set of strokes that composes the distribution originates from a region at a well-defined distance from the sensor. If there are multiple regions of lightning, however, then the range estimate given by this simple model will correspond to an average distance to these regions, weighted by the number of detected strokes from each region.

A 12-dB decrease in the received signal levels for lightning at the 50- and 200-km ranges is consistent with the theoretical value of 1 for the attenuation rate coefficient p in Eq. (2) (151). Generally, however, $p > 1$ is found due to *site error* effects that are related, in

part, to reradiation of the lightning signals by man-made structures, to topographic variations along the propagation path (3;22, p. 160), and to distortion of the magnetic field in the vicinity of the direction-finding antenna (152). For example, the NLDN uses $p = 1.13$ in a more complicated version of Eq. (2), which was determined by Orville (132) to estimate the peak current I_{peak} from the peak level of the received signal (see later). In an earlier study of NLDN observations, Idone et al. (133) found $p = 1.09$ for a power law of the form of Eq. (2); in contrast, Mazeroski et al. (56) found bearing-dependent values of $p < 1$ for power laws of this form for the GP-1 SSL installation at Lawrence, Kansas (see later).

As noted earlier, accurate location of the source of lightning depends on accurate determination of the bearing. Although broadband magnetic direction finders can, in principle, provide accuracies of a degree or less (7), site errors also can significantly affect the bearing calculated at one receiving site. Because these bearing errors generally do not change with time, they can be accounted for during calibration of the instrument and corrections can be applied before the signal propagation model, Eq. (2), is used for ranging (56,152). Multistation networks use triangulation or time-of-arrival methods to obtain a best position by finding the location that minimizes the variability in the estimates given by pairs of receiving stations (22, pp. 160–162) (see earlier). The best estimates are obtained by combining two independent estimates via least-squares optimization, as is done by the Improved Accuracy from Combined Technology (IMPACT) sensors used in the NLDN network (13) (see later).

A description of several examples of real-time, operational lightning detection, direction finding, and locating systems follows in the next sections. Of the many systems summarized in Tables 1 and 2, those chosen for description sample different characteristics of the lightning signal to obtain information tailored to the particular applications of that system. The accuracies and ranges vary with each system and are constrained in part by the particular lightning characteristics sampled. Some systems use more than one characteristic, and this strategy generally improves the location or detection accuracy noticeably.

THE LAUNCH PAD LIGHTNING WARNING SYSTEM: A LIGHTNING WARNING DEVICE

The United States Air Force (USAF) 45th Weather Squadron uses a network of 31 electric field mills, over an area of roughly 20 × 20 km, to measure the electric field potential gradient at the surface, in support of the space program at Cape Canaveral Air Force Station (CCAFS) and Kennedy Space Center (KSC). This network of field mills, communication, analysis, and display equipment is known as the *Launch Pad Lightning Warning System* (LPLWS). A map of the field mill locations is shown in Fig. 12. The electric field potential gradient is sampled at 50 Hz, and 1-minute averages are displayed and isoplethed in kV/m (Fig. 13). The electric field potential gradient has the same magnitude as the electric field but has the opposite sign. Thus LPLWS displays the fair-weather field as about +200 V/m, but is positive. Strip charts that show the time series of the field mills are also available.

The primary purpose of the LPLWS is to help evaluate the lightning launch commit criteria (153) to guard against rocket-triggered lightning or to determine if potentially hazardous conditions are safe. The LPLWS also helps in lightning forecasting: the building electric fields from locally developing thunderstorms provide some warning of the onset of lightning. The decaying phase of the thunderstorm can sometimes be seen on the field mill display, especially in the slowing of the lightning rate on the strip charts, and the end-of-storm oscillation (92,109) can sometimes be detected on the strip charts to help forecast the end of lightning. In addition, the electric charge in anvil and debris clouds and their lightning threat can be detected, although screening layers at the thunderstorm tops must be considered. The two cases of anvil cloud and debris cloud lightning are especially difficult forecast challenges. Although much less frequent than lightning from cellular thunderstorms, cloud–ground lightning from an anvil can be surprising, given the long distances it can strike from the parent thunderstorm. Cloud–ground lightning from anvil clouds has been documented more than 90 km from the parent thunderstorm, and in-cloud lightning more than 130 km. Lightning from a debris cloud is even less frequent than that from an anvil cloud but can be surprising, given the long time it can strike after the parent thunderstorm has dissipated. Cloud–ground lightning more than $1\frac{1}{4}$ hours after the last flash in the parent thunderstorm has been documented.

The LPLWS also provides lightning detection and has limited lightning location capability. The LPLWS detects and locates all types of lightning, including lightning aloft. If a field mill changes by 200 V/m or more, positively or negatively, between 50 Hz measurements, then the LPLWS defines that a lightning event has occurred and activates its lightning algorithm. Every field mill that detected the instantaneous pulse, regardless of magnitude, is used to isopleth the overall electric field change. Note that because the isopleth maximum is used, the location does not necessarily correspond to the field mill reporting the largest change; that is, the interpolation scheme can produce greater electric field changes not collocated with the field mills themselves. The location of the maximum field change, is determined by the isoplething, is taken to represent the center of mass of total change in the electric charge in the lightning flash, that is, the middle of the lightning flash projected onto the ground. A 32 × 32 one-kilometer grid is used to plot the inferred position of the lightning flash. Letters in alphabetical sequence are used to plot the lightning; the first lightning flash is plotted as "A," the next as "B," and so forth (Fig. 14). The lightning is plotted at the closest grid point, not the actual maximum in isopleths. This approach introduces an average positioning error of 354 m for purely vertical lightning, and an unknowable range of 0 to 707 m. But a much greater location problem occurs for lightning that has any horizontal extent. Consider a flash of anvil

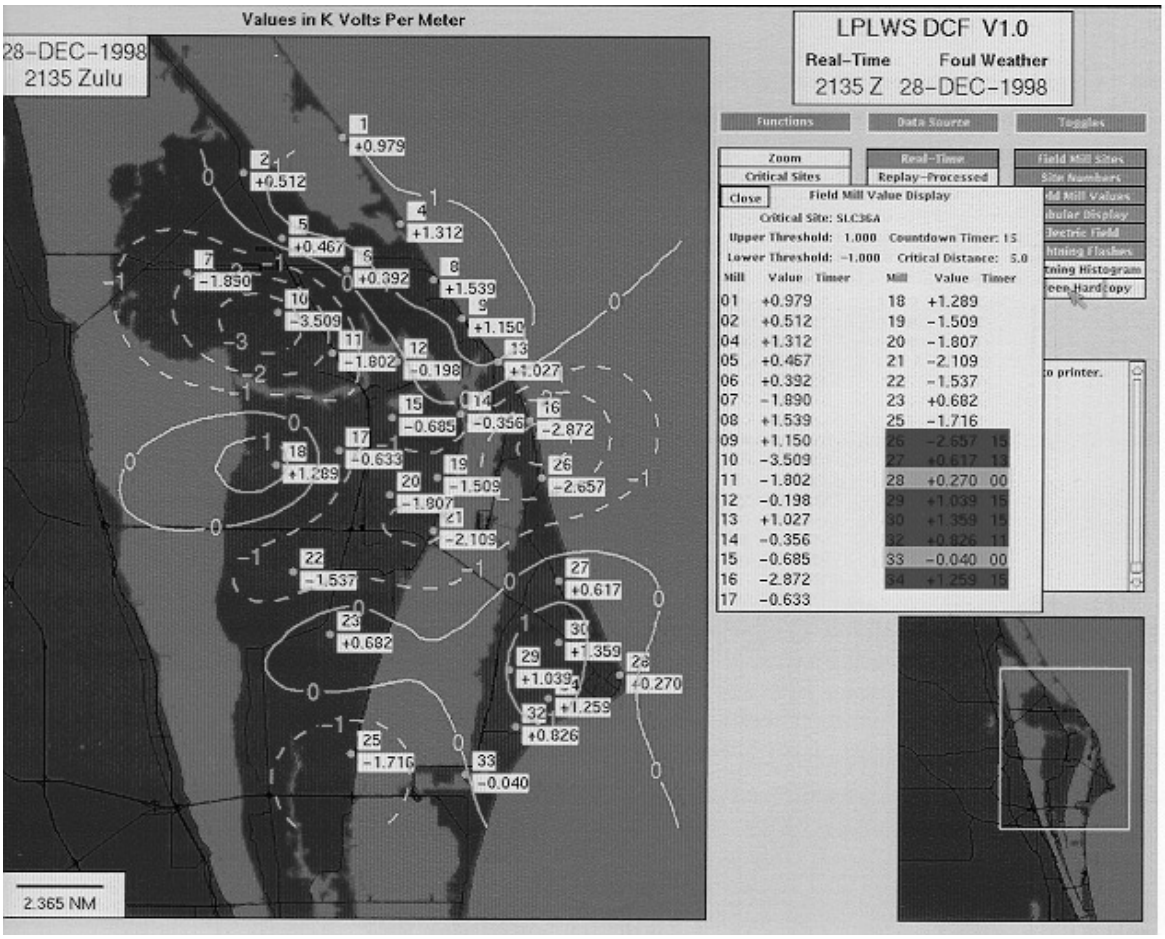


Figure 13. LPLWS example of a lightning event at 21 : 35 UTC 28 December 1998. The electric field potential gradient is sampled at 50 Hz, and one-minute averaged electric field potential gradient (kV/m) values are isoplethed. Fair-weather field values are displayed as +0.2 kV/m because the electric field potential gradient and the electric field have opposite signs. The field mill locations are denoted by the integers and their observations by decimal values. (Courtesy USAF, 45th Weather Squadron, used with permission.) See color insert.

lightning at a 9-km height, more than 50-km long, over the middle of the network. LPLWS will plot that flash as a single point in the middle of the network. The three-dimensional nature of the lightning is completely lost; its altitude and horizontal extent are unknown. The plotting grid also displays strong lightning outside the LPLWS network erroneously on the edge of the network. LPLWS was meant to detect and locate lightning only over the field mill network but typically detects lightning as far as ~20 km outside the network. Fortunately, users of LPLWS are accustomed to this problem and know how to discount the erroneously plotted lightning locations.

The LPLWS is a reasonably good detector of all types of lightning, and has a detection efficiency of 90%. But its location accuracy is considered poor. Since the implementation of the Lightning Detection and Ranging system (see later) and other lightning systems at CCAFS/KSC, the lightning detection and location features of LPLWS have been considered merely a third backup capability by the 45th Weather Squadron.

ELECTRICAL STORM IDENTIFICATION DEVICE: A RADIO-OPTICAL FLASH COUNTER

The Electrical Storm Identification Device (ESID) and the similar ASOS (Automated Surface Observing System) Lightning Sensor (ALS) developed by Global Atmospherics, Inc. (GAI) (Tables 1 and 2) are two of the most widely used single-station flash counter (SSC) systems. The system has been specifically designed for ease of installation and operation to protect people and assets. The ESID is a portable, lightweight unit that measures two feet high and one foot in diameter, and it contains a solar-powered sensor as well as a battery backup within the display unit to ensure uninterrupted operation of the system during power outages. The ESID has appealed to a broad group of users, including private companies, individuals, and the United States Golf Association (USGA) for real-time use in its golf championships. The Thunderstorm Sensor Series (TSS) versions of the ESID have undergone extensive testing for certification on ASOS platforms (154), and the results have

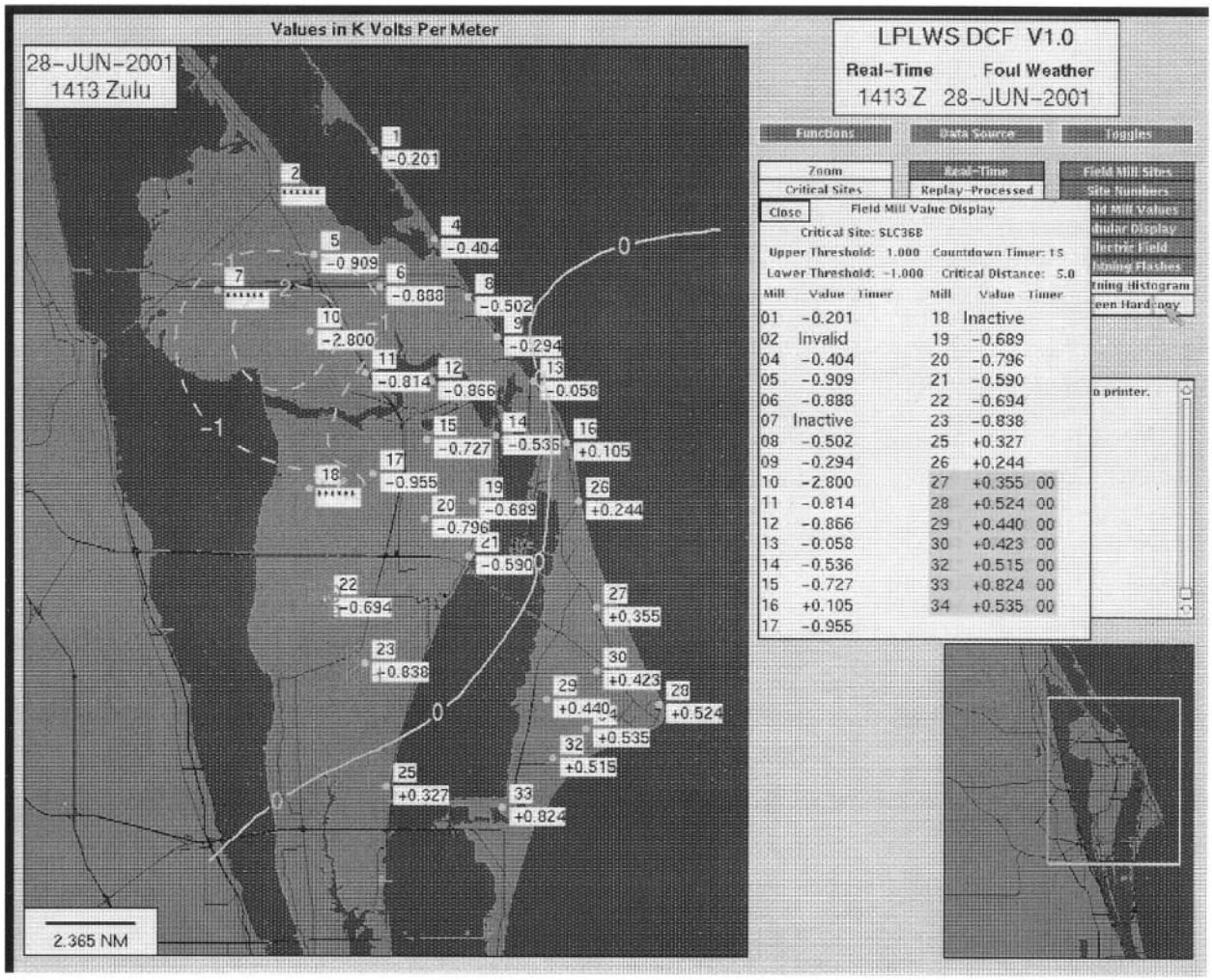


Figure 14. Sample of LPLWS display at 14 : 13 UTC 28 June 2001. The red letters show where LPLWS infers the center of a lightning flash of all types projected onto the ground. One-minute averaged electric field potential gradient (kV/m) values are isoplethed, and dashed curves indicate negative values. The field mill locations are denoted by the integers and their observations by the decimal values (kV/m). (Courtesy USAF, 45th Weather Squadron, used with permission.) See color insert.

been published in reports available from the ASOS Program Office of the United States National Weather Service (155).

Lightning Signal Discrimination

Several components of the ESID enable sensing both return and intracloud (including cloud–cloud and cloud–air) lightning strokes (54). Both radio and optical signals from a stroke must be received for lightning detection, a requirement that greatly reduces the *false alarm rate*, the ratio of the number of incorrect detections to the total number of detections (154;156, pp. 240–241). The device is omnidirectional and so cannot resolve the bearing of a given stroke. Rather, emphasis is placed on range estimation of return strokes. Range estimation of intracloud strokes, however, is not available from the earlier TSS520 model ESID; the newer model TSS924 is used on ASOS platforms as the ALS. Because the ESID is a single-point sensor, range cannot be estimated by triangulation but is derived from an analysis of the optical and radio signals generated by the stroke.

First, the optical sensor on the ESID detects and processes the waveforms from the light emitted by a lightning stroke (154). When compared with the inverse-range attenuation rate $1/r$ of the electric field (see earlier) (5, p. 353; 106, p. 545), the optical field attenuates more rapidly with range r and obeys the inverse-squared relationship in which the acquired optical signal level is proportional to $1/r^2$. This attenuation rate limits the range of the ESID to line-of-sight measurements when the atmospheric visibility is not seriously degraded by rainfall. The system response also degrades as the flash rate increases owing to *processor dead time*, the time when the system is analyzing a recently received signal and so cannot detect another (154).

Second, the instrument senses the electric field. Because the ESID detects lightning at close range, both

the near-field, or electrostatic, and far-field, or radiation, components of the emitted electromagnetic radiation must be considered (73, pp. 61–63). To do so, the decay time constant is changed to allow the antenna to measure both of these components of the electric field. The electrostatic component has lower temporal frequency, of the order of milliseconds to seconds, and essentially reproduces the total field change from an entire flash (22, pp. 106–107). When the decay time constant of the circuit signal is reduced and forces the field signal level to decay faster, the electric field waveform, or radiation component, is measured (75, p. 191) (see earlier section). The *E*-field antenna now serves as a fast antenna, of the type used in combination with gated, magnetic direction finders (5,11;117, p. 221), which allows observation of the higher frequency components to distinguish the type of flash using time-domain waveform analysis (see earlier section). The simultaneous sensing of both components of the *E*-field by one antenna allows the ESID to acquire the rapidly occurring field variations of a complete lightning flash with adequate time resolution.

The ESID registers a lightning stroke *only* if the optical threshold is exceeded within 100 μs of the electric field threshold time. The patented signal processing circuits in the ESID, which reject background noise, also aid in virtually eradicating false alarms for properly installed sensors, as affirmed by the manufacturer, GAI (157).

Once a return stroke is detected, its range is then estimated by data-processing electronics within the unit, and data are transmitted to the display unit via a fiber-optic communications link. Range estimates are given in one of three intervals: 0–5, 5–10, and 10–30 miles (0–8, 8–16, 16–50 km) from the device. The ESID can also automatically issue audible and visual alarms according to a user-selected criterion. Flash count information is also displayed on the unit.

The capacity of the ESID to estimate the proximity of lightning, as well as its ability to withstand occasionally harsh outdoor weather elements, made it a viable choice as an operational lightning sensor in the early 1990s. During this time, a network of Automated Observing Weather Systems (ASOS) was being installed as the result of the modernization plan of the United States National Weather Service. These instruments, positioned at hundreds of airports across the country, relay critical surface weather observations to ground operations, eliminating the need for human observers. The ALS reports that a thunderstorm is in the vicinity of the station (using the designator *TS* in the coded message) if the sum of the cloud–ground flashes in a 10-statute-mile radius and cloud discharges at any range is two or greater. A flash cannot be older than 15 minutes to be included in this count. This so-called two-stroke algorithm (158) provides more reliable reports than simply requiring one stroke, as demonstrated in a performance test conducted by the Raytheon STX Corporation in the spring and summer of 1997.

Sensor Performance

Studies in the mid-1990s compared the performance of ESID with that of human observers and with that of magnetic direction finder networks that locate lightning via triangulation (see earlier section). A study at an airport in 1993 concluded that the ESID could detect a greater number of return strokes than humans because humans use thunder to detect lightning (154). Ambient aircraft noise may have occasionally muffled distant thunder that the human observer would have otherwise detected. In this early study, the ESID performed quite well with respect to both human and network detection, particularly within the first range-bin of 0 to 5 miles. The more extensive 1997 study by Raytheon STX used 10 separate installations and found that the two-stroke algorithm produced ESID thunderstorm reports 88% of the time that humans reported them and 89% of the time that the National Lightning Detection Network (NLDN) reported them. ESID reported 25% more thunderstorm minutes than humans and almost $2\frac{1}{2}$ times more minutes than the NLDN, whose event coverage was low during the study. Only 0.8% of the ESID-reported events were defined as false alarms because neither humans nor the NLDN reported lightning when an ESID did. Most of the events missed by the ESID were due to the conservative two-stroke requirement for logging a thunderstorm report. Although humans and ESID reported thunderstorms about the same number of times as the NLDN, there were many events when a human or an ESID detected lightning and the NLDN did not. This discrepancy is explained by the fact that the NLDN reports only cloud–ground strokes (see later), but humans and ESID report both cloud–ground and cloud discharges.

Tests such as these provide strong evidence that the ESID sensor provides a viable means for automatically detecting thunderstorms. Currently, 27 ASOS sites have such ALS detectors (158). An automated system that can report both range and direction in at least the eight octants remains a requirement of the FAA because this amount of detail is expected in the standard hourly reports of atmospheric conditions produced by any surface reporting station. A more advanced ESID, that can resolve the bearing of cloud–ground strokes and estimate the range of intracloud lightning strokes is planned.

To help meet this FAA requirement, an alternative method for reporting thunderstorms at ASOS sites is being implemented currently to supplement the observations from the ALS. Dubbed the Automated Lightning Detection and Reporting System (ALDARS), this approach reports when the NLDN detects cloud–ground lightning within 30 nautical miles of an ASOS site and reports the location to that site (158,159). As of 6 October 2000, 785 of the 886 ASOS sites had operational version 2.60 ALDARS installations; for the current status of these installations, see (160).

THE GREAT PLAINS-1: A SINGLE-STATION LOCATOR

The Great Plains-1 (GP-1) system is a gated, single-station magnetic direction finder of the type used in the National Lightning Detection Network (NLDN) (6,7,11,56,57,161) (see later). The GP-1 consists of the usual crossed-loop magnetic field (*B*-field) and electric field (*E*-field) antennas that detect the electromagnetic signal from a lightning

stroke. Time-domain waveform analysis is used to identify which received electromagnetic signals are from negative return strokes (see earlier section and Figs. 6 and 7). For each stroke, the bearing calculation is based on the peak levels *ew* and *ns* of the two components of the *B*-field signals (Fig. 2). The system estimates the locations of active areas of cloud–ground lightning using a signal propagation model of the form of Eq. (2) that relates measures of the received peak signal to the estimated range to that region. Of course, this approach can give reasonable results only when there is a single well-defined region of lightning in the bearing of interest.

Signal Propagation Model

Because a typical distribution of lightning peak signal levels, or magnitudes, is right-skewed, and peak signal levels from a given storm often vary by an order of magnitude (13,162), a viable signal propagation model must be based on a set of observations from a region of cloud–ground lightning rather than on the peak signal level from an individual stroke (56,57). To relate a given set of observations to the range to the source, the changes in the shape of the lightning peak signal distribution must be known as the range changes. Once this has been done, lightning range is estimated by matching the observed distribution over a fixed time interval, such as 30 minutes, to a reference distribution of known distance from the receiver. Such reference distributions are found by analyzing a *developmental database*, in which the GP-1 observations in each 1° bearing are sorted into range-bins; these are sufficiently wide distance intervals to provide an adequate sample for an entire lightning season. The ranges to the lightning are provided by the NLDN observations for the same time and region. Separate range-estimating equations of the form of Eq. (2) are needed for each 1° bearing because site errors lead to varying signal attenuation rates that are produced when the electromagnetic waves propagate over surfaces that have different conductivities or varying topographic relief (3;22, p. 160) or when the magnetic field becomes distorted in the vicinity of the direction-finding antenna (152). Site errors also are manifested as bearing errors that must be corrected using a two-cycle sinusoidal function determined for each GP-1 site (152).

Signal propagation models for each 1° bearing are determined that relate the decile value from the lightning peak signal level distribution to the range. To find the 10 decile values $d_i, i = 1, \ldots, 10$, the set of observed peak signal levels over the sampling interval of interest is sorted from smallest to largest value (156, pp. 22–23). The first decile d_1 is the value in the list that is one-tenth of the way up from the first, smallest value. The fifth decile d_5 is the same as the median because it is the value at the midpoint in the list, and the tenth decile d_{10} is the largest value. Deciles conveniently standardize the description of the distribution for samples of varying size.

Deciles also provide a convenient, detailed way to express how the distribution changes with range. Such changes can be well approximated by the Weibull distribution, an idealization that has parameter values that can be varied to produce distributions ranging from Gaussian to highly skewed (38,162). To illustrate how the received peak signal level distributions vary with range, two possible Weibull distributions representing lightning peak signal levels from sources at two different ranges are shown in Fig. 15. Here the distribution of the *B*-field peak magnitude that represents the closer source (solid curve) has a broader shape and is shifted toward larger values than the distribution that represents the more distant source (dashed curve). These particular distributions are chosen based on the expectation that, owing to additional signal attenuation as range increases, the more distant case would yield fewer strokes that have relatively large peak levels but would yield more strokes that have relatively small peak levels. These two distributions are expressed as probability density functions, which are normalized so that the area under each function is 1. Here, the 10 decile values $d_i, i = 1, \ldots, 10$ are given by the

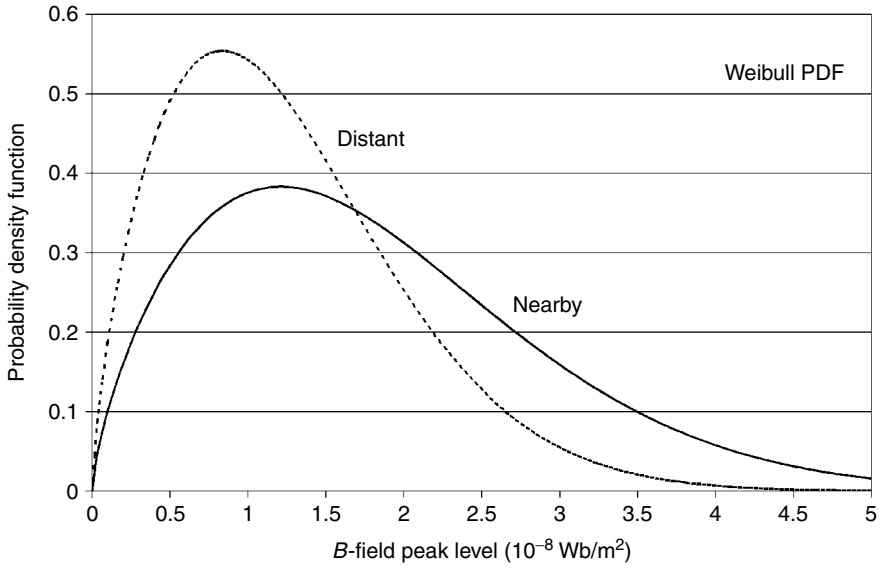


Figure 15. Two idealized distributions of received *B*-field peak signal levels given by the Weibull probability density function. Note that the more distant sources (dashed curve) produce fewer strong, but more weak, signals than the nearby sources (solid curve), as would be expected owing to greater attenuation of the original lightning signal as range increases. The cumulative distribution functions for these two cases are shown in Fig. 16.

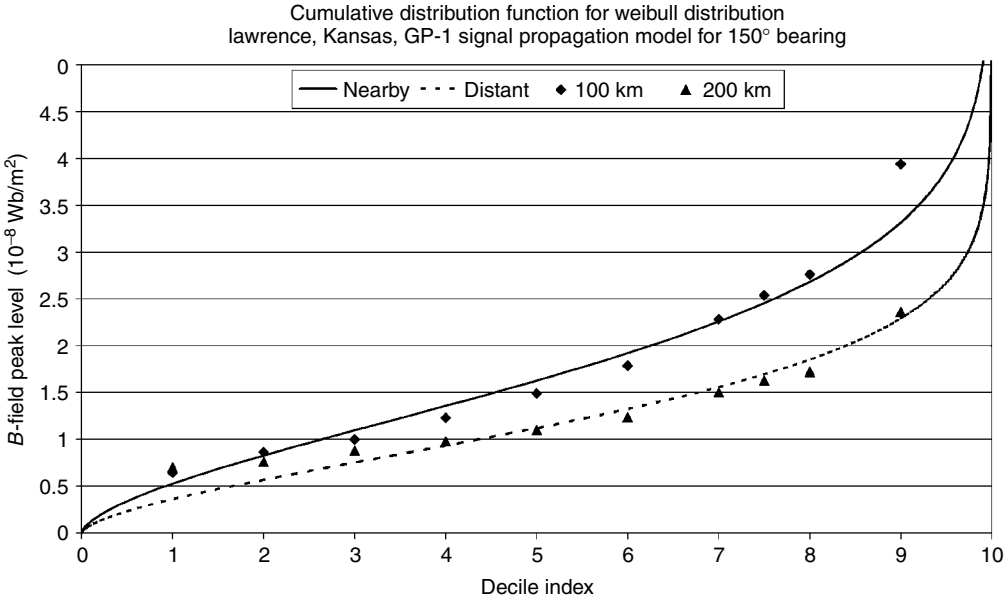


Figure 16. The cumulative distribution function for the B-field peak signal level as a function of the decile index i for each of the two Weibull distributions in Fig. 15 for the nearby (solid curve) and distant (dashed curve) cases. Also shown are the decile values d_i for two sets of points given by the signal propagation model at a bearing of 150° for the GP-1 locator at Lawrence, Kansas, one for the 100-km range (diamonds), and the other for the 200-km range (triangles). Note that both the slope and magnitude of the decile values are greater for the closer sources and that the cumulative distribution function from the Weibull distribution models the received peak signal level fairly well. Equations of the form of Eq. (2) model these variations better using coefficient D_i and exponent p_i values that are functions of the decile index i and bearing.

values of B for which the area under the curve between 0 and B is $i/10$. When these decile values are graphed as a function of the decile index i (Fig. 16), the curves are called *cumulative distribution functions*.

It is apparent from Fig. 16 that for each decile index i, the decile value d_i for the more distant stroke (dashed curve) is less than the decile value for the nearby stroke (solid curve). Owing to the different shapes of the distributions in Fig. 15, the difference between decile values increases with the decile index i and is manifested in Fig. 16 by the greater slope and magnitude of the curve for the nearby distribution that represents the closer source. Also shown in Fig. 16 are two sets of decile values for the signal propagation model at a bearing of 150°, one for a range of approximately 100 km from the GP-1 installation in Lawrence, Kansas (diamonds), and the other at a range of approximately 200 km (triangles). Note that the signal propagation model at each range follows the changes in magnitude and shape of each idealized distribution fairly well; in particular, the two models follow the expected trends as range increases. The deviations from the Weibull distribution are sufficient, however, to warrant use of the signal propagation model.

Once the decile variations are found for each bearing and range-bin, equations of the form of Eq. (2) are found but are rewritten as $r_i = D_i y_i^{-1/p_i}$. Here y_i is the decile value for index i obtained from the distribution of the received peak signal level, converted from V to Wb/m^2 by using Eq. (1) in (6), and r_i is the range estimate for that index, expressed in kilometers (161). The signal propagation model is expressed in terms of the values of D_i and p_i that are bearing- and decile-dependent, where $p_i < 2/3$ typically, and yield useful ranges between roughly 40 and 360 km if there is a sufficient sample over a reasonable time interval (56,57). How well the signal propagation model tracks the decrease of peak signal levels with range is expressed by the *coefficient of determination*, or squared correlation coefficient, R^2 (156, p. 167); the better the representation of the variation of decile value with range, the closer to 1 the value of R^2. For the GP-1 site in Lawrence, Kansas, the average value of R^2 is 0.78 across all bearings and decile values but is greater than 0.93 for deciles 6 through 9 that describe the upper half of the distribution. The first two decile values that represent the weakest peak signal levels that contribute to the distribution provide the poorest set of equations, and so their range estimates are often ignored. All of these results are consistent with the results in Fig. 16 that reveal an increasing separation between decile values as the decile index increases.

Estimating Range

Given that the rate of movement of thunderstorm areas is as high as 40 km/h, sampling periods of the order of 20 or 30 minutes must be used to determine the peak signal level distribution, so that the storms remain at approximately the same range. Although the signal propagation model is developed by using a large database of observations from an entire lightning season, it is applied to samples that are considerably smaller in size, varying from only 15 to

40 strokes per 1° bearing. For any given 20- to 30-minute sample containing thunderstorms in various portions of their life cycle, the resulting distribution is likely to deviate from that in the longer term developmental distribution that is uniformly composed of thunderstorms in all stages of their life cycles.

Therefore, to obtain the best possible range estimate, care must be taken to ensure that the small-sample distribution follows the form of a large-sample distribution. As can be seen from Fig. 16, the decile values from the long-term distributions represented by the triangles and diamonds each follow a cumulative distribution function that has a nonzero slope. In contrast, short-term distributions can have lengthy portions of zero slope, as illustrated by the 30-minute distribution denoted by diamonds and labeled "Duplicates" in Fig. 17. Those intervals of zero slope are caused by duplicate values in the distribution that overemphasize some portions of the long-term distribution for that range, which means that other portions are underemphasized. During some portions of the life cycle of a lightning-producing cell, duplicate peak signal levels are common and can be separated in time by many minutes; this means that they do not come from the same lightning flash (56,57). The distortion of the distribution due to retention of the duplicate peak signal levels can lead to range estimation errors that in some cases are unacceptably large (56). One way to adjust the short-term distribution and thereby obtain distribution estimates whose cumulative distribution functions have nonzero slopes is to eliminate the duplicate peak signal levels; *duplicate peak signal levels* are defined as those that have the same value within the resolution of the receiver. Experience has shown that when there are many duplicate peak signal levels to eliminate, doing so allows the sample distribution to replicate the long-term one better (Fig. 17) and consequently, to yield better ranging results.

When the ranging algorithm is implemented, a minimum of 11 samples in a bearing is required to have a sufficiently long sorted list that defines the decile values and so produces a potentially usable range estimate. Even when eliminating the duplicate peak signal levels to form a corrected distribution, not all decile values from the adjusted distribution will necessarily yield acceptable range estimates. Other quality control measures must be implemented to lower the uncertainty of the range estimate to an acceptable level. One such measure is to require that a decile value fall between its smallest and largest value in the developmental database for that 1° bearing before it is used to produce a range estimate from the equation $r_i = D_i y_i^{-1/p_i}$. These smallest and largest values

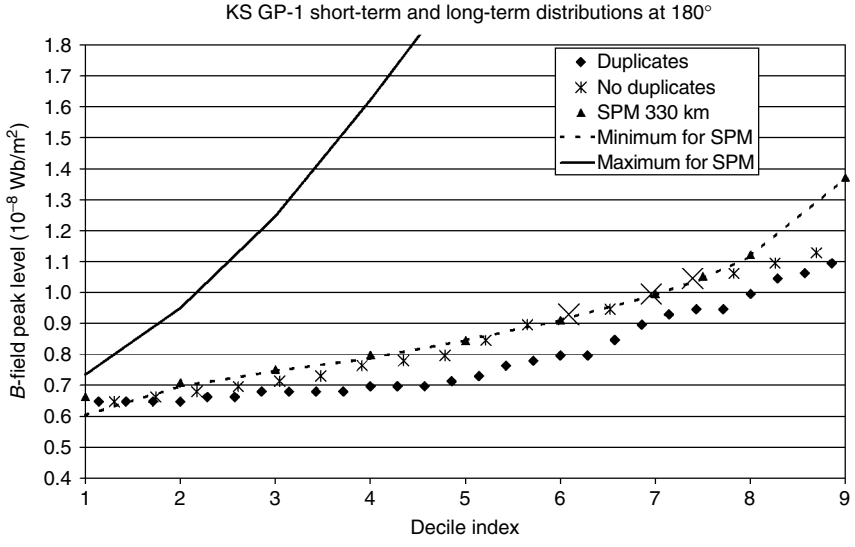


Figure 17. The decile values for the 30-minute period ending at 09 : 00 UTC 27 August 1999 at a bearing of 180 ° when all 35 samples are included in the sorted distribution (diamonds) and when all duplicates are removed, giving 23 samples in the sorted distribution (asterisks and Xs). Note that in contrast to the case including duplicate values, the case without duplicates follows the values from the signal propagation model quite well at a 330-km range (triangles), the range given by the average of all NLDN locations at the 180 ° bearing during this period. Clearly, removing the duplicates from the short-term distribution produces a better representation of the long-term distribution for the given range. The minimum values (dotted curve) and maximum values (solid curve) of the deciles that span the ranges between 40 and 360 km in the developmental database are also shown. A decile value from a distribution obtained during a 20- or 30-minute period is used to provide a range estimate from the corresponding signal propagation model power law of the form of Eq. (2) if that value falls within this envelope. All but one of the decile values for the case with duplicates fall outside this envelope and so are not used to give a range estimate, whereas three for the case without duplicates fall within the envelope (decile indices 6, 7, and 7.5, denoted by X). These three give the corresponding range estimates of 332, 315, and 333 km that are averaged to produce the estimated range of 327 km, which is very nearly the same value of 330 km that is given by the NLDN. (NLDN data used with permission of Global Atmospherics, Inc.)

are given by the dashed and solid curves in the example in Fig. 17, and they define an envelope within which a decile value must fall to be acceptable. Because each decile index i in a 1° bearing has its own model equation, as many as 10 independent range estimates r_i can be created from each short-term sample. Only those estimates given by acceptable deciles are averaged to produce the final estimate, and they can be updated with each new observation.

For the example based on a 20-minute sample given in Fig. 17, the actual range to the lightning at the 180° bearing was given as 330 km by the NLDN. At this range, the equations for the first and second deciles are not reliable because they have quite low R^2 values. Consequently, a fair comparison of the results is best done using the eight range estimates given by the third through ninth deciles and the third quartile, which has a decile index of 7.5. None of these decile values fall within the envelope of acceptable values when all observations are retained (diamonds in Fig. 17); but if used, then the eight range estimates r_i produce an average of 391 ± 50 km, a value considerably greater than the actual range of 330 km. Once the duplicates are removed from the short-term distribution, the range estimates provided by the eight decile values give a much improved average range of 336 ± 32 km, which is consistent with the fact that the cumulative distribution function for the distribution without duplicates (asterisks and Xs in Fig. 17) closely follows the cumulative distribution function given by the signal propagation model at a range of 330 km (triangles). If only the decile values that are within the envelope are used to calculate the average, in this case the sixth and seventh deciles and the third quartile (large Xs in Fig. 17), then the estimate is improved a bit further to 327 ± 10 km.

Across all 1° bearings, the mean absolute error in locating lightning at a range between 40 and 360 km from the receiver is reduced to 52 ± 14 km, once the following three quality control procedures are applied to the original short-term samples: the duplicate peak signal levels are eliminated, only the decile values that occur in their corresponding developmental database ranges are used to provide individual range estimates, and at least three of these acceptable range estimates must be found before the average of these preliminary estimates is computed to produce the final range estimate (56,57).

An example from 1 June 1999 is provided in Fig. 18, based on observations obtained by the GP-1 locator at Lawrence, Kansas, which is at the center of the map, as indicated by the black cross. The locations of negative return strokes south of Lawrence in southeastern Kansas (black dots), at a range of about 125 km from the receiver, are placed well when compared with NLDN observations (white dots). The ranges to the more distant strokes in western Kansas, Oklahoma, and Nebraska are not estimated because the GP-1 receiver acquires an insufficient number of high-quality signals. At these ranges, however, the system still functions well as a magnetic direction finder.

One significant limitation in applying the preceding algorithm comes from the assumption that all lightning at a bearing occurs in a well-defined region at a fixed distance from the GP-1 site. This assumption is often poor, because multiple regions of lightning often exist in a bearing. For example, parallel squall lines of thunderstorms separated by one- to two-hundred kilometers are often found ahead of cold fronts in the Great Plains; in addition, widely separated areas of isolated thunderstorms typically occur in warm, unstable air masses in the spring and summer. In such situations, the range to the lightning given by the signal propagation model will be a weighted average of the detections from the various regions. The resulting plots of estimated lightning locations will be unsatisfactory in these cases.

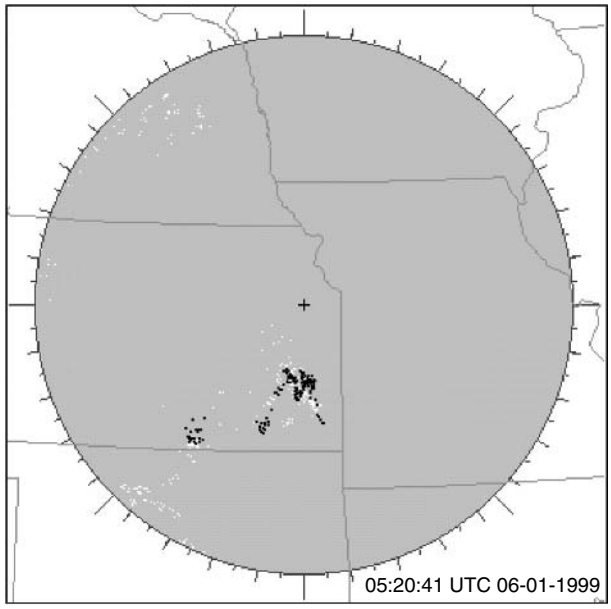


Figure 18. A typical screen display from the GP-1 locator in Lawrence, Kansas, based on a 20-minute sample. The locations given by the signal propagation model (black dots) agree well with those given by the NLDN (white dots) in a region south of Lawrence (black cross) on 1 June 1999 at 05 : 21 UTC. The NLDN data in Oklahoma, Nebraska, and western Kansas are beyond the range of the GP-1 installation. The GP-1 signal propagation model of the form of Eq. (2) generally provides good ranging between 100 and 300 km. (NLDN data used with permission of Global Atmospherics, Inc.)

SATELLITE-BASED SINGLE-STATION LOCATORS

In the mid- to late 1990s, two satellites carried instruments developed by the United States National Aeronautics and Space Administration (NASA) for detecting lightning from space using the visible light emitted by a lightning flash; one instrument, the Lightning Imaging Sensor (LIS), is still operational. The data from these satellites have been vital in establishing a consistent worldwide lightning climatology, especially over data-sparse areas such as oceans or sparsely populated land or developing countries. Although scattering from water droplets and ice crystals attenuates more light emitted by lightning channels in the lower portions of a cloud, the vast majority of lightning flashes originate in channels extending through the middle and upper portions of a cloud. Such bright flashes are detected easily from space (44,163). These two satellites are successors to the first optical instruments used for detecting lightning from space, the Operational

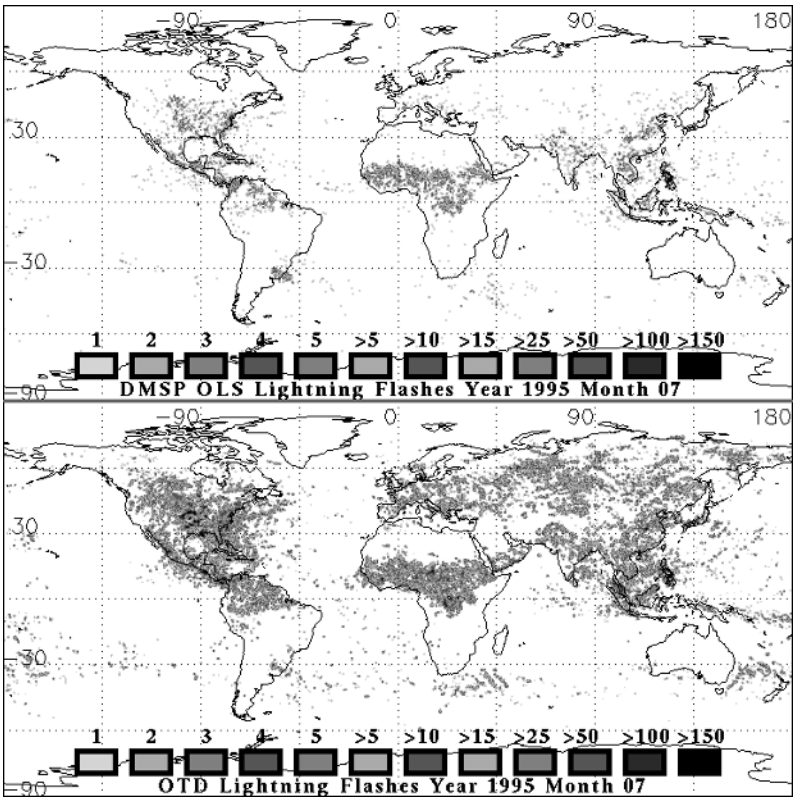


Figure 19. A comparison of lightning observations during July 1995 given by the Operational Linescan Sensor (OLS) and the newer Optical Transient Detector (OTD) (166). Each color represents the number of detections at each pixel during the entire month. The OLS measurements are obtained at local midnight by a human analyst who must manually determine the latitude, longitude, and time of each streak produced by a lightning flash on a visible image. The OTD observations at any location between 75° N and 75° S were obtained automatically twice in each 24-hour period, both day and night. Note that both sensors captured the fundamental fact that lightning is more likely to occur over land than over the ocean. Because these observations were taken during the Northern Hemisphere summer, the lightning flashes occur primarily in the tropics and Northern Hemisphere. (Figures provided by the NASA Lightning Imaging Sensor science team at Marshall Space Flight Center, AL, used with permission.) See color insert.

Linescan System Sensor (OLS) (43,164,165). The OLS could only distinguish the signature of lightning at nighttime, and thus lightning was only detectable during the local orbit crossings that occurred at local midnight or near dawn and dusk (44) (Fig. 19). The OLS sensor continues to fly on United States Defense Meteorological Satellite Program (DMSP) satellites. A lightning data archive derived from the smoothed 2.5-km resolution visible imagery filmstrips is available for some years from 1973 to 1996 at the Global Hydrology and Climate Center (166).

The NASA Optical Transient Detector (OTD) was the first optical instrument on a satellite, the Microlab-1, that could detect lightning during both day and night (45–47). The OTD was a rapidly developed prototype of the LIS, having a lower spatial resolution of 8 km at nadir from its altitude of 750 km (versus 4 km for LIS from its 350 km altitude) and a lower sensitivity (LIS sensitivity is 1.6 times that of OTD). LIS is still operational aboard the Tropical Rainfall Measurement Mission (TRMM) satellite (31,34,48). These two sensors are in precessing orbits (55-day repeat cycle for OTD; 24-day repeat cycle for LIS) on low earth-orbiting satellites. The BLACKBEARD VHF sensor on board the ALEXIS satellite operated by the Los Alamos National Laboratory (52,53) and the instrument on the FORTÉ satellite operated by the Department of Energy that uses a combination of optical (based on the LIS/OTD design) and VHF sensors to distinguish lightning flashes (49–51), are also aboard low earth-orbiting satellites (Tables 1 and 2). The snapshot from low earth orbit limits their utility in certain research and application areas. For that reason, NASA's next generation optical detector, the Lightning Mapping Sensor (LMS), has been developed (43,163,167) for geosynchronous earth orbit, and the NASA instrument development team is planning for an anticipated flight opportunity for it in June 2004.

Optical Transient Detector

The OTD was a flight-qualified engineering prototype of the LIS and was developed at NASA's Marshal Space Flight Center in Huntsville, Alabama (45). Its development and delivery required only about 9 months, which is an extremely short period for a satellite-mounted instrument. It was launched aboard the MicroLab-1 satellite on 3 April 1995 into a low earth orbit at an altitude of 750 km with a 70° inclination that permitted observations of lightning between latitudes 75° S and 75° N (163,168) (Figs. 19 and 20). The MicroLab-1 completed an orbit of the earth once every 100 minutes. OTD exceeded its expected one- to two-year design life, and data collection ended after five years in April 2000 (46).

Instrument Characteristics. Two distinct but integral units composed the OTD instrument. The sensor, which was essentially a camera, was approximately 20 cm in diameter and 40 cm tall. The accompanying processing electronics equipment was about 10 cm high by 25 cm in length and width. The two modules weighed 18 kilograms, which was about one-quarter of the total weight of the MicroLab-1 satellite. In many ways, the optical equipment on the OTD was highly conventional. The optical sensor was similar to a television camera with lenses, a narrowband interference filter to reject the bright sunlit daytime backgound, a detector array, and a mechanism to convert electronic output into useful data.

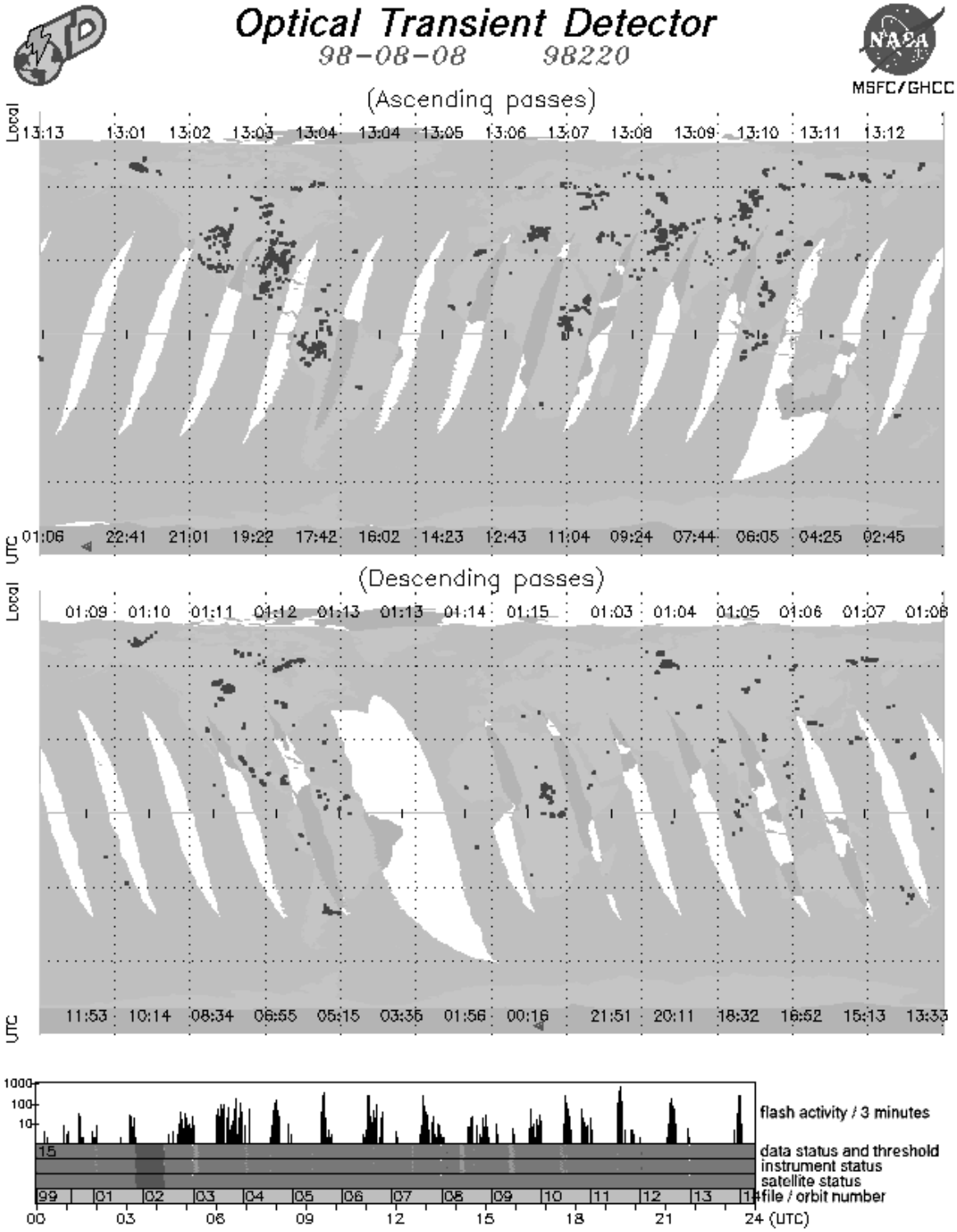


Figure 20. Lightning observations made globally by the OTD during 8 August, 1998 (172). Each location was sampled as often as twice a day, once during northward-moving, ascending passes and once during southward-moving, descending passes; the blue regions were sampled on this day. The local and UTC times of each pass are indicated, respectively, above and below the maps. On this day, the ascending passes were at about 1 P.M. local time, and the descending passes were at about 1 A.M. local time. Because thunderstorms are more likely to be created by heating of the ground during daytime, more lightning discharges were detected during the ascending passes in the upper map. The graph at the bottom of the figure summarizes the daily flash activity in 3-minute intervals and gives data quality flags; the maximum activity, approaching 1,000 flashes per 3-minute period, occurred during the ascending pass over South America. (Figure provided by the NASA Lightning Imaging Sensor science team at Marshall Space Flight Center, AL, used with permission.) See color insert.

Because it was aboard a low earth-orbiting satellite the instantaneous field of view of the OTD was rather limited compared with that possible from a geosynchronous orbit. The field of view of the OTD, 1,300 × 1,300 km, was about 1/300 of the area of the earth's surface (169). The nadir spatial resolution of 8 km (46) is consistent with the average area visibly illuminated by a lightning flash (163), and the temporal resolution of the sensor was 2 ms (46) to minimize the contribution of the sunlit background (i.e., noise). Location errors typically were between 20 and 40 km and had a median value of 50 km; 25% of the errors were greater than 100 km (46).

Capable of detecting lightning during both day and night, the OTD represented a significant advance in lightning imaging technology compared with previous satellite-based platforms such as the OLS. The detection efficiency was 40 to 65%, including both cloud–ground and intracloud (including cloud–cloud and cloud–air) strokes (169); as determined by comparison with the National Lightning Detection Network (NLDN) (see later), the detection efficiency of cloud–ground strokes has been estimated to be 46 to 69% (46). There are four major sources of variance in the detection efficiency:

1. the instrument sensitivity, by design, is an implicit function of the background scene variance; this design allows daytime detection while preserving the signal-to-noise ratio;
2. there is minor (10–15%) variability in the instrument lens/filter response across its angular field of view;
3. there is variability in the pixel ground footprint size across the field of view, which couples to detection efficiency; and
4. the direct line of sight from the sensor to the cloud varies from 0° to about 50° across the field of view, and less light may be scattered from clouds at angles off-normal from cloud top (170).

These four sources do not explain the large range of values cited in (46), however. That large variance was due to uncertainty in the empirical validation technique. As described in a paper in preparation by Boccippio, the actual instantaneous uncertainty or variability in detection efficiency (e.g., one standard deviation) from items (1–4) in the list above has been modeled as 1–20%.

There were several products in which the detected lightning could be reported (47,169,171). The OTD compared the difference in luminance of adjacent frames of optical data. Should the difference in luminance be large enough, an *event*, the finest resolution, was recorded. One or more adjacent events in a 2-ms time frame composed a *group*. Subsequently, one or more groups in a small time period were classified as a *flash*. Flashes were then grouped into *areas*, the lowest resolution, if there was sufficient space between existing flashes.

The OTD's predicted lifetime of about 2 years far exceeded longevity expectations. Data were transmitted each day from the OTD to an intermediate station in Fairmount, West Virginia, followed by transmission to the Global Hydrology and Climate Center in Huntsville, Alabama, where the data were processed, analyzed, and distributed (172).

Applications of Observations. Several important discoveries and lightning databases resulted from analysis of OTD data, including the creation of the most comprehensive maps of global lightning distribution (Fig. 19). In addition, analysis of the OTD data resulted in a revision of the estimate of the global lightning flash rate to approximately 40 flashes per second, a significant decrease from the previous estimate of 100 flashes per second that had been accepted since the mid-1920s (168,169). This flash rate is based on the estimate of 1.2 billion flashes across the globe in the period from September 1995 to August 1996, a value determined by adjusting the number of flashes observed by the OTD to account for the known detection efficiency of the instrument (168). Moreover, the flash rate in the Northern Hemisphere is considerably greater than that in the Southern Hemisphere; rates during the Northern Hemisphere summer can approach 54 flashes per second. Approximately 75% of the lightning flashes occur in the tropics and subtropics between 30° S and 30° N (168).

OTD data analysis also demonstrated that there is far less lightning over the ocean than over the continents (Fig. 19); the global average rate over all of the oceans is only seven flashes per second, but the rate over all of the land is 24 to 49 flashes per second, depending on latitude (168). Although this discrepancy had been hypothesized for some time, based, for example, on studies in the 1950s showing that land is preferred to the ocean for the location of nighttime radio noise at 1 MHz (81, p. 178), the OTD datasets provided the first conclusive evidence for this hypothesis. One strong contributor to the disparity in lightning frequencies results from the higher updraft speeds over land that result from stronger surface heating by the sun; these stronger updrafts lead to greater charge separation in convection over land than that in convection over the ocean.

OTD data also have been used to ascertain a lightning-discharge signature indicative of continuing current in cloud–ground discharges (see earlier). This signature is useful for predicting lightning strike thresholds for the ignition of wildfires (173).

From an operational meteorological standpoint, OTD data analysis has laid the groundwork for what may eventually be an effective tornado warning system based on lightning detected from space (173). In one of many similar examples, on 17 April 1995, the OTD observed that lightning rates in a supercell storm over Oklahoma first rose rapidly, then dropped sharply. Several minutes later, the storm produced a tornado (47). A similar result was observed using the surface-based SAFIR interferometric network that detects intracloud lightning by mapping the three-dimensional orientation of the lightning channel (65) (see later). Strong updrafts result in vertical stretching and an increase in vertical vorticity, which is a measure of the spin of the air about a vertical axis; this increase is a key element in tornado formation because a tornado is an intense vertical vortex. Strong updrafts result in heavy precipitation loading, which, in turn, lead to increased charge separation, and so to increased lightning rates. The strong updraft then weakens, the precipitation loading weakens, and the lightning rate decreases; at this time, the vertical

vorticity has reached a maximum, and the conditions are then optimal for tornado formation. If this hypothesis is validated, then the variation in the frequency of lightning may provide up to ten minutes additional tornado warning time beyond what is currently available by using Doppler radar. Other recent work using ground-based sensors is confirming this tornado signature (174). Lightning flash rate time lines, especially when they include lightning aloft, may also be useful in downburst, and even hail, nowcasting (see also later section).

Despite its advantages, the OTD dataset did have several limitations. Its 40 to 69% detection efficiency was somewhat low and varied between day and night (168). The Lightning Imaging Sensor has improved detection efficiency and sensitivity (48). Also, the OTD was onboard the MicroLab-1, a low earth-orbiting satellite. Because the elevation of the orbit was quite low, the instantaneous viewing area was rather small. The small viewing area and rapid orbit of the satellite made the OTD unsuitable for real-time forecasting. The large location error of 20 to 40 km or more made it an unsuitable instrument for ground-truth studies (31). Consequently, OTD datasets have been applied to lightning climatological research rather than to operational forecasting.

Lightning Imaging Sensor

The LIS was launched from the Tanegashima Space Center in Japan aboard the Tropical Rainfall Measuring Mission (TRMM) satellite platform on 28 November 1997 (175). It is one of five instruments aboard the TRMM platform, which is intended to study tropical rainfall patterns and variability, especially over the ocean where conventional observations are unavailable. As a result, lightning is detected from an altitude of 350 km between 35° N and 35° S (48) (Fig. 21).

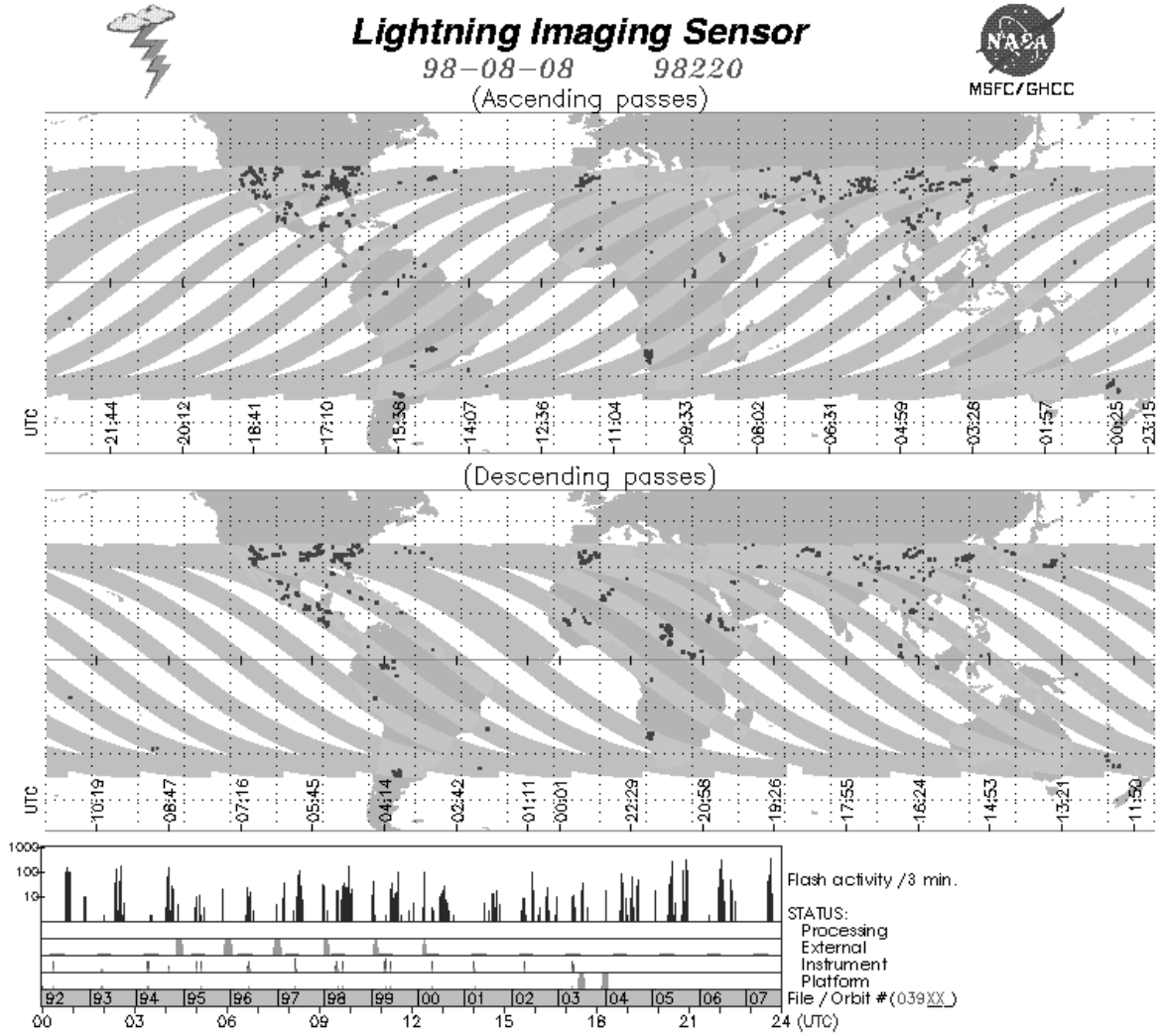


Figure 21. As in Fig. 20, except for the Lightning Imaging Sensor (LIS) (172). The ascending orbit passing over the Gulf of Mexico and Florida is shown in Fig. 22, and the lightning observations near the Kennedy Space Center are shown in Fig. 24a. Note that some parts of the globe were sampled at nearly the same time by both the OTD (Fig. 20) and the LIS. (Figure provided by the NASA Lightning Imaging Sensor science team at Marshall Space Flight Center, AL, used with permission.) See color insert.

Instrument Characteristics. The LIS optical equipment is the same as that in the OTD: a staring imager has a diameter of about 30 cm and height of about 50 cm (48,163). The sensor can detect total lightning, day and night, and uses an optical lens and detection array, as well as a signal processor. Perhaps the most important element of both the OTD and LIS sensors is the *Real Time Event Processor* (RTEP). The RTEP is a sophisticated electronic unit that allows the instruments to filter sun glint and diffuse solar radiation reflected from clouds. By comparing the duration of the reflected sunlight (long) with the duration of a lightning flash (short), the RTEP can remove the background signal level (reflected sunlight) and identify lightning flashes as superimposed transients. As both sensors use RTEPs, the increased sensitivity by LIS is due to other hardware changes.

The altitude of the TRMM satellite on which the LIS is integrated is 350 km compared with 750 km for the MicroLab-1 satellite on which the OTD was located. With an instantaneous viewable area of 600 × 600 km, the LIS detects only a small portion of the earth at any given time. Spatial resolution is between 4 and 7 km, the first at the nadir and the second at the edge of the field of view near 35° N and 35° S (34,48,175). This marks an improvement over the 8-km nadir spatial resolution of the OTD (46). Estimated location accuracy is better than 10 km (176). The LIS detection efficiency has been empirically estimated as 1.6 times greater than OTD (177), a value consistent with modeling of the instrument response. Because the TRMM satellite is not geostationary, the LIS cannot view a single storm for more than 90 s before the storm leaves its field of view (Fig. 22). Although the TRMM satellite orbit is restricted to the tropics and subtropics (35° N to 35° S), this is the region in which the majority of lightning does occur (168).

Applications of Observations. Data from the LIS are being used to generate more complete lightning climatologies in the tropics and subtropics. A comparison of the total number of flashes observed in 1999 by the OTD and LIS instruments is given in Fig. 23 (172). Note that the color scales on the high end for the LIS detections in Fig. 23b are double those for the OTD in Fig. 23a. In addition to the fact that the absolute numbers from the LIS are higher than those for the OTD, the LIS detected relatively more flashes in South America than the OTD, primarily due to the amount of viewing time at midlatitudes. LIS datasets also confirmed that rainfall rates are

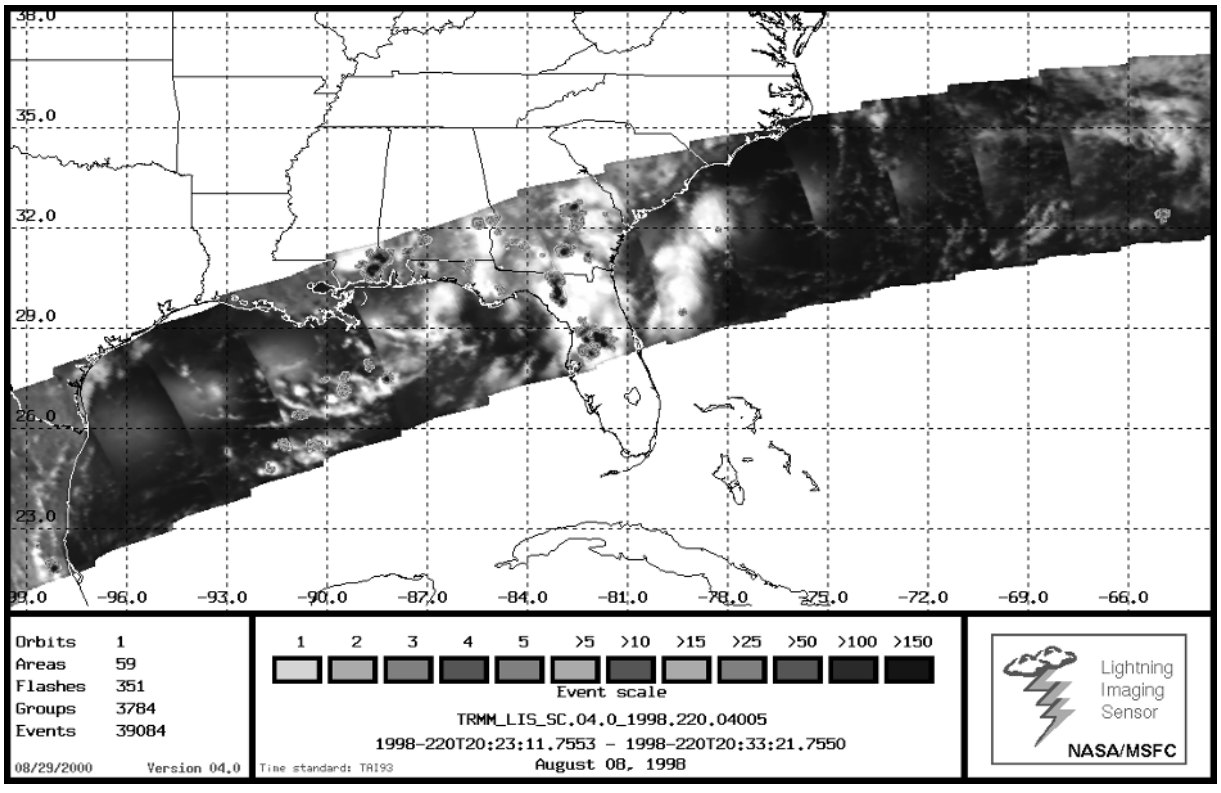
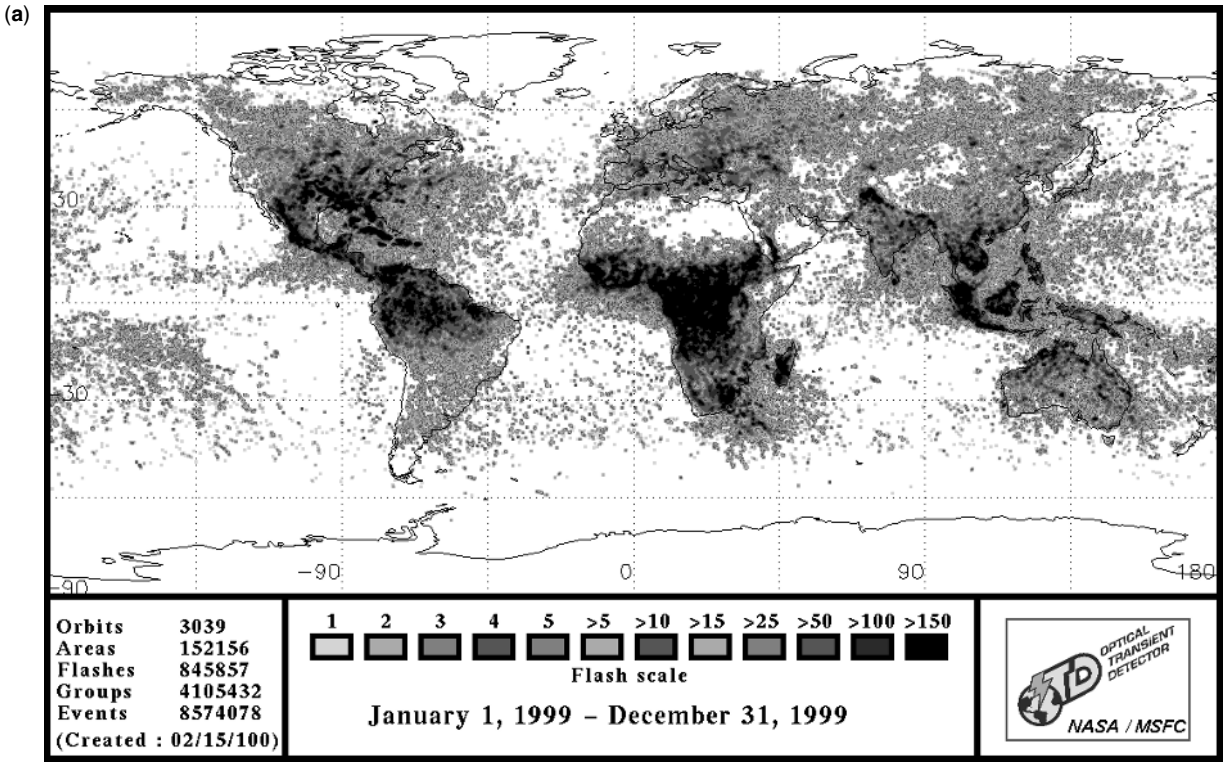


Figure 22. LIS background images taken between 20 : 23 : 11 and 20 : 33 : 21 UTC on 8 August 1998 showing the cloud cover in white, ocean glint, and the lightning events detected by the LIS in color; the color bar indicates the number of events that occurred at each pixel. This image is obtained at the website (172) by clicking the mouse on Florida on the background image that is displayed after clicking on Florida on the global image in Fig. 21. Although more than 39,000 events are detected during this portion of the orbit, these events are confined to an extremely small percentage of the total area covered by clouds. The region of activity in central Florida is depicted in Fig. 24a. (Figure provided by the NASA Lightning Imaging Sensor science team at Marshall Space Flight Center, AL, used with permission.) See color insert.



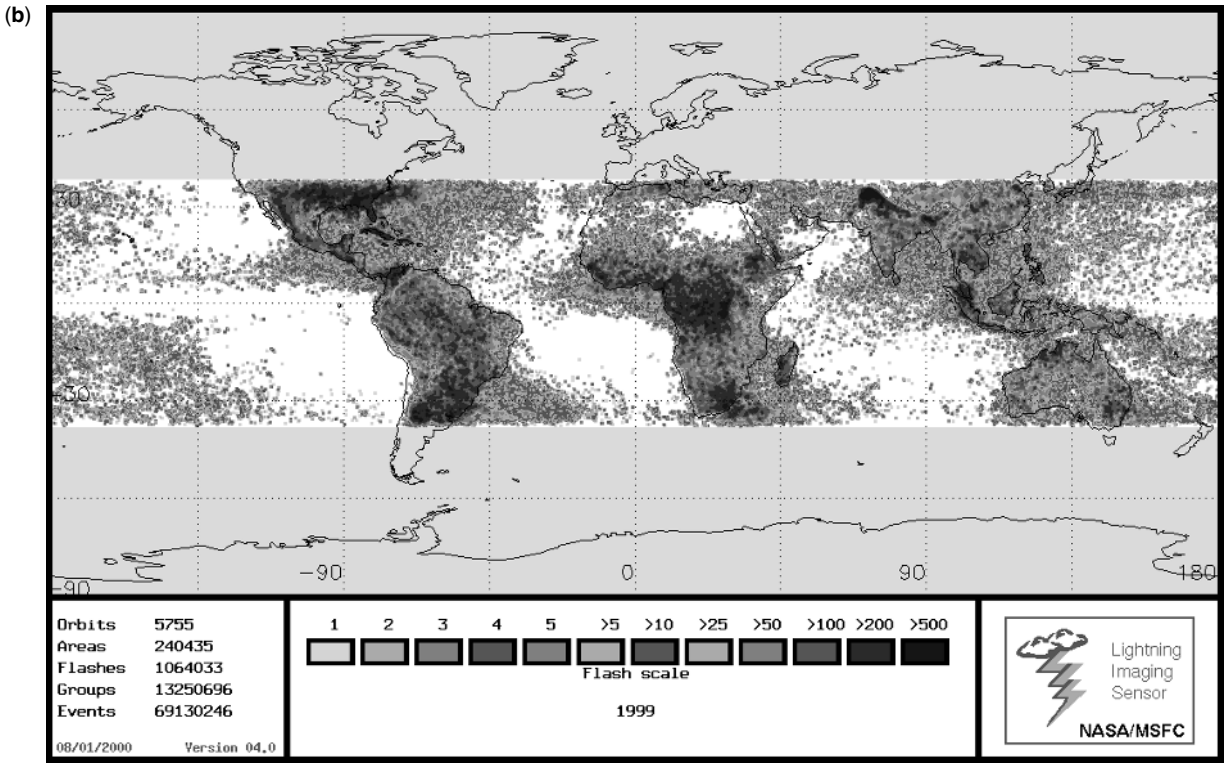


Figure 23. Total number of flashes per unit area in 1999 observed by the OTD (**a**) and LIS (**b**) instruments (172). Note that the color scales on the low end are the same in both figures, but that those on the high end are about twice as large for the LIS as they are for the OTD. Although the overall flash density pattern is the same for both instruments, there are notable differences, such as in South America where the LIS detected far more flashes than the OTD. (Figures provided by the NASA Lightning Imaging Sensor science team at Marshall Space Flight Center, AL, used with permission.) See color insert.

highest in areas of frequent lightning (48,173) and have helped document changes in geographical distributions of wintertime lightning frequencies along the United States Gulf Coast during an El Nino (178).

Because of its value in providing reliable total lightning observations, the LIS data are being used in many cross-sensor research studies (31,34). In particular, an extensive program is under way at the Kennedy Space Center (KSC) to compare the lightning observed by the LIS with that by the National Lightning Detection Network (NLDN) network locator (see later) and the Lightning Detection and Ranging (LDAR) network mapper (31,176) (see later). Indeed, NASA maintains an excellent browse gallery on the World Wide Web (179) where contemporaneous images from these sensors are posted, together with a corresponding WSR-88 precipitation radar image. Figure 24 shows a set of these four images centered on KSC that were taken during an LIS overpass at approximately 20:30 UTC on 8 August 1998; this is the same overpass as depicted in Fig. 22, where it is clear that only a small region covered by clouds is electrically active. Although both the LIS visible wavelength and LDAR VHF electromagnetic wavelength total lightning systems detect all high-frequency stages of the lightning event, the NLDN displays only the cloud–ground strokes (see later). Thus the number of observations reported by the total lightning systems, LIS and LDAR, is larger than the number given by the cloud–ground system, NLDN. Nevertheless, it is clear from Fig. 24 that the three different lightning detection systems agree quite well in placing the lightning locations. In particular, note that the most intense activity detected by the LIS instrument (Fig. 24a) coincides with the locations of the cloud–ground strokes given by the NLDN (Fig. 24b). It is clear that lightning channels span regions larger than those revealed by the locations of the cloud–ground strokes alone. For example, both LIS and LDAR detect lightning west-northwest of the KSC at a 60-km range, but the NLDN does not (Figs. 24a,b,c) LIS validation with a New Mexico Tech-developed Lightning Mapping Array deployed in northern Alabama will continue through the remaining duration of TRMM.

Certainly the LIS sensor represents a marked improvement in lightning detection efficiency from space, but its inability to be used for real-time forecasting and analysis underscores the need for real-time, large-scale detection from space. To this end, the Lightning Mapping Sensor has been developed for launch aboard a geostationary GOES satellite when the opportunity arises (43,163,167,180).

THE NATIONAL LIGHTNING DETECTION NETWORK™: A PLAN VIEW LOCATING SYSTEM

Although networks of cathode-ray direction finders (CRDF) were used during the middle of the twentieth century for estimating the location of cloud–ground lightning using triangulation techniques based on magnetic direction finding (MDF) (Fig. 3), the National Lightning Detection Network™ (NLDN) operated by Global Atmospherics, Inc. (GAI) is the first modern real-time lightning location network to offer lightning locations to a broad set of users in real time (1). This system detects the location of cloud–ground strokes based on the peak signal levels received in the VLF/LF portion of the electromagnetic spectrum (Fig. 11). Cloud–ground and intracloud (including cloud–cloud and cloud–air) strokes are distinguished by applying time-domain waveform analysis (Figs. 6 and 7), and the polarity of the cloud–ground, or return, strokes is determined (see earlier sections). As reviewed in (13) and (14), the NLDN began real-time operation in 1989 when gated, wideband regional MDF networks in the western and midwestern United States (11,181) were merged with a network operated in the eastern United States by the State University of New York at Albany (182). Originally the network was created in the 1980s to support the needs of electric utilities and forestry agencies, but since that time, the user base has broadened considerably to span a wide variety of commercial and government applications (1,13).

Optimizing Stroke Locations

In the mid-1990s, the MDF network was combined with the time-of-arrival (TOA) network known as the Lightning Position and Tracking System (LPATS) (183) (Fig. 3), which had been operated since the late 1980s by Atmospheric Research Systems, Inc. (ARSI), to form the modern NLDN (14). Finally, in the late 1990s, the network was merged with the Canadian Lightning Detection Network (CLDN) to form the North American Lightning Detection Network (NALDN) (12). Today, many similar VLF/LF MDF or TOA networks are in operation throughout the world (Tables 1 and 2). These networks achieve similar location accuracies and detection efficiencies for both negative and positive return strokes in cloud–ground lightning flashes; because the NLDN is the first and perhaps the most extensively studied of these networks, it is described here as an example of a network locating (NL) system.

The hybrid sensors in the NLDN that combine both the MDF and TOA approaches are called Improved Accuracy from Combined Technology (IMPACT) sensors (13,14). Site error is the largest contributor to erroneous measurements using the MDF triangulation method (see earlier section). Both distortions of the magnetic field near the antenna (152) and man-made and natural features that affect the propagation of electromagnetic signals from lightning cause site error, which is both site-specific and bearing-dependent. Once the site error for a specific location is determined, it generally remains constant, thereby allowing routine adjustments of observations to account for it (13;22, p. 160; 56). The most recent IMPACT ESP (Enhanced Sensitivity and Performance) model of the sensor yields bearing uncertainties of less than 1° (184).

As discussed in early sections, TOA location techniques compare the time when three or more sensors detect an electromagnetic signal from the same source such as a lightning discharge. The majority of errors associated with the TOA technique are due to imprecise time synchronization of the sensors. Thus, an extremely accurate and well-calibrated timing system is required for the detection network to implement the TOA technique effectively. Clock synchronization errors must be less than 10^{-6} s. To achieve this accuracy, the NLDN uses the global

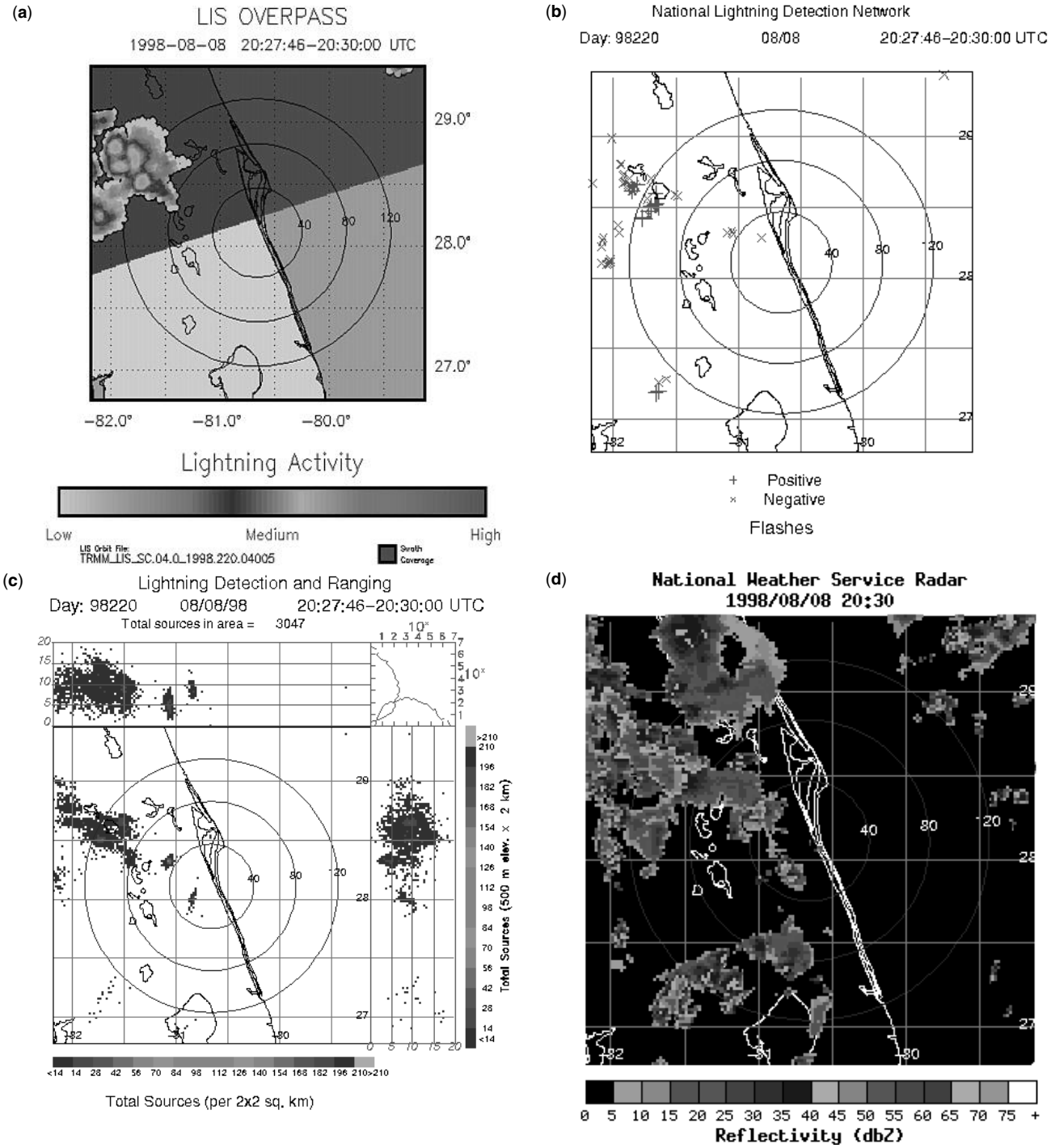


Figure 24. A set of four contemporaneous depictions of the lightning near the Kennedy Space Center (KSC) in Florida during an LIS overpass between 20:27:46 and 20:30:00 UTC on 8 August 1998 (179). KSC is at the center of each image; latitude and longitude grids and range circles are provided every 40 km. LIS data, confined by the satellite track to the green region, are shown in (**a**), NLDN data in (**b**), LDAR data in (**c**), and National Weather Service WSR-88 data from Melbourne, Florida, in (**d**). The total lightning sensors LIS and LDAR show excellent agreement west-northwest of KSC; the cloud–ground sensors in the NLDN show activity in the same region, although this activity appears to be less only because the NLDN instruments do not detect the cloud discharges. The intensity of the radar echo in (**d**) is indicated by the reflectivity (dBZ). Note the close correspondence between the locations of the greatest amount of lightning and the locations of highest reflectivity given in red (>50 dBZ). (Figures provided by the NASA Lightning Imaging Sensor science team at Marshall Space Flight Center, AL, used with permission; NLDN data shown with permission of Global Atmospherics, Inc.) See color insert.

positioning system (GPS) for timing. Additional site errors leading to inaccurate placement of the TOA hyperbola (Fig. 3) arise from changes in the electromagnetic signal as it travels from the source to the sensor location (22, pp. 161–162). Communication errors between the sensor and the satellite relaying the observations to the central processor may be caused by attenuation due to heavy rain or by system saturation during high data rates; these errors can also degrade the performance of the system in real time. These problems are corrected when the data are reprocessed for use in system recalibration and meteorological research studies (13).

When the network was upgraded in 1995 using instruments that had better ranging, the total number of sensors was reduced from 130 to 106 (13) (Fig. 25). Today 59 LPATS-III TOA sensors and 47 IMPACT MDF/TOA sensors across the United States compose the NLDN (12). The Canadian CLDN adds 55 of the newer LPATS-IV and 26 of the newer IMPACT ESP sensors to those in the NLDN to form the NALDN (12). Using these newer models, GAI reports that more than 98% of the detected strokes are correctly identified as cloud–ground or intracloud strokes, and the return stroke polarity is assigned correctly 99.5% of the time for lightning ranges out to 600 km (184). Achieving such high rates is important because the focus of the network is on accurate reporting of the location and polarity of cloud–ground strokes. Network estimates of 85% flash detection efficiency and 500-m location accuracy for cloud–ground strokes are supported by ground validation studies based on independent video recording of lightning flashes within the network (12,185,186). This is a marked improvement in the 65 to 80% detection efficiency achieved by the MDF triangulation version of the network before 1995 (13).

To achieve these network accuracies, the NLDN determines the final location estimate of a lightning stroke by combining the location estimates given separately by several sets of MDF and TOA sensors (Figs. 3 and 26). A chi-square test provides the nonlinear least-squares technique for determining the optimum stroke location (13). This location comes from the optimal combination of MDF triangulation bearings and TOA times that are given by minimizing the value of an appropriate chi-square function, defined by (22, pp. 160–162)

$$\chi^2 = \sum_i \left(\frac{\phi_i - \phi_{mi}}{\sigma^2_{az,i}} \right)^2 + \sum_j \left(\frac{t_j - t_{mj}}{\sigma^2_{tj}} \right)^2 . \qquad (3)$$

For the ith MDF triangulation station, ϕ_i is the bearing of the calculated stroke location in the trial solution, ϕ_{mi} is the bearing measured by the station, and $\sigma^2_{az,i}$ is the expected azimuthal error in the measurement. For the jth TOA station, t_j is the time at which the lightning signal arrives in the trial solution, t_{mj} is the time of signal arrival measured at the station, and σ^2_{tj} is the expected error in the time measurement. Combining the measurements from the two independent systems by minimizing the value of χ^2 produces a least-squares optimal location estimate that overcomes many of the problems inherent in systems that use only one location technique (13). This location process is illustrated in Fig. 26, in which three bearing and seven TOA estimates are combined to provide a strike point at latitude 30.5148° N and 97.8550° W in central Texas on 2 November 2000. The bearing estimates given by stations **Z6**, **Be** and **Z0** are indicated by the intersecting radials, and the time-of-arrival estimate given by each station is indicated by a circle centered on the station whose radius is given by the product of the speed of light and the difference between the time of the stroke and the time of the sensor measurement.



Figure 25. Distribution of IMPACT MDF/TOA hybrid (triangles) and LPATS-III TOA (circles) sensors that form the National Lightning Detection Network (NLDN) in the United States. (Used with permission of Global Atmospherics, Inc.)

Determining Stroke Current

The NLDN reports peak currents I_{peak} for cloud–ground strokes; negative values are assigned to negative return strokes, and positive values to positive return strokes. The method for determining I_{peak} is based on the theoretical *transmission line model* (TLM) that describes the time-invariant waveform of the far-field radiation produced by a vertical lightning channel (14,112,133). Soon after the upward streamer attaches to the stepped leader, a typical first return stroke produces a peak E-field signal level E_{peak} of 5 to 10 V/m at a range r of 100 km (187), with an upward propagation velocity v in the channel of roughly one-half the speed of light c (129). For μ_0 equal to the permeability of free space, $\mu_0 = 4\pi \times 10^{-7}$ Wb A^{-1}m^{-1} (188, p. 316), the TLM relating E_{peak} and I_{peak} is given by (189)

$$E_{\text{peak}} = \frac{-\mu_0 v I_{\text{peak}}}{2\pi r}. \qquad (4)$$

For example (1), an E-field peak level $E_{\text{peak}} = 8$ V/m at $r = 100$ km when $v = (1/2)c = 1.5 \times 10^8$ m/s yields $I_{\text{peak}} = -27$ kA.

To apply the TLM for determining I_{peak}, first the value of the measured signal is converted to a *range-normalized signal strength* (RNSS), given by the *signal propagation model* (13,133)

$$\text{RNSS} = a \exp\left(\frac{r - 100 \text{ km}}{10^5 \text{ km}} \right) \left(\frac{r}{100 \text{ km}} \right)^p y, \qquad (5)$$

where y is the received E-field or B-field signal, in either V m^{-1} or Wb m^{-2}, r is the range in km, and a and p are

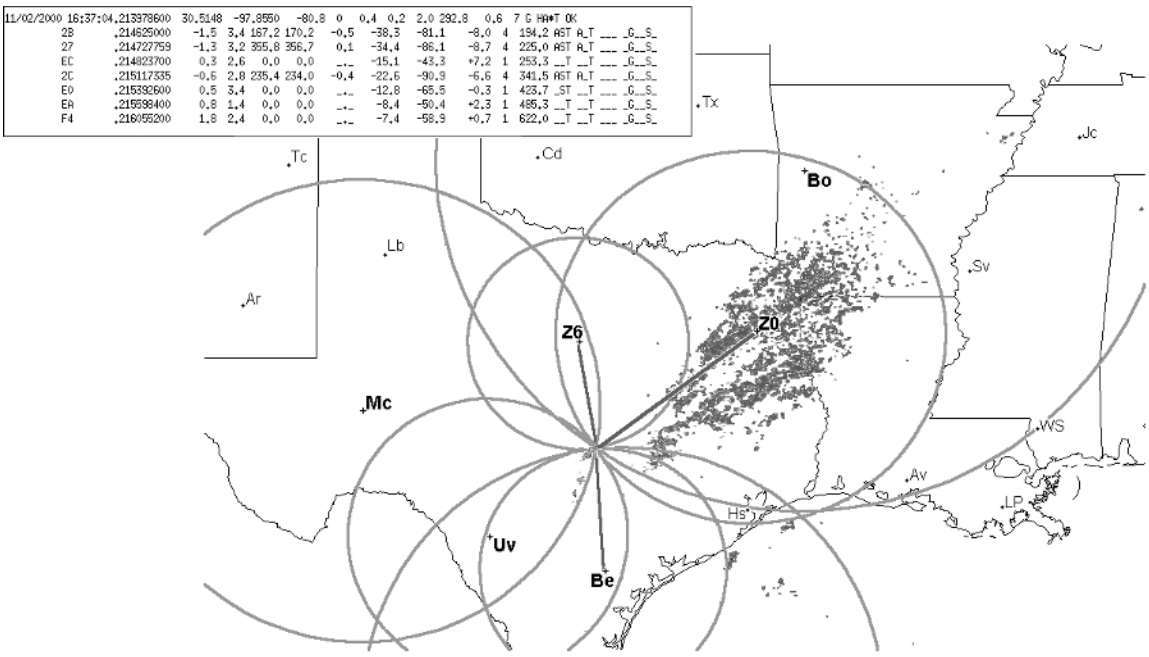


Figure 26. An example from 2 November 2000 of an optimal cloud–ground lightning strike point given by combining measurements from seven sensors in the NLDN network. The location of the lightning stroke, whose parameter values are summarized in the table above the map, is indicated by the intersection of radials and circles at latitude 30.5148° N and 97.8550° W in central Texas. Recent lightning locations across the mapped region are indicated by the green symbols, + for positive and − for negative polarity. The locations of the measuring stations in the network are indicated by two-letter codes. The seven that were used here are denoted in bold letters on the map, but their hexadecimal equivalents are used in the table (**Z6** and 2B, Stephenville, TX; **Be** and 27, Beeville, TX; **Z0** and 2C, Overton, TX; **Uv** and EC, Uvalde, TX; **Mc** and E0, McCarney, TX; **Bo** and F4, Booneville, AR; and EA, Post Isabel, TX, which is to the south of the map). Three olive-colored bearing radials given by MDF stations **Z6**, **Be**, and **Z0** and the seven brown time-of-arrival circles intersect at the estimated strike location. Each time-of-arrival circle is centered on the measuring station and has a radius given by the product of the speed of light and the difference between the time of the stroke and the time of the sensor measurement. From left to right, the parameter values in the table are given by GAI's LP2000 manual as stroke summary in the first line — date; corrected time to the nearest nanosecond; latitude, degrees; longitude, degrees; average range-normalized signal strength ($\overline{RNSS}$), LLP units (1 LLP = 1.5×10^{-10} Wb m^{-2} = 4.5×10^{-2} V m^{-1}) and stroke polarity (− for negative and + for positive); maximum multiplicity of the stroke; semiminor axis of the confidence ellipse, km; semimajor axis of the confidence ellipse, km; eccentricity of the ellipse; angle of orientation of the ellipse, degrees; reduced chi-square; number of sensors used; stroke type, G for cloud–ground and C for cloud; type of start position used in the iteration to minimize location errors, H for hyperbolic and T for triangulated; information used in optimizing the stroke location, A for angle, S for signal, T for time, * if not used; and quality check indicator, OK if all checks met. Each remaining line is for one of the sensors used — hexadecimal ID; sensor report time, decimal fraction of a second; difference in time between the actual stroke time and the sensor's measured time, microseconds; rise time, microseconds; bearing if an MDF station, navigational degrees; corrected bearing, navigational degrees; angle deviation between the corrected bearing and the bearing derived from the calculated stroke position, degrees; measured signal (corrected), LLP units; RNSS, given by product of the measured signal and [(range in km)/100]$^{1.13}$, LLP units; difference between station RNSS and average $\overline{RNSS}$, LLP units; multiplicity; range to stroke, km; configured usable information, A for angle, S for signal, T for time, underscore if not used; information used to determine location, A for angle, S for signal, T for time, underscore if not used; and last two columns other descriptive codes. (Used with permission of Global Atmospherics, Inc.) See color insert.

positive constants; the value of 100 km in the denominator of the term raised to the pth power is used as the standard range normalization. Over a range of 500 km, the exponential term in Eq. (5) represents a small, less than 0.5%, correction to the power-law term. Normally, RNSS is expressed in LLP units, which are related to the magnetic and electric field strengths via 1 LLP = 1.5×10^{-10} Wb m^{-2} = 4.5×10^{-2} V m^{-1} (14) (*LLP units* derive their name from *Lightning Location and Protection*, the original company operating the NLDN). Presently, all stations within 625 km of a stroke calculate the value of RNSS using the same parameter values of $p = 1.13$

and 10^5 km in Eq. (5), although work is ongoing to find site-specific values for these two parameters (13,132). The value of RNSS given by Eq. (5) for each station is then averaged to give $\overline{RNSS}$, expressed in LLP units (see table in Fig. 26). Next, this average is converted to I_{peak} in units of kiloamperes (kA) via the simple equation $I_{\text{peak}} = 0.185\overline{RNSS}$, which is based on a linear regression analysis (13); this equation produces an estimate of 14.9 kA for the negative return stroke in Fig. 26. This equation also revises that given in the study by Idone and collaborators (133) that was created for the NLDN before the TOA sensors were added to the network.

The NLDN detects individual cloud–ground lightning strokes that compose a lightning flash (see earlier sections). It reports both the first and subsequent strokes in a flash, although the detection efficiency of the subsequent strokes is much reduced from the detection efficiency of the first strokes (13). Summaries of lightning occurrences, such as those plotted in Figs. 24b and 27, are also available in which the latitude, longitude, time, stroke polarity, and peak current I_{peak} of the first strokes are reported; normally this first stroke produces the largest peak current, and so characterizes the flash well (13;74, p. 121; 131). The algorithm that groups detected strokes into flashes uses both the time of detection of the strokes and the distance these strokes are from each other. As noted earlier, a flash as defined by the NLDN consists of those strokes, of either polarity, that occur within one second of the initial stroke, are within 10 km of the initial stroke, and are no more than 0.5 s apart in time (13). Separations of this magnitude are considered because as many as 50% of the negative return strokes can have multiple attachments to the ground (127,128). An NLDN flash may contain as many as 15 strokes. In part due to the 50–60% detection efficiency of subsequent strokes, the average multiplicity found is only about 2 (13), which is approximately half the expected number based on previous studies (141). Strokes that have higher peak currents are more likely to be part of multistroke flashes; for example, strokes whose mean peak current is 40 kA have a mean multiplicity of 3.3 (12). When more than 15 strokes occur that meet the single-flash criteria, the additional strokes are arbitrarily assigned as part of a second flash. The location of the flash is given by the latitude and longitude of the first stroke in the flash (13).



Figure 27. An example of real-time cloud–ground lightning stroke locations given by the NLDN in the United States for the period 19:00 to 21:00 UTC 27 August 2000. Yellow pixels indicate strike locations that are current; red, locations that are about an hour old; and blue, locations that are about two hours old. Global Atmospherics, Inc. provides these data through the World Wide Web-based Lightning Explorer service that is accessed via *http://www.lightningstorm.com*. This example is made available free and is updated every 15 minutes; more detailed times and locations are available for a fee. A line of thunderstorms is clearly seen stretching down the East Coast of the United States; scattered thunderstorms are occurring elsewhere, particularly in the forest-fire-prone western United States. (Used with permission of Global Atmospherics, Inc.) See color insert.

Applications of Observations

The NLDN has been used extensively to study lightning climatology in the United States (12,190,191). A total of 13.4 million flashes was reported by the early version of the NLDN in the United States in 1989, and the peak of 3.6 million ground flashes occurred in July (190). After the system was upgraded in 1994, the number of observed flashes increased from a preupgrade mean of 16.7 million ground flashes per year to 20.6 million flashes in 1995 (14). Furthermore, the positive polarity flash count increased from 700,000 flashes per year to 2.1 million in 1995. These changes have been attributed to changes in the cloud–ground stroke selection algorithm, which is reviewed later (15). From the upgraded system in 1998, the maximum *ground flash density* value of 16 flashes/km^2/yr occurs in Florida; the minimum value is much less than 1 flash/km^2/yr in the northwestern United States and Canada (12). There is considerable year-to-year variation in ground flash density, but in general there are many more ground flashes per unit area per year in the midwestern United States and central Florida than in other portions of the country; in some years, the density values in the Midwest have approached those seen in Florida (191). Moreover, positive cloud–ground strokes are far more likely in the central United States and Florida than elsewhere, and their relative numbers are affected by the types of aerosols in the air in which the thunderstorms form (147).

As a percentage of total cloud–ground strokes, positive cloud–ground strokes are found relatively more frequently in the winter months, perhaps approaching 20% of the total, and in the central United States (191). Overall, about 10% of the cloud–ground strokes are positive, although this percentage may be high by a factor of 2 or more, owing to the suspected acceptance of some intracloud strokes as positive return strokes by the algorithm used by the NLDN for time-domain waveform analysis (13–15). This overly high acceptance rate of positive ground flashes is indicated, for example, by work at the Kennedy Space Center that compared results given by the CGLSS MDF triangulation network, a VLF/LF system described at the end of this section that is a small-scale version of the NLDN (Tables 1 and 2), with those obtained by

the Lightning Detection and Ranging (LDAR, see later) system; LDAR can map the entire lightning channel and so can indicate if the stroke remains aloft or approaches the ground. In addition, as part of the 1995 system upgrade, the minimum waveform width for acceptance of a return stroke was decreased from 11 μs before 1995 to 7.4 μs after (113), the sensor gain was increased by 50%, and the field overshoot tolerance (see earlier) was increased from 0.85 times the initial peak value to 1.2 times (14). These changes in the cloud–ground stroke acceptance algorithm led to an increase in the number of relatively weak positive return strokes (13,15); some are apparently misdiagnosed large-amplitude intracloud pulses known as narrow positive bipolar pulses (NPBP) (114,115). Because intracloud strokes tend to have much smaller waveform widths — of the order of 5 μs — than return strokes (Figs. 6 and 7), setting the minimum waveform width criterion too low will undoubtedly lead to the misclassification of some intracloud strokes as return strokes. Indeed, apparent positive return strokes whose peak currents are less than 10 kA are reclassified during reprocessing as intracloud strokes (13).

Many practical forecast applications are emerging from the NLDN archived databases. Flow regime stratified climatology is being used in Florida by the United States Air Force (USAF) 45th Weather Squadron and the National Weather Service office in Melbourne, Florida, to predict local thunderstorm probability (66). In addition, studies of lightning strike distance distributions are influencing lightning safety designs. Finally, NLDN data provide an important source for cross-sensor studies that assess the quality of lightning detection and location by other systems (Fig. 24) (31,34,47).

CGLSS — A Small-Scale Version of the NLDN

Around the Cape Canaveral Air Force Station (CCAFS) and the John F. Kennedy Space Center (KSC) at Cape Canaveral, Florida, the USAF operates a small lightning location network that is a small-scale version of the NLDN (16). Known as the Cloud to Ground Lightning

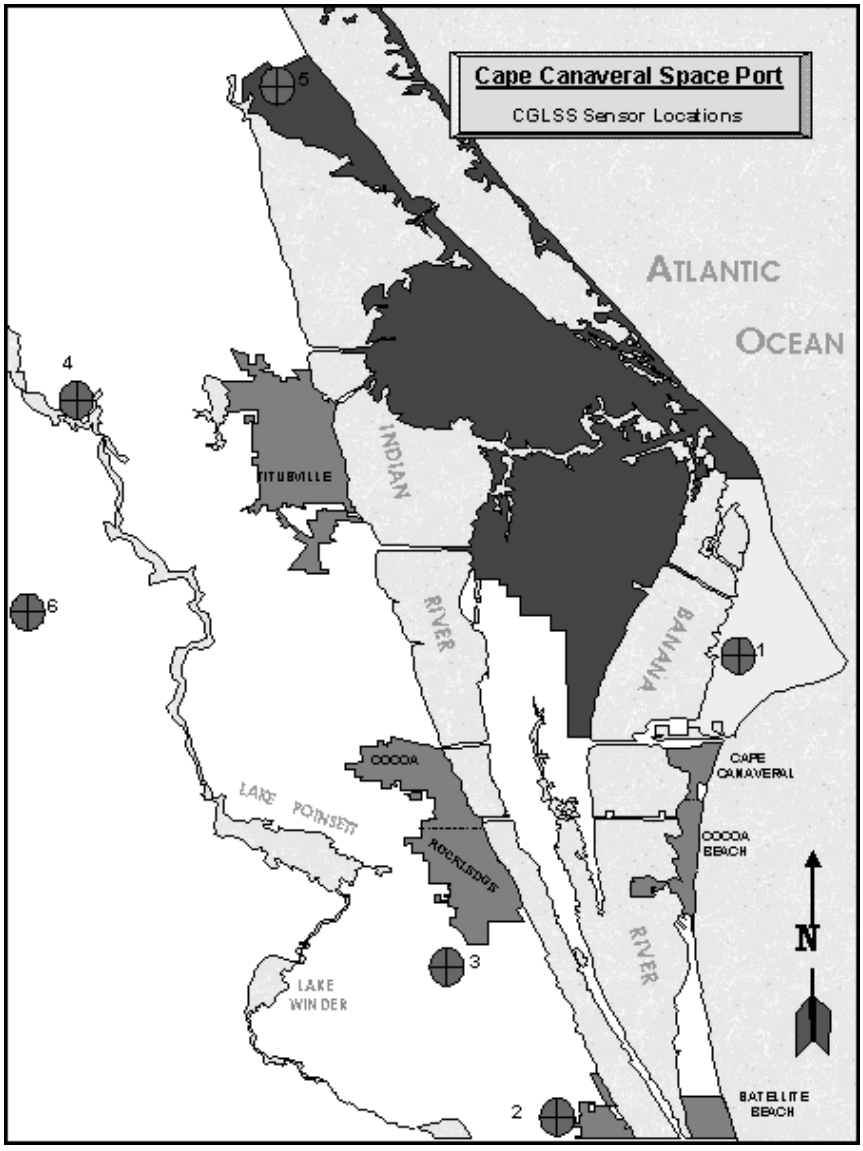


Figure 28. Map of the region around Kennedy Space Center (KSC) showing the six sites at which there are VLF/LF IMPACT sensors composing the MDF triangulation CGLSS network, which is a small-scale version of the NLDN. The sensors form a circle of approximately 20 km in radius, and so the system loses effectiveness beyond roughly 100 km. The ground strike point locations given by CGLSS are used to supplement the lightning observations given by the LDAR system that locates the lightning channel aloft via locating bursts of short-duration radio-frequency (RF) sources (Courtesy USAF, 45th Weather Squadron, used with permission.) See color insert.

Surveillance System (CGLSS), this network is used to estimate the electromagnetic pulse (EMP) hazard to payload, launch vehicles, and support electronics. If lightning strikes closely enough or strongly enough, then the affected electronics must be inspected for possible damage. The CGLSS is also used to help 45th Weather Squadron forecasters issue lightning watches and warnings for personnel safety and resource protection. CGLSS consists of six IMPACT sensors (Fig. 28) that are extremely well maintained and calibrated to remove local site effects. At CCAFS/KSC, there is a large network of video cameras and some lightning protection devices that together allow extremely precise location and timing of cloud–ground lightning and so allow calibration of the sensors. Consequently, location accuracy close to 250 m and detection efficiency close to 98% are achieved within the CGLSS network. This compares favorably with the 500-m accuracy and 90% detection efficiency given by the NLDN (13). Owing to the placement and number of sensors in the network, the CGLSS performance loses effectiveness beyond 100 km, in contrast to that provided by the much larger NLDN, which provides coverage throughout the United States (13).

LIGHTNING DETECTION AND RANGING: A TOA THREE-DIMENSIONAL NETWORK MAPPER

A lightning discharge has various components that are detectable across many frequencies and across a wide range of distances (Fig. 11). The systems reviewed in earlier sections use the electromagnetic radiation emitted by lightning in the very low frequency/low frequency (VLF/LF) bands that span the approximate range of 1 to 400 kHz. The ability to use such sferics has been well established for more than a half century (1), and is best suited for detecting cloud–ground lightning at ranges from 100 to as much as 1,000 km (13,84,184,192). The effectiveness of wideband magnetic VLF methods is seriously degraded at close ranges (within 100 km), in part because of the possibility that the system is prematurely triggered by stepped leaders in nearby lightning (Fig. 8). Arnold and Pierce (193) found that an intracloud *K*-discharge associated with dart leaders at 20 km also produces a peak signal level that is comparable in size to that of a return stroke discharge at 100 km. Wideband VLF/LF techniques focus on detecting cloud–ground strokes, but many applications such as space operations require knowing whether any intracloud (including cloud–cloud and cloud–air) or prestroke discharges are occurring that commonly precede the first cloud–ground strokes by several minutes. Knowledge of these cloud discharges is needed in this case because such discharges pose a hazard to missiles launched in environments free of cloud–ground strokes (16). Locating such cloud discharges requires mapping the entire lightning channel.

Thus this operational requirement for locating lightning channels within thunderstorms in real time and at close range has necessitated the development of sensing systems that use very high frequency (VHF) radiation associated with prestroke sferics. As summarized earlier, these initial prestroke discharges, which are best detected across a frequency range that spans a few to several hundred MHz, are the negative stepped leader processes within a cloud (Fig. 11) that are the precursors to both cloud–ground and intracloud lightning (86). Indeed, such prestroke discharges within a cloud may occur between seven and 20 minutes before the first cloud–ground flash (65,81,194–196). One study (197) found an average lead time of only 4.5 min, however, and a leadtime of less than one minute — the Lightning Detection and Ranging (LDAR) system display rate — in 23% of the cases. Thus, in nearly one-quarter of the cases, there is essentially no warning of a cloud–ground stroke. Nevertheless, the sources of VHF radiation, which can be detected at line-of-sight distances (84), when mapped in three-dimensional space, may be used to help warn of the possibility of cloud–ground discharges.

In this section, the long-baseline time-of-arrival (TOA) VHF technique is reviewed (22, pp. 152–153) as it is applied by the LDAR system that is operated by the Kennedy Space Center (KSC) at the Cape Canaveral Air Force Station (CCAFS), Florida, and is used by the 45th Weather Squadron, United States Air Force (USAF), Patrick Air Force Base, Florida, to support space operations (28–31,87). A similar, but deployable, VHF ground-based mapping system, the Lightning Mapping Array (LMA) or Lightning Mapping System (LMS), is operated by New Mexico Tech (32,33) and is used in a variety of field programs that study lightning (34,107). Finally, an interferometric ground-based VHF mapping system, Surveillance et Alerte Foudre par Interférométrie Radioélectrique (SAFIR) (24,25), is discussed later.

VHF TOA Systems

Zonge and Evans (194) were among the first people to use an array of antennas operating at frequencies from a few megahertz to more than 100 MHz to detect electromagnetic radiation from a growing rain shower 10 to 15 minutes before the first discharge. The development of accurate time interval counters, whose resolution is of the order of microseconds, enabled Oetzel and Pierce (192) to use pairs of receivers spaced less than 300 m apart, giving a *baseline length* of 300 m, to detect VHF emissions. Multiple receivers are needed in a TOA system, because the difference in the time of arrival of a signal between any pair of receivers produces a locus of points traced by a hyperbola (87) (Fig. 3). Ambiguity in the origin of the discharge is usually eliminated when a third receiver is employed, because for most cases, the location is given by the unique intersection of the hyperbolas. A fourth receiver is often included in the system to resolve those cases where multiple intersections are possible (22, pp. 160–161). With five receivers in the shape of a cross, the location in three-dimensional space can be ascertained (22, pp. 152–153; 198,199). With more sensors, multiple solutions allow improvements in data quality control, internal consistency checks, and location accuracy via statistical techniques (87).

The sensor configuration chosen by Oetzel and Pierce (84) used a short baseline length of 300 m because they could achieve a time difference resolution of only about 10 μs. As a result, their method was limited

primarily to detecting overhead discharges at a range greater than the baseline length. This requirement did not present a problem because VHF discharges originate at 5 to 10 km above the surface, in the middle or upper levels of clouds (101). Further developments in accurate time measurement enabled researchers to expand the baseline length to tens of kilometers (198) and to use five antennas. Expanding the baseline length enabled researchers to map stepped leaders over a wider range (up to several hundred kilometers), rather than only those overhead. In addition, further studies (200,201) found that the optimum frequency for detection is between 30 and 85 MHz. This bandwidth captures the distinctive spike in amplitude given by the stepped leader and has the least interference from local television stations and other man-made sources.

Further research relating VHF emissions to the origins of the lightning process led to development of VHF detection networks in South Africa (199,202), Ontario, Canada (200,203) and KSC, Florida (201), in the middle 1970s. Observations from all locations showed that the highest concentration of stepped leader pulses, up to 2,500 discharges per single cloud–ground stroke (199), is above the freezing level and most concentrated between −5 and −15 °C (199,202,203). The high number of stepped leader pulses found to occur in the mixed-phase portion of a cloud — where high concentrations of large liquid drops, rime-coated particles, and graupel occur — is an essential ingredient in the early stages of the lightning process and so in the genesis of a flash.

The LDAR System

LDAR is one of the earliest VHF detection networks still in operation; it is presently providing support at CCAFS/KSC to the United States space program. This system detects VHF radiation in the 6-MHz bandwidth centered at 66 MHz (204) and provides a real-time view of the lightning process to forecasters. The location of the KSC is ideal for testing the limits of an LDAR system (30,31), because central Florida has an average of 8–16 cloud–ground strikes per km^2 per year (12,205), the highest concentration in the United States.

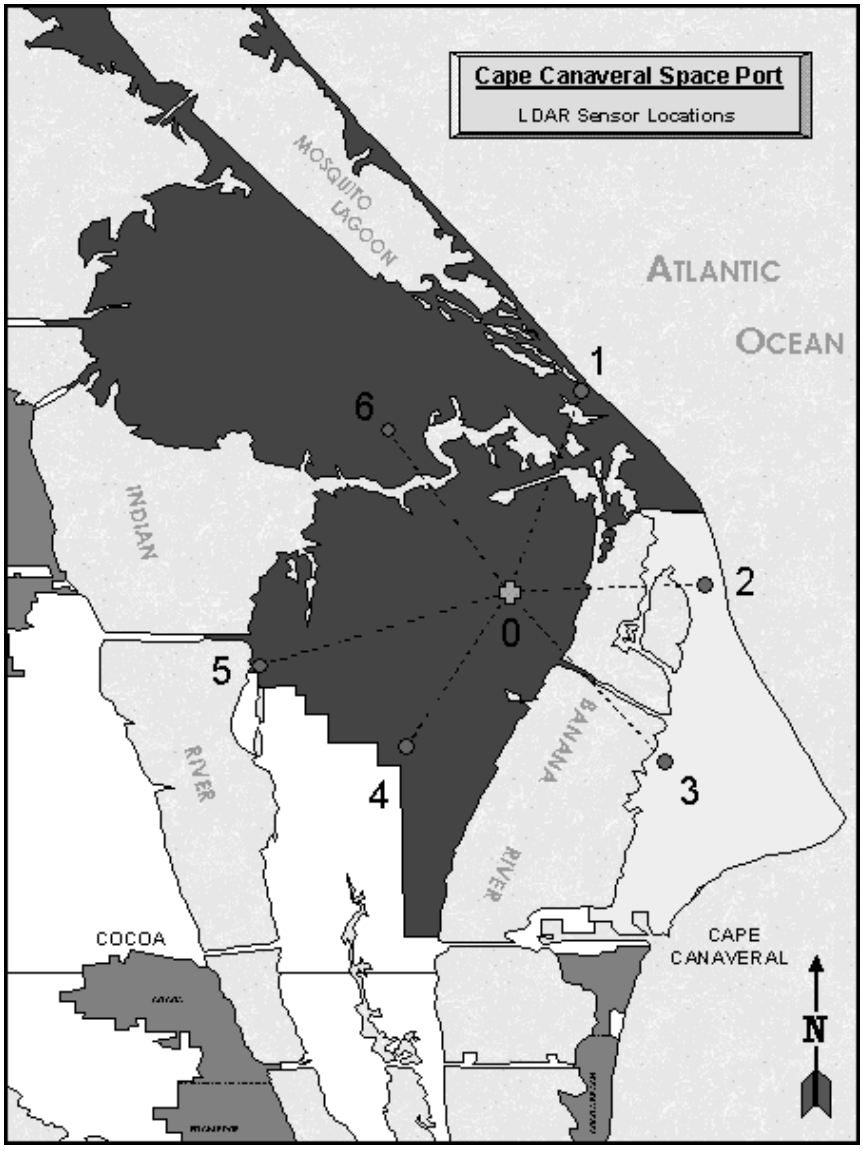
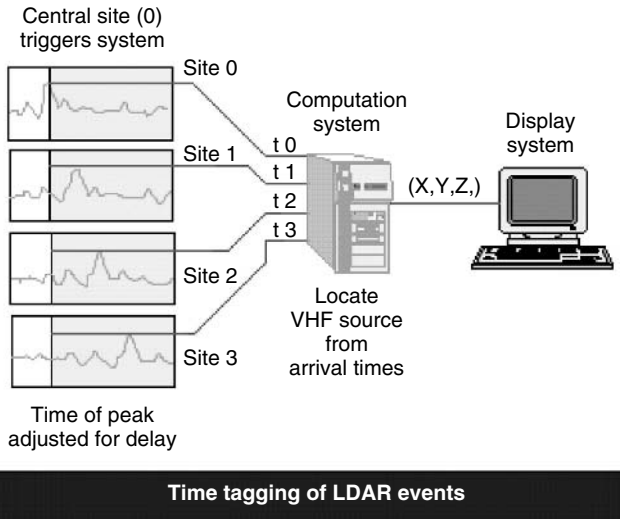


Figure 29. Map of the region around the Kennedy Space Center (KSC) showing the seven sites where there are VHF antennas in the LDAR network. The sensors form a circle approximately 8 km in radius. Reception of a lightning waveform at site 0 triggers the opening of a 100-μs window during which the other six sites report whether and when they receive the same signal, offset perhaps in time (Fig. 30). (Courtesy USAF, 45th Weather Squadron, used with permission.) See color insert.

System Characteristics. The arrangement of seven sensors around the KSC, shown in Fig. 29 from (196), is akin to the multiple-sensor arrangement of Proctor (199). Six outlying sensors form a circle whose radius is approximately 8 km around the central site (site 0) and so provide a web of coverage around the KSC complex (87,196,204). Because the power received by the VHF emissions decreases 10 dB for every 71 km (30,206), the KSC network is optimized to track thunderstorms in the immediate vicinity of KSC, up to 150 km away in some cases (Fig. 24c). VHF emissions are received and processed at the six outlying sites and then transmitted to the central site.

When site 0 receives a VHF emission exceeding a predetermined threshold, the system is triggered, and the central site opens a 100-μs data window to determine the times relative to the trigger time, known as *time tagging*, of LDAR events between all pairs of receivers (Fig. 30); time tagging can be accomplished at 10-ns resolution using the global positioning system (GPS) (29,30,196). During this window, the system determines the peak signal level at each of the seven sites, tags it, and adjusts for the time delay during transmission to the central site. In principle, detection at only four sites is required, three outlying sites and the central site 0, to map an event (204). In practice, detection at all seven sites is required for better quality results (87). The antennas are divided into two optimal "Y" configurations, antennas (0, 1, 3, 5) and (0, 2, 4, 6), and the central site is common to both. If the solutions given by these two configurations agree to within 5% or 350 m, whichever is greater, then the mean of the two solutions is used as the stepped leader location. If this double solution fails, then a computationally more expensive solution is used, using all 20 possible combinations of four antennas (29,87). If these 20 solutions satisfy the quality control criteria, then a weighted solution is used for the stepped leader location (204). If the 20 solutions do not satisfy the criteria, then no solution is displayed. The system can process up to 10^4 pulses per second (22, p. 153).

Flash detection efficiency and discharge location accuracy have been examined in a number of cross-sensor studies. One study that compared detections by a network of field mills with those by LDAR found that both systems reported a comparable number of flashes (38). The location accuracy of the LDAR system was tested by using aircraft that had a 66-MHz pulse generator as ground truth and by matching the LDAR location estimates to the coordinates of the aircraft by using the GPS (29). Figure 31 shows the three-dimensional accuracy with range that was obtained. Within the sensor baseline length of 10 km, the median errors are less than 100 m when the signal originates above 3 km, and are slightly higher when the signal originates below 3 km; at a 40-km range, the median error is ∼900 m. Much of this error is along a radial from the central site, and considerably less error is in the azimuthal component. Similar location error estimates have been obtained theoretically (87). Other test aircraft were used to examine the detection rates of LDAR. In an aircraft data set of more than 300,000 events, 97.4% were detected by the LDAR system. For those events within 25 km of the central site, the detection rate was more than 99%. Finally, flash detection efficiencies in the 90- to 200-km range were studied using observations from the satellite-based Lightning Imaging Sensor and the National Lightning Detection Network. LDAR flash detection efficiency in the medium to far range (50 to 300 km) varies from greater than 90% in the 90–100 km range but decreases to less than 25% at a 200-km range. Flash location errors in the medium to far range increase from 1 km at a 50 km range to ±7 km between the 175- and 250-km range (31).



Time tagging of LDAR events

Figure 30. Illustration of the time-tagging process used by LDAR to determine the time of arrival (TOA) of the lightning waveform. Site 0 in the center of the array (Fig. 29) detects a lightning waveform, whose peak occurs at time t0, and then opens a 100-μs window during which the other sites report detections. The waveforms are offset in time; site 1 reports a peak at time t1, site 2 at t2, etc. The differences between these peak times at the sites on the array perimeter and the time t0 provide the time tags used in the TOA calculations that determine the coordinates (x, y, z) of the source by finding where the hyperbolic surfaces intersect (Fig. 3). (Excerpt from LDAR Computer-Based Training, USAF/NASA/NWS Applied Meteorology Unit, ENSCO, Inc., used with permission.) See color insert.

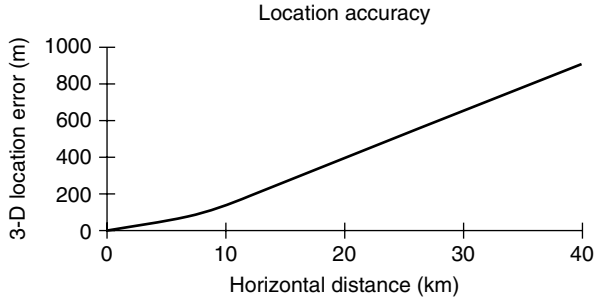


Figure 31. Three-dimensional location accuracy achieved by the LDAR network as a function of range from the central site 0. Within the sensor baseline of 10 km, the median three-dimensional error is ∼100 m when the source is above 3 km and increases to ∼900 m at a 40-km range. Most error is in the radial direction. (Excerpt from LDAR Computer-Based Training, USAF/NASA/NWS Applied Meteorology Unit, ENSCO Inc., used with permission.)

Aircraft flying in clouds in the vicinity of the network sometimes generate false alarms. Corona discharge from *triboelectrification*, the production of electrostatic charges by friction, of the aircraft can be detected by LDAR. Fortunately, the footprint of an aircraft on the display terminal has a distinctive three-dimensional, temporal evolutionary pattern that is easily discernible from natural events to the practiced eye. When these detections are removed from the dataset, the *false alarm rate*, which is the ratio of the number of incorrect detections to the total number of detections (156, pp. 240–241), is of the order of 0.01%.

Display Examples. A real-time example of two thunderstorms, as seen on the LDAR terminal at KSC, is shown in Fig. 32. In this figure, the lower left panel is a zoomed

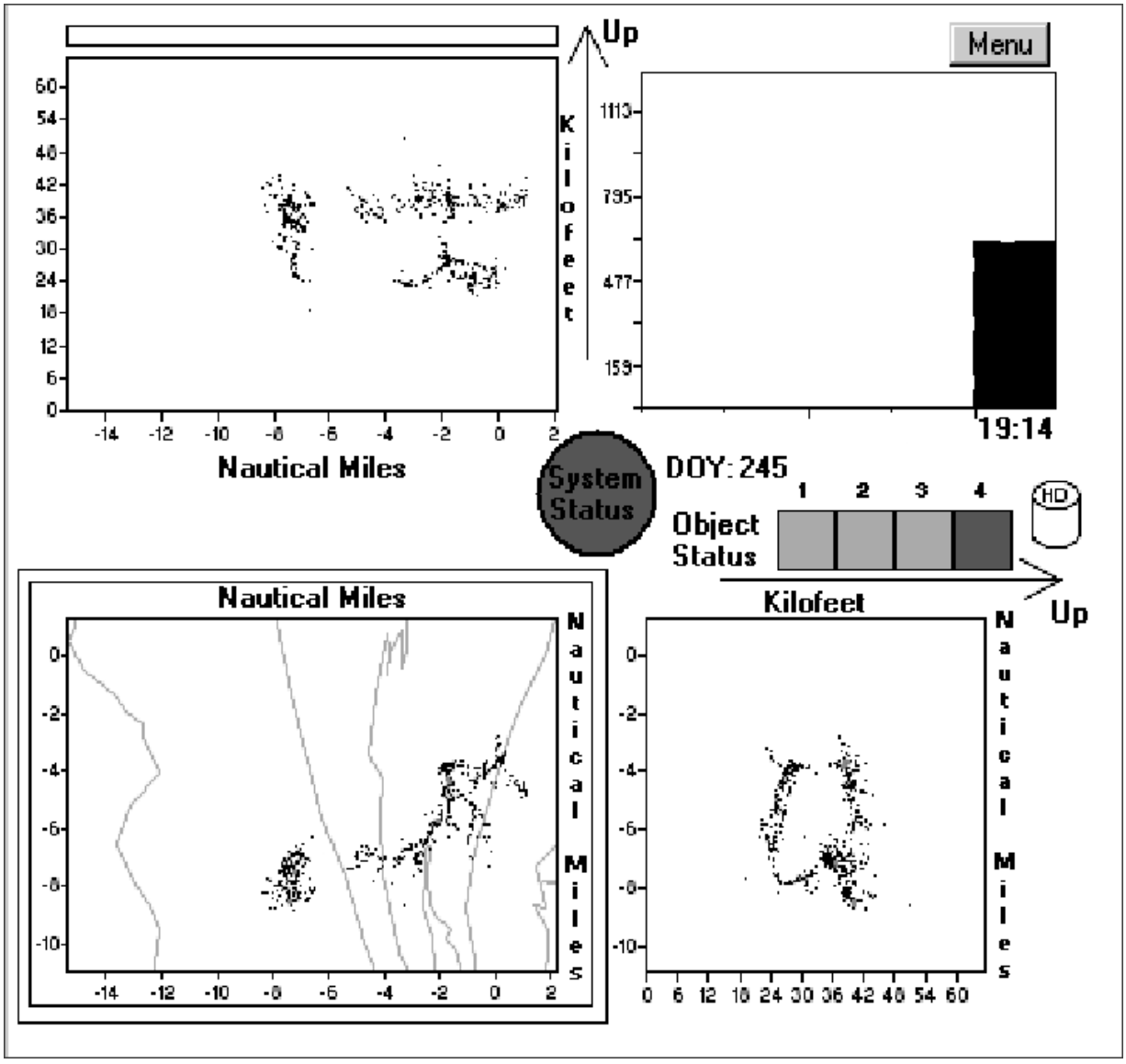


Figure 32. LDAR display illustrating a cloud–ground stroke. The lower left panel is a zoomed *x*,*y* depiction of part of the KSC complex approximately 15 nautical miles south of the central site, where *x* increases to the east and *y* increases to the north. The upper left panel is a view in the *x*,*z* plane as seen from the south, and the lower right panel is a cross-sectional view in the *y*,*z* plane as seen from the west; here *z* is the altitude above the ground in kilofeet. Each dot represents a pulse from a radio-frequency (RF) source, and a coherent collection of dots maps the lightning channel. Quasi-horizontal collections indicate cloud discharges, as seen between −4 and 0 nautical miles in the upper left panel; vertical collections indicate cloud–ground strokes, as seen at −7 nautical miles. The cloud–ground stroke here does not extend completely to the surface because the near-surface elevations are below the line of sight of the antennas in the network. The upper right panel provides a 1-minute count of LDAR events. (Excerpt from LDAR Computer-Based Training, USAF/NASA/NWS Applied Meteorology Unit, ENSCO, Inc., used with permission.) See color insert.

x, y, or *plan view*, depiction of part of the KSC complex approximately 15 nautical miles south of the central site, where x increases to the east and y increases to the north. The upper left panel is a *cross-sectional view* in the x, z plane as seen from the south, and the lower right panel is a cross-sectional view in the y, z plane as seen from the west; here z is the altitude above the ground. Conceptually, the upper left and lower right panels can be folded upward out of the page, so as to create a box in the lower left corner, in which the LDAR detections are projected onto the three sides. Each dot represents an *event*, which is normally taken as the tip of a negative stepped leader pulse within the lightning channel (86). The collection of dots maps the lightning channel; quasi-horizontal collections indicate cloud discharges and vertical collections indicate cloud–ground strokes. The upper right panel provides a 1-minute count of the number of LDAR events detected.

In the example in Fig. 32, there are two convective areas. The one to the west near −7 nautical miles is more concentrated in the x, y view (lower left panel) and has a distinctive cloud–ground flash seen by the alignment of the pixels in the x, z view (upper left). The signals originate in the mixed-phase region near 40,000 feet and then form a cone as the channel propagates toward the surface. Note the lack of data below 6,000 feet — consequence of the network's inability to detect events toward the horizon farther away from the central site. The CGLSS VLF/LF network of sensors (Fig. 28) that is collocated with the LDAR network identifies the strike point of the cloud–ground flashes, and this point is included on the real-time displays now in use. The cloud event in the right portion of the panels in Fig. 32 is diffuse in all projections, indicative of a decaying storm that has no concentration in signal and so no well-defined stroke channel.

A unique observation by the LDAR system at CCAFS/KSC is one in which a cloud–air discharge descended, well away from the storm edge, to the surface as a cloud–ground stroke (Fig. 33). In this figure, the cloud–air discharge, seen in the upper x, z panel, emanates from the concentrated region of signals approximately at the 25,000-foot level. After propagating westward for approximately 5 nautical miles, the discharge forks into a southward and upward moving cloud–air branch and a northwestward and downward moving cloud–ground stroke. Such *bolt-out-of-the-blue* strokes have also been seen by using the similar LMA/LMS deployable research system operated by New Mexico Tech (32,107).

The LDAR system enables forecasters and researchers to analyze the origins of all types of lightning flashes and to visualize their development and propagation in three-dimensional space. Although the CCAFS/KSC network is the only VHF network in continuous real-time use in the United States, its location in a region of Florida that has a high-thunderstorm frequency allows the development and examination of large data sets.

The KSC plans several upgrades to LDAR within the next several years. First, an additional antenna will be deployed near Orlando, Florida, approximately 40 nautical miles to the west of the KSC complex. This eighth antenna, it is hoped, will provide better warnings of lightning over the peninsula of Florida, where the concentration

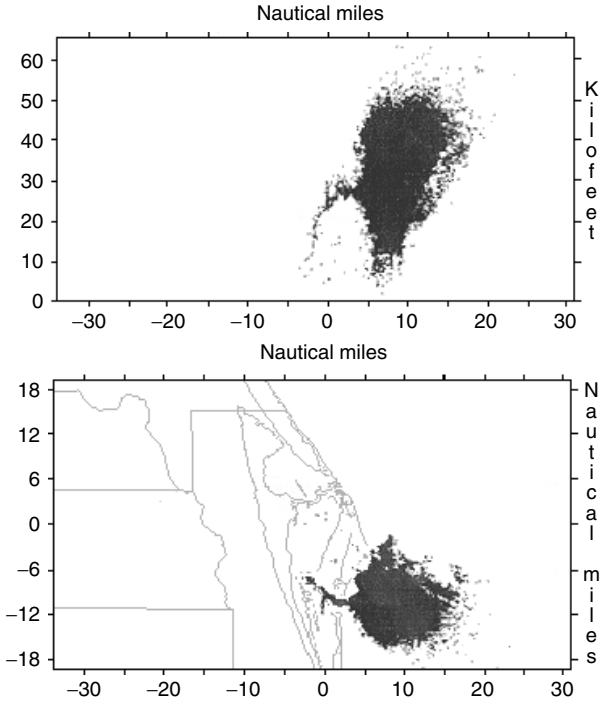


Figure 33. LDAR display of a bolt-out-of-the-blue. The x,z cross section depicted in the top panel and the x,y plan view in the bottom panel show a lightning channel extending horizontally about 5 nautical miles from the main cluster of events, where it splits; one branch extends southward and upward as a cloud–air flash, and the other branch extends northwestward and downward as a cloud–ground stroke. (Excerpt from LDAR Computer-Based Training, USAF/NASA/NWS Applied Meteorology Unit, ENSCO, Inc., used with permission.)

of storms is higher. This region is upstream of the KSC in the southwesterly flow regime that is predominant in the summertime (207). Second, studies are under way to attempt to link LDAR flash rates to the likelihood of severe weather, such has been done with OTD and SAFIR. Finally, Global Atmospherics, Inc., the NLDN provider, has expressed interest in developing LDAR commercially (30). Perhaps the future of LDAR is to provide public warnings of lightning.

SAFIR: AN INTERFEROMETRIC THREE-DIMENSIONAL NETWORK MAPPER

A second approach for mapping a lightning channel in three dimensions uses interferometric analysis of the radiated VHF signals to determine the bearing to the source (24,25,27). In contrast to the TOA mapping technique used by the Lightning Detection and Ranging (LDAR) system that detects discrete events lasting from 1 to 20 μs associated with negative stepped leaders, the interferometric (ITF) mapping technique detects the more continuous bursts of radio noise of <10-nanosecond (denoted ns) duration (27), which can last for tens to hundreds of microseconds; these are associated with stepped leaders, dart leaders, and K-processes or Q-trains (86,126,137). Interferometric techniques have been applied throughout the VHF

frequency band (30–300 MHz) in a number of independent efforts, and many have produced highly successful location results (24,25,27,65,126,137,208–212).

Radio Interferometry

Interferometric analysis requires at least one pair of antennas to determine the direction angle θ of the radial to the source of the lightning signal; two pairs of two antennas that have perpendicular baselines are needed to determine both θ and the angle of elevation of the radial in three dimensions (22, p. 155). As reviewed earlier in the section "Interferometry" and illustrated in Fig. 4, the output from the two antennas separated by a distance D — called the *baseline length* — are sent to a phase detector, which in turn produces a signal that is a function of the phase difference ϕ between the two signals (212). The classic phase detector acts by amplifying and mixing the output of each antenna using a local oscillator before multiplying these signals together (22, p. 156).

Assuming that the radiation propagates as a plane wave, one finds that the phase difference ϕ is a simple function of the direction angle θ, the wavelength of the radiation λ, and the baseline length D (22, p. 155; 137):

$$\phi = \frac{2\pi D}{\lambda} \cos\theta. \tag{6}$$

The antenna configuration described in Fig. 4 is for the special case when $D = \lambda/2$ for which $\phi = \pi \cos\theta$. Because bearings of either $+\theta$ or $-\theta$ are associated with the same phase angle ϕ, bearing ambiguity is inevitable. In addition, phases that necessarily vary only between 0 and 2π will not be associated with a unique value of $|\theta|$ if the baseline length D of the interferometer is large relative to the wavelength λ of the sampled radiation. Indeed, such ambiguities occur if $D > \lambda/2$, as is often the case in antenna configurations in interferometers. For 100 MHz VHF radiation traveling at the speed of light, $c = 3 \times 10^8$ m/s, the wavelength λ is 3 m. If the baseline length is $D = 4\lambda = 12$ m, then the equal phases ϕ of $\pm 2\pi$, $\pm 4\pi$, $\pm 6\pi$ and $\pm 8\pi$ are given by eight direction angles θ : 0°, 41.4°, 60°, 75.5°, 105.5°, 120°, 138.6°, and 180°. These cycles of repeating phase differences are referred to as *fringes* (22, p. 155; 137); in general there are $2D/\lambda$ such fringes, or eight in the preceding example, and one for the case illustrated in Fig. 4.

Because it is necessary to unravel the preceding direction angle ambiguities produced by baseline lengths that are much longer than the wavelength of the radiation, one might expect that short baselines would be optimal. It can be demonstrated, however, that baseline length is inversely proportional to the angular resolution that can be achieved from a measurement (22, p. 155). Reading the hands of a clock to tell the time provides a good analogy. The hour hand unambiguously gives the time of day; in principle, if it could be read with sufficient accuracy, then it would yield the exact time. Although the minute hand does not uniquely reveal the time by itself, the hand allows greater resolution of it. Using the two hands simultaneously provides the required temporal accuracy, and this combination suggests an appropriate baseline configuration for interferometric measurements (137,211). Two collinear baselines are used; one has a baseline length $D = \lambda/2$, resulting in one fringe and no ambiguities, but poor resolution, and one has a longer baseline length $D = 4\lambda$, resulting in one complete fringe over the resolution of the short-baseline antenna (22, pp. 156–157; 137,213). Other combinations of short and long baselines have been used to good advantage (108).

Once accurate direction angles and angles of elevation are established to define a radial to the radiation source, a second array of antennas located some distance away is used to obtain a second radial. These two radials establish the lightning source location in three dimensions by triangulation. Because the absolute uncertainty in location increases with the range from the antennas, it can prove difficult to achieve sufficient vertical resolution of the source locations for lightning channel mapping unless the antenna arrays are separated no more than a few tens of kilometers from each other and from the radiation source (22, p. 157). For example, three-dimensional mapping of spider intracloud lightning using interferometers is described in (108,109), as well as in (214), in which an array of two interferometers that are separated by 40 km were used. Plan view radiation location can be established over much greater ranges, sometimes as far as hundreds of kilometers (65).

The capabilities of an interferometric system are largely defined by parameters related to the configuration of their antenna arrays. The arrangements and baseline lengths are adjusted to best suit the type of, and the range to, the lightning phenomena under study. Also of importance are the choice of center frequency and bandwidth over which the receivers operate. Electromagnetic radiation at different frequencies corresponds to different physical phenomena and is also affected differently by local sources of noise. Because the operating frequencies f are related to the necessary dimensions D of the antenna baselines, which are a few multiples, or near-multiples (108), of the sampled wavelength $\lambda = c/f$, where c is the speed of light, systems at the higher end of the VHF range (300 MHz) will have shorter baselines (on the order of several meters) than systems operating near the lower end (30 MHz).

The bandwidth over which the signal is analyzed affects the accuracy of the results. Averaging results over broader frequency ranges should result in more accurate measurements of the phase differences, much as could be achieved by using multiple baselines (215). In addition, the time interval over which an incoming signal is averaged should help to improve the accuracy in determining the direction angles. Increasing the bandwidth and the sampling time interval, however, introduces uncertainties. Increasing the bandwidth can contribute to signal decorrelation across the array, which is not a problem when signals arrive at right angles to the baseline, but becomes an ever increasing problem as signals arrive at angles more closely parallel to the baseline (137). Emissions at different frequencies f are associated with different wavelengths λ and so yield different phase angles ϕ, except when they are perpendicular to the baseline (Fig. 4). As a result, at different sampling frequencies, waves appear to come from

different directions θ [Eq. (6)], even if their arrivals at the two receivers are nearly simultaneous. Increasing the time interval over which signals are averaged improves accuracy, but also diminishes the temporal resolution of the system and so increases the likelihood of averaging two distinct signals as one (126). Interferometric techniques are also somewhat limited by the electronic techniques required to calculate the phase difference between two signals (22, pp. 156–157).

Interferometric lightning location techniques have some unique advantages over those used by the other types of lightning location systems reviewed in this article. Interferometric techniques depend only on the phase of the incoming signal, not on the amplitude of the waveform (24). Thus, because propagation effects and noise can easily lead to significant difficulties in identifying the cause of particular signal amplitudes and waveforms (137), eliminating these variables greatly reduces errors introduced by local noise and propagation effects (212). VHF interferometric systems can typically achieve high temporal resolutions, of the order of tens of nanoseconds or less (27,213), and can thereby resolve nearly all stages of a lightning discharge. Using such high sampling rates, an interferometric system would have no problem distinguishing two separate storms in the same direction angle. Moreover, true three-dimensional characterization of the lightning sources is possible for systems that have high spatial resolution and short distances between detection stations. Even when vertical resolution cannot be established, the combination of high temporal resolution and the ability to detect many different types of lightning emissions can enable an interferometric system to obtain the plan view geometry of a lightning stroke that might extend horizontally for many kilometers.

The SAFIR System

Surveillance et Alerte Foudre par Interférométrie Radioélectrique (SAFIR) is an operational, real-time interferometric system that samples over 1-MHz bandwidths at a selectable center frequency between 110 and 118 MHz (22, p. 157). The temporal resolution of the system is high enough ($\sim$10 μs) to eliminate the possibility of unwanted interference between disparate sources of radiation. SAFIR, originally developed by the French Office National d'Etudes et de Recherches Aérospatiales (ONERA), is now a product of the company Dimensions SA; the research version of the SAFIR system is called the *ONERA three-dimensional interferometric mapper*, or *ONERA-3D* (27,108,214). In February 2000, Dimensions SA was purchased by Vaisala, an international company based in Finland that specializes in producing environmental and meteorological measuring equipment. The original purpose of SAFIR was to provide assistance to the European space program (22, p. 157), but in the last decade, Vaisala Dimensions SA has developed SAFIR into a commercially available system for detecting lightning discharges. SAFIR is used widely in Europe by many weather agencies and forecasting centers, the European Space Center in French Guiana, numerous hydrology offices, the aviation industry, the military, and several power companies (65), and it has also been used in Japan (26) and the United States (214) for fundamental lightning research and for comparisons with the performance of other lightning location techniques (86). Real-time lightning networks that use the same instruments as SAFIR are summarized in Table 1.

System Characteristics. The combination of high temporal resolution and three-dimensional radiation source location allows SAFIR to offer a unique set of detection parameters (65). SAFIR maps the three-dimensional structure of lightning channels that sometimes extend for tens of kilometers (24,32,33). SAFIR can detect all types of lightning phenomena that emit VHF radiation, in particular negative stepped and dart leaders, cloud–cloud recoil streamers, streamer–leader transitions, and possibly even stratospheric discharges (27,86,137). Cloud–ground and intracloud (including cloud–cloud and cloud–air) strokes are identified preliminarily via the orientation of the lightning channel. Time-domain waveform analysis allows distinguishing between intracloud and cloud–ground strokes by using low-frequency (LF) sampling techniques. Such analysis is used to provide an independent position analysis, peak current value, peak current derivative, stroke polarity, wave rise time, wave decay time, stroke energy estimate, and a neutralized charge estimate. As for the LDAR/CGLSS system, combining the LF and VHF techniques yields higher detection efficiencies, stroke-type classification, and a more precise location of cloud–ground strokes than either technique alone.

A wide input dynamic range, 100 dB, that is achieved at the detection stations allows detection of all lightning discharges over a broad area (65). Electrostatic field analyses have estimated that all of the SAFIR systems currently installed can achieve a 95% detection rate over a 200-km range from each detection station for low-elevation sources and over a 400-km range for high-elevation sources. The total lightning detection range that can be achieved by three or four detection stations is shown in Fig. 34.

As for LDAR, the accuracy of the source locations given by the SAFIR system has been verified using aircraft (65). This accuracy was verified by using a helicopter, equipped with a VHF transmitter radiating pulsed broadband VHF signals simulating lightning radiation. A helicopter-mounted radiation source was located by the SAFIR system and this was compared with the location given by tracking radar. Each interferometric detection station achieved a verified angular error of 0.5° in placing the radials in three dimensions (65), and the subsequent three-dimensional location on a short-baseline SAFIR system (30 km between detection stations) was 500 m rms and 750 m rms for a medium-baseline standard system (120 km between detection stations).

Application Examples. One example of the utility of the system is demonstrated by its ability to distinguish between cloud–ground and intracloud lightning phenomena (65). Figures 35a,b,c depict the lightning strokes from a storm as it progressed from the southwest to the northeast over the Netherlands (65). Each color represents a 1-hour increment; blue indicates the oldest, and

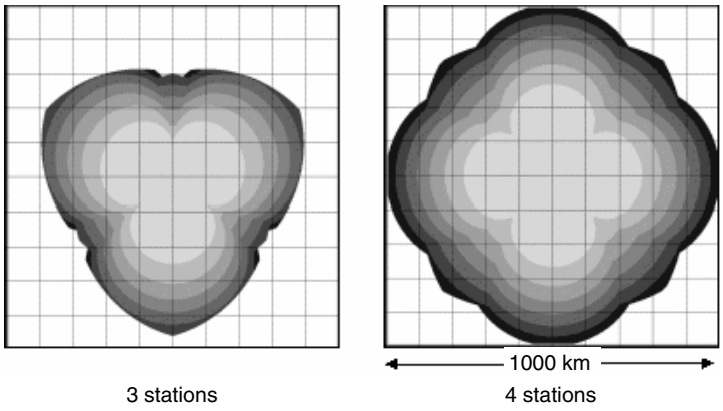


Figure 34. Detection ranges available for two SAFIR systems; that on the left uses three detection stations at the vertices of an equilateral triangle whose sides are 200 km long, and the other on the right uses four stations at the vertices of a square whose sides are 200 km long. The gray shading indicates how the location accuracy decreases as the range from the array of stations increases. [From (65)]. (Used with the permission of VAISALA DIMENSIONS SA.)

red the newest, observations. Figure 35a shows the complete locations of all lightning discharges. Note that some lightning is mapped as linear streaks, rather than just points in space, indicating that SAFIR has captured some of the horizontal extent of individual discharges. Figure 35b displays only the locations and polarity of the cloud–ground strokes that were detected and located by the combined interferometric/LF system. Figure 35c shows how the numbers of total lightning detections and cloud–ground strokes varied throughout the storm. The combination of these three figures reveals that the location of a large majority of the lightning during the first four hours of the storm was limited to cloud discharges. Shortly after the intracloud lightning strokes reached a maximum value, the rate of occurrence of ground strokes increased rapidly. These results suggest that observations of intracloud frequency can be used to help guide predictions of cloud–ground strokes, which are the type of

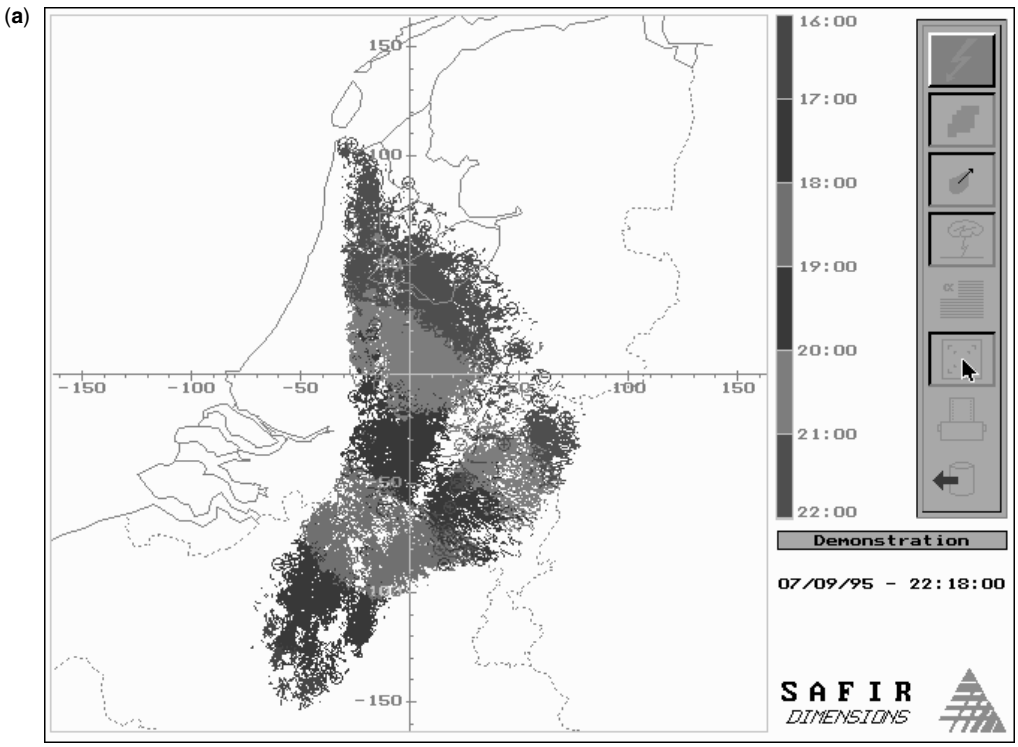
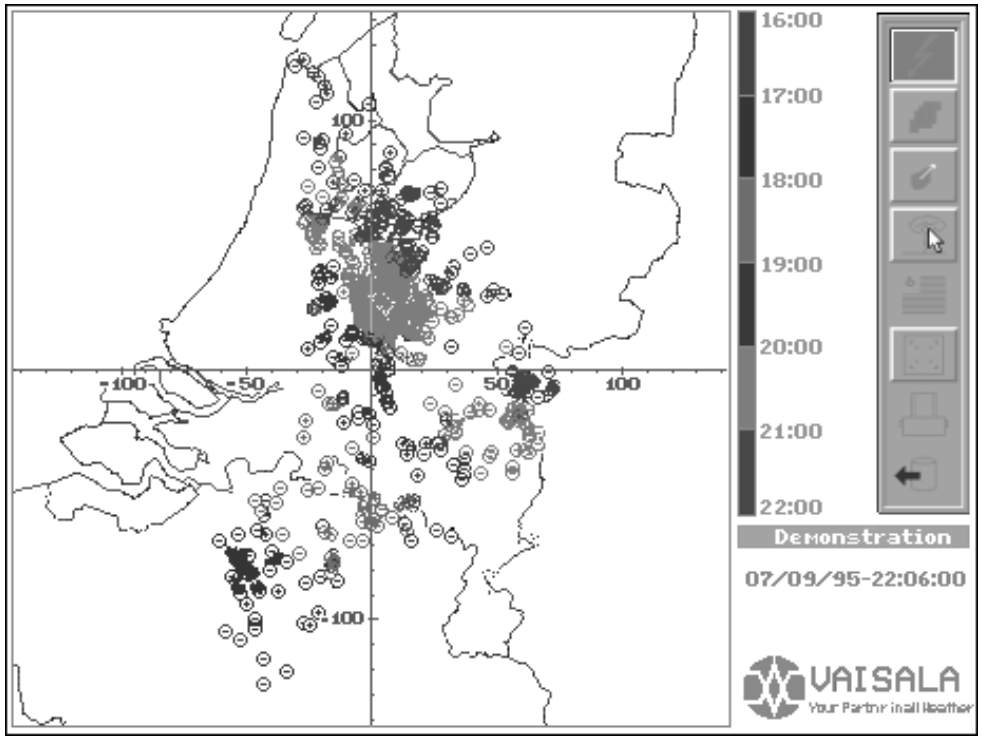


Figure 35. SAFIR plan view displays of total lightning detections (**a**), cloud–ground detections (**b**), and total number of strokes of both types (**c**). The colors in (**a**) and (**b**) indicate the time of occurrence of the lightning, in 1-hour increments; dark blue to the southwest in the lower left is the oldest, and red to the northeast in the upper right is the newest. Initially, there were few cloud–ground strokes. Between the third and fourth hour, however, the number of intracloud strokes decreased and presaged the increase in the number of cloud–ground strokes during hour 4. Using the trends in the number of intracloud strokes together with a projection of the storm movement based on the rate of advance of the lightning areas may lead to improved warning of cloud–ground strokes. [From (65)]. (Used with the permission of VAISALA DIMENSIONS SA.) See color insert.

(b)



(c)

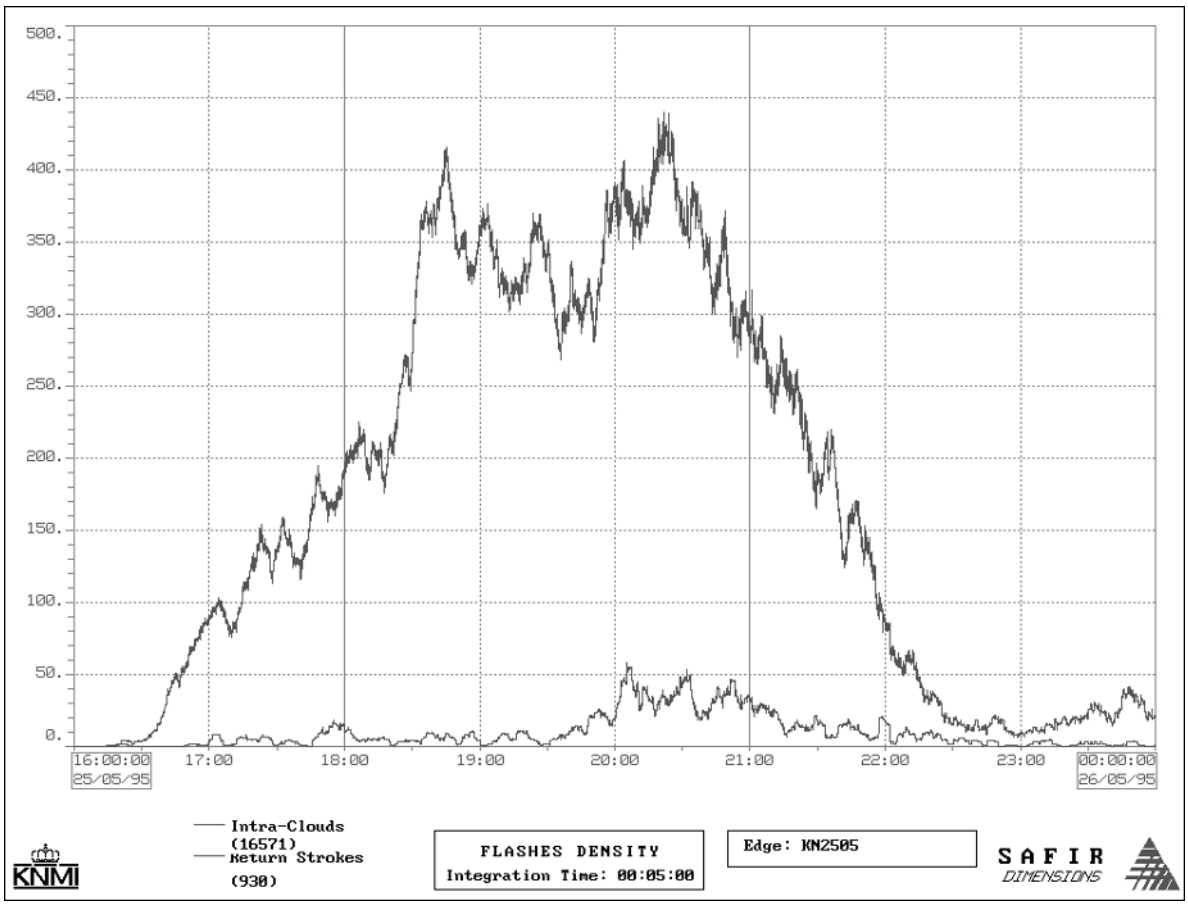


Figure 35. (*Continued*)

lightning hazard detected by VLF/LF systems such as the NLDN. Based on the information SAFIR revealed about this storm, it would have been possible to identify a dangerous and growing threat several hours in advance and then to predict the locations most likely to be severely affected by intense cloud–ground strokes.

A second and somewhat similar case study confirmed the utility of SAFIR as a predictor of other sorts of thunderstorm hazards (65). The study compared the number of intracloud strokes during a storm with the location of damaging downbursts from thunderstorms, which are regions of air that rapidly move downward until they reach the ground and then spread horizontally to cause straight-line wind damage. This study found that the peak in intracloud lightning precedes the most intense portion of the downburst by an average of 7 minutes. This result is consistent with similar studies using OTD data (47).

ABBREVIATIONS AND ACRONYMS

A	ampere
ac	alternating current
ALDARS	automated lightning detection and reporting system
ALEXIS	array of low energy X-ray imaging sensors
ALS	ASOS lightning sensor
AM	amplitude modulation
ARSI	Atmospheric Research Systems, Inc.
ASOS	automated surface observing system
ATD	arrival time difference
B-field	magnetic field
c	speed of light
C	coulomb
CCAFS	Cape Canaveral Air Force Station
CG	cloud–ground
CGLSS	cloud to ground lightning surveillance system
CGR3	cloud–ground ratio 3
CIGRE	Conference Internationale des Grands Reseaux Electriques
CLDN	Canadian Lightning Detection Network
cm	centimeter
CRDF	cathode ray direction finder; cathode ray direction finding
dB	decibels
dBZ	decibels relative to 1 mm^6/m^3; unit of radar reflectivity
dc	direct current
DF	direction finder; direction finding
DMSP	defense meteorological satellite program
DOE	department of energy
EDOT	*E*-field change sensor array
E-field	electric field
EFMS	electric field measurement system
ELF	extremely low frequency
EMP	electromagnetic pulse
ESID	electrical storm identification device
ESP	enhanced sensitivity and performance
EW	east/west
F	farads
FAA	federal aviation administration
FORTÉ	fast on-orbit recording of transient events satellite
GAI	Global Atmospherics, Inc.
GHz	gigahertz
GP-1	Great Plains-1
GPATS	global position and tracking systems®
GPS	global positioning system
h	hours
HF	high frequency
Hz	hertz
IC	intracloud
IMPACT	improved accuracy from combined technology
ITF	interferometer, inteferometry
kA	kiloamperes
kHz	kilometers
km	kilometers
KSC	Kennedy Space Center
kV	kilovolts
LANL	Los Alamos National Laboratory
LDAR	lightning detection and ranging
LF	low frequency
LIS	lightning imaging sensor
LLP	Lightning Location and Protection, Inc.
LMA	lightning mapping array
LMS	lightning mapping system
LPATS	lightning position and tracking system
LPLWS	launch pad lightning warning system
LRLDN	long-range lightning detection network
m	meters
MDF	magnetic direction finder; magnetic direction finding
MF	medium frequency
MHz	megahertz
min	minutes
ms	milliseconds
NALDN	North American Lightning Detection Network
NASA	National Aeronautics and Space Administration
NL	network locator
NLDN	National Lightning Detection Network™
nm	nanometer
NM	network mapper
NNBP	narrow negative bipolar pulse
NPBP	narrow positive bipolar pulse
ns	nanosecond
NS	north/south
NWS	National Weather Service
OLS	operational linescan system
ONERA	Office National d'Etudes et de Recherches Aérospatiales
OTD	optical transient detector
RF	radio frequency
RNSS	range-normalized signal strength
RTEP	real time event processor
s	seconds
SAFe	electric field measuring system marketed by dimensions SA
SAFIR	Surveillance et Alerte Foudre par Interférométrie Radioélectrique
SOLLO	sonic lightning location
SSC	single-station flash counter
SSL	single-station locator
TIPPs	transionospheric pulse pairs
TLM	transmission line model
TOA	time of arrival
TREMBLE	thunder recording employed in mapping branched lightning events
TRMM	tropical rainfall measurement mission satellite
TSS	thunderstorm sensor series
UHF	ultrahigh frequency

ULF	ultralow frequency
USAF	United States Air Force
USGA	United States Golf Association
UTC	Universal Time Coordinated
V	volts
VHF	very high frequency
VLF	very low frequency
Wb	webers
WD	lightning warning systems
WWW	World Wide Web
yr	year
μs	microseconds
Ω	ohms
2-d	two-dimensional
3-d	three-dimensional

BIBLIOGRAPHY

1. E. P. Krider, in J. R. Fleming, ed., *Historical Essays on Meteorology 1919–1995*, American Meteorological Society, Boston, 1996, pp. 321–350.
2. R. A. Watson-Watt and J. F. Herd, *J. Inst. Electr. Eng.* **64**, 611–622 (1926).
3. F. Horner, *Proc. Inst. Electr. Eng.* **101**, Part III: 383–390 (1954).
4. F. Horner, *Proc. Inst. Electr. Eng.* **104B**, 73–80 (1957).
5. E. T. Pierce, in R. H. Golde, ed., *Lightning*, vol. 1, Academic Press, NY, 1977, pp. 351–384.
6. E. P. Krider and R. C. Noggle, *J. Appl. Meteorol.* **14**, 252–256 (1975).
7. E. P. Krider, R. C. Noggle, and M. A. Uman, *J. Appl. Meteorol.* **15**, 301–306 (1976).
8. E. P. Krider, C. D. Weidman, and R. C. Noggle, *J. Geophys. Res.* **82**, 951–960 (1977).
9. C. D. Weidman and E. P. Krider, *J. Geophys. Res.* **83**, 6,239–6,247 (1978).
10. C. D. Weidman and E. P. Krider, *J. Geophys. Res.* **84C**, 3,159–3,164 (1979).
11. E. P. Krider, R. C. Noggle, A. E. Pifer, and D. L. Vance, *Bull. Am. Meteorol. Soc.* **61**, 980–986 (1980).
12. K. L. Cummins, R. B. Pyle, and G. Fournier, *Proc. 11th Int. Conf. Atmos. Electr.*, National Aeronautics and Space Administration, CP-1999-209261, Guntersville, AL, 1999, pp. 218–221.
13. K. L. Cummins et al., *J. Geophys. Res.* **103D**, 9,035–9,044 (1998).
14. R. S. Wacker and R. E. Orville, *J. Geophys. Res.* **104D**, 2,151–2,157 (1999).
15. R. S. Wacker and R. E. Orville, *J. Geophys. Res.* **104D**, 2,159–2,162 (1999).
16. W. P. Roeder, B. F. Boyd, and D. E Harms, *Proc. Conf. Lightning and Static Electr.*, National Interagency Coordination Group, Orlando, FL, 2000.
17. A. C. L. Lee, *J. Atmos. Oceanic Technol.* **3**, 630–642 (1986).
18. A. C. L. Lee, *Q. J. R. Meteorol. Soc.* **115**, 1,147–1,166 (1989).
19. N. Daley et al., *Preprints, 2000 Int. Lightning Detect. Conf.*, Global Atmospherics, Inc., Tucson, AZ, 2000; *http://www.glatmos.com/news/ildc_schedule.htm*.
20. J. A. Cramer and K. L. Cummins, *Proc. 11th Int. Conf. Atmos. Electr.*, National Aeronautics and Space Administration, CP-1999-209261, Guntersville, AL, 1999, pp. 250–253.
21. A. J. Sharp, *Proc. 11th Int. Conf. Atmos. Electr.*, National Aeronautics and Space Administration, CP-1999-209261, Guntersville, AL, 1999, pp. 234–237.
22. D. R. MacGorman and W. D. Rust, *The Electrical Nature of Storms*, Oxford University Press, NY, 1998.
23. R. S. Massey et al., *Proc. 11th Int. Conf. Atmos. Electr.*, National Aeronautics and Space Administration, CP-1999-209261, Guntersville, AL, 1999, pp. 684–687.
24. P. Richard, *Preprints, 16th Conf. Severe Local Storms and Conf. Atmos. Electr.*, American Meteorological Society, Boston, 1990, pp. J21–J26.
25. P. Richard, *Proc. 9th Int. Conf. Atmos. Electr.*, vol. III, International Commission on Atmospheric Electricity, St. Petersburg, Russia, 1992, pp. 925–928.
26. K. Kawasaki et al., *Geophys. Res. Lett.* **21**, 1,133–1,136 (1994).
27. G. Labaune, P. Richard, and A. Bondiou, *Electromagnetics* **7**, 361–393 (1987).
28. C. Lennon and L. Maier, *Proc. Int. Aerospace and Ground Conf. Lightning Static Electr.*, National Aeronautics and Space Administration, CP 3106, Cocoa Beach, FL, 1991, vol. II, pp. 89-1–89-10.
29. L. Maier, C. Lennon, T. Britt, and S. Schaefer, *Proc. 6th Conf. Aviation Weather Syst.*, American Meteorological Society, Boston, 1995, pp. 305–309.
30. D. J. Boccippio, S. Heckman, and S. J. Goodman, *J. Geophys. Res.* **106D**, 4,769–4,786 (2001).
31. D. J. Boccippio, S. Heckman, and S. J. Goodman, *J. Geophys. Res.* **106D**, 4,787–4,796 (2001).
32. W. Rison, R. J. Thomas, P. R. Krehbiel, T. Hamlin, and J. Harlin, *Geophys. Res. Lett.* **26**, 3,573–3,578 (1999).
33. P. R. Krehbiel et al., *Proc. 11th Int. Conf. Atmos. Electr.*, National Aeronautics and Space Administration, CP-1999-209261, Guntersville, AL, 1999, pp. 376–379.
34. R. J. Thomas et al., *Geophys. Res. Lett.* **27**, 1,703–1,706 (2000).
35. E. A. Jacobson and E. P. Krider, *J. Atmos. Sci.* **33**, 103–117 (1976).
36. D. E. Harms, B. F. Boyd, R. M. Lucci, and M. W. Maier, *Preprints, 10th Symp. Meteorol. Obs. Instrum.*, American Meteorological Society, Boston, 1998, pp. 317–322.
37. E. P. Krider, *J. Geophys. Res.* **94D**, 13,145–13,149 (1989).
38. J. S. Nisbet et al., *J. Geophys. Res.* **95D**, 5,417–5,433 (1990).
39. D. Mackerras, *J. Geophys. Res.* **90D**, 6,195–6,201 (1985).
40. D. MacKerras et al., *J. Geophys. Res.* **103D**, 19,791–19,809 (1998).
41. Y. Yair, Z. Levin, and O. Altaratz, *J. Geophys. Res.* **103D**, 9,015–9,025 (1998).
42. V. Cooray, *J. Geophys. Res.* **91D**, 2,835–2,842 (1986).
43. H. J. Christian, R. J. Blakeslee, and S. J. Goodman, *J. Geophys. Res.* **94D**, 13,329–13,337 (1989).
44. S. J. Goodman and H. J. Christian, in R. J. Gurney, J. L. Foster, and C. L. Parkinson, eds., *Atlas of Satellite Observations Related to Global Change*, Cambridge University Press, Cambridge, England, 1993, pp. 191–219.
45. H. J. Christian et al., *Proc. 10th Int. Conf. Atmos. Electr.*, Osaka, Japan, 1996, pp. 368–371.
46. D. J. Boccippio et al., *J. Atmos. Oceanic Technol.* **17**, 441–458 (2000).
47. E. Buechler, K. T. Driscoll, S. J. Goodman, and H. J. Christian, *Geophys. Res. Lett.* **27**, 2,253–2,256 (2000).

48. H. J. Christian et al., *Proc. 11th Int. Conf. Atmos. Electr.*, National Aeronautics and Space Administration, CP-1999-209261, Guntersville, AL, 1999, pp. 746–749.
49. A. R. Jacobson, S. O. Knox, R. Franz, and D. C. Enemark, *Radio Sci.* **34**, 337–354 (1999).
50. D. M. Suszcynsky et al., *J. Geophys. Res.* **105D**, 2,191–2,201 (2000).
51. A. R. Jacobson et al., *J. Geophys. Res.* **105D**, 15,653–15,662 (2000).
52. D. N. Holden, C. P. Munson, and J. C. Davenport, *Geophys. Res. Lett.* **22**, 889–892 (1995).
53. R. S. Zuelsdorf, R. C. Franz, R. J. Strangeway, and C. T. Russell, *J. Geophys. Res.* **105D**, 20,725–20,736 (2000).
54. G. Sampson, ASOS Lightning Sensor Assessment, Western Region Technical Attachment No. 97-31, NOAA, Salt Lake City, Utah, 1997; *http://www.wrh.noaa.gov/wrhq/97TAs/TA9733/TA97-33.html*.
55. R. Markson and L. Runkle, *Proc. 11th Int. Conf. Atmos. Electr.*, National Aeronautics and Space Administration, CP-1999-209261, Guntersville, AL, 1999, pp. 188–191.
56. D. W. Mazeroski, *MS Thesis, Penn State University*, 2000.
57. D. W. Mazeroski, H. N. Shirer, and H. W. Shirer, *Preprints, 2000 Int. Lightning Detection Conf.*, Global Atmospherics, Inc., Tucson, AZ, 2000, *http://www.glatmos.com/news/ildc_schedule.htm*.
58. V. A. Rafalsky, A. P. Nickolaenko, and A. V. Shvets, *J. Geophys. Res.* **100D**, 20,829–20,838 (1995).
59. P. J. Medelius, *System Locates Lightning Strikes to Within Meters*, NASA Technical Briefs Online, KSC-11785, Associated Business Publications, NY, 2000; *http://www.nasatech.com/Briefs/July00/KSC11992.html*.
60. E. R. Williams, S. G. Geotis, and A. B. Bhattacharya, in D. Atlas, ed., *Radar in Meteorology*, American Meteorological Society, Boston, 1990, pp. 143–150.
61. R. Kithil, *Proc. Int. Aerospace and Ground Conf. Lightning Static Elec.*, US. Navy, Naval Air Warfare Center, Aircraft Division, NAWCADPAX — 95-306-PRO, Williamsburg, VA, 1995, pp. 9-1–9-10 (70–79). Also published on the National Lightning Safety Institute web site at *http://www.lightningsafety.com/nlsi_lls/sec.html*.
62. R. Kithil, *Proc. 11th Int. Conf. Atmos. Electr.*, National Aeronautics and Space Administration, CP-1999-209261, Guntersville, AL, 1999, pp. 204–206.
63. R. L. Holle and M. A. Cooper, *Preprints, 2000 Int. Lightning Detection Conf.*, Global Atmospherics, Inc., Tucson, AZ, 2000; *http://www.glatmos.com/news/ildc_schedule.htm*.
64. W. P. Roeder, *Lightning Safety web page, 45th Weather Squadron, USAF, Patrick AFB and Cape Canaveral AFS*, Florida, 1999; *http://www.patrick.af.mil/45og/45ws/LightningSafety/index.htm*.
65. Vaisala Dimensions SA, SAFIR web page, Vaisala Company, Meyreuil, France, 2000; *http://www.eurostorm.com/*.
66. East-Central Florida Meteorological Research web site, National Weather Service, Melbourne, Florida, 2000, *http://WWW. SRH.NOAA.GOV/MLB/rsrchamu.html*.
67. S. J. Goodman et al., *Proc. 11th Int. Conf. Atmos. Electr.*, National Aeronautics and Space Administration, CP-1999-209261, Guntersville, AL, 1999, pp. 515–518.
68. E. M. Bazelyan and Y. P. Raizer, *Lightning Physics and Lightning Protection*, Institute of Physics, Bristol, 2000.
69. R. L. Gardner, ed., *Lightning Electromagnetics*, Hemisphere, NY, 1990.
70. R. H. Golde, ed., *The Physics of Lightning*, vol. 1, Academic Press, NY, 1977.
71. D. J. Malan, *Physics of Lightning*, The English Universities Press, London, 1963.
72. V. A. Rakov and M. A. Uman, *Lightning: Physics and Effects (Encyclopedia of Lightning)*, Cambridge University Press, NY, 2002.
73. M. A. Uman, *Lightning*, McGraw-Hill, NY, 1969; republished by Dover, NY, 1984.
74. M. A. Uman, *The Lightning Discharge*, Academic Press, Orlando, FL, 1987.
75. H. Volland, *Handbook of Atmospherics*, vol. 1, CRC Press, Boca Raton, FL, 1982.
76. H. Volland, *Handbook of Atmospherics*, vol. 2, CRC Press, Boca Raton, FL, 1982.
77. L. Wåhlin, *Atmospheric Electrostatics*, Research Studies Press, Letchworth, England, 1986.
78. M. Yamshita and K. Sao, *J. Atmos. Terrestrial Phys.* **36**, 1,623–1,632 (1974).
79. M. Yamshita and K. Sao, *J. Atmos. Terrestrial Phys.* **36**, 1,633–1,641 (1974).
80. M. A. Uman, Y. T. Lin, and E. P. Krider, *Radio Sci.* **15**, 35–39 (1980).
81. F. Horner, in J. A. Saxton, ed., *Advances in Radio Research*, vol. 2, Academic Press, NY, 1964, pp. 121–204.
82. B. D. Herrman, M. A. Uman, R. D. Brantley, and E. P. Krider, *J. Appl. Meteorol.* **15**, 402–405 (1976).
83. E. A. Lewis, R. B. Harvey, and J. E. Rasmussen, *J. Geophys. Res.* **65**, 1,879–1,905 (1960).
84. N. Cianos, G. N. Oetzel, and E. T. Pierce, *J. Appl. Meteorol.* **11**, 1,120–1,127 (1972).
85. D. E. Harms et al., *Preprints, 28th Conf. Radar Meteorol. Obs. Instrum.*, American Meteorological Society, Boston, 1997, pp. 240–241.
86. V. Mazur et al., *J. Geophys. Res.* **102D**, 11,071–11,085 (1997).
87. W. J. Koshak and R. J. Solakiewicz, *J. Geophys. Res.* **101D**, 26,631–26,639 (1996).
88. W. P. Roeder and C. S. Pinder, *Preprints, 16th Conf. Weather Anal. Forecasting*, American Meteorological Society, Boston, 1998, pp. 475–477.
89. S. Heckman, *Proc. 11th Inter. Conf. Atmos. Electr.*, National Aeronautics and Space Administration, CP-1999-209261, Guntersville, AL, 1999, pp. 719–721.
90. A. A. Few and T. L. Teer, *J. Geophys. Res.* **79**, 5,007–5,011 (1974).
91. R. H. Holzworth, in H. Volland, ed., *Handbook of Atmospheric Electrodynamics*, vol. 1, CRC Press, Boca Raton, FL, 1995, pp. 235–266.
92. P. R. Krehbiel, in Geophysics Study Committee, ed., *The Earth's Electrical Environment*, National Academy Press, Washington, D.C., 1986, pp. 90–113.
93. R. Mühleisen, in H. Dolezalek and R. Reiter, eds., *Electrical Processes in Atmospheres*, Dr. Dietrich Steinkopff, Darmstadt, 1977, pp. 467–476.
94. H. Volland, in L. J. Lanzerotti and M. Hill, eds., *Atmospheric Electrodynamics, Physics and Chemistry in Space*, vol. 11, Springer-Verlag, Berlin, 1984.
95. C. T. R. Wilson, *Philos. Trans. R. Soc. London, Ser. A* **221**, 73–115 (1920).

96. R. G. Roble and I. Tzur, in Geophysics Study Committee, ed., *The Earth's Electrical Environment*, National Academy Press, Washington, D.C., 1986, pp. 206–231.
97. T. C. Marshall and N. Stolzenburg, *J. Geophys. Res.* **106D**, 4,757–4,768 (2001).
98. E. R. Williams, *J. Geophys. Res.* **94D**, 13,151–13,167 (1989).
99. M. Stolzenburg, W. D. Rust, and T. C. Marshall, *J. Geophys. Res.* **103D**, 14,097–14,108 (1998).
100. P. R. Krehbiel, M. Brook, and R. A. McCrory, *J. Geophys. Res.* **84C**, 2,432–2,456 (1979).
101. D. E. Proctor, *J. Geophys. Res.* **96D**, 5,099–5,112 (1991).
102. Thunderstorm Technology, NASA/KSC Launch Pad Lightning Warning System web page, Huntsville, AL, 1997, *http://www.tstorm.com/lplws.html*.
103. S. A. Pierce, *J. Appl. Meteorol.* **9**, 194–195 (1970).
104. S. A. Prentice and D. MacKerras, *J. Appl. Meteorol.* **16**, 545–550 (1977).
105. I. I. Kononov, I. A. Petrenko, and I. E. Yusupov, *Proc. 25th Int. Lightning Protection Conf.*, Rhodos, Greece, 2000, 7 pp.; Also published at *http://thunderstorm.newmail.ru/publications/iclp2000/ICLP2000.htm*.
106. G. N. Oetzel and E. T. Pierce, in S. C. Coroniti and J. Hughes, eds., *Planetary Electrodynamics*, Gordon and Breach, NY, 1969, pp. 543–571.
107. P. R. Kriebel et al., *EOS* **81**, 21, 22, 25 (2000).
108. X. M. Shao and P. R. Krehbiel, *J. Geophys. Res.* **101D**, 26,641–26,668 (1996).
109. V. Mazur, X. M. Shao, and P. R. Krehbiel, *J. Geophys. Res.* **103D**, 19,811–19,822 (1998).
110. M. A. Uman et al., *J. Geophys. Res.* **80**, 373–376 (1975).
111. E. P. Krider, in Geophysics Study Committee, ed., *The Earth's Electrical Environment*, National Academy Press, Washington, D.C., 1986, pp. 30–40.
112. M. A. Uman et al., *Radio Sci.* **11**, 985–990 (1976).
113. C. Lucas and R. E. Orville, *Mon. Weather Rev.* **124**, 2,077–2,082 (1996).
114. O. Pinto Jr., et al., *Proc. 11th Int. Conf. Atmos. Electr.*, National Aeronautics and Space Administration, CP-1999-209261, Guntersville, AL, 1999, pp. 62–64.
115. J. C. Willett, J. C. Bailey, and E. P. Krider, *J. Geophys. Res.* **94D**, 16,255–16,267 (1989).
116. M. Ishii and J. -I. Hojo, *J. Geophys. Res.* **94D**, 13,267–13,274 (1989).
117. M. Le Boulch, J. Hamelin, and C. Weidman, in R. L. Gardner, ed., *Lightning Electromagnetics*, Hemisphere, NY, 1990, pp. 211–255.
118. Y. Villanueva, V. A. Rakov, and M. A. Uman, *J. Geophys. Res.* **99D**, 14,353–14,360 (1994).
119. D. M. LeVine, *J. Geophys. Res.* **85C**, 4,091–4,905 (1980).
120. D. A. Smith et al., *J. Geophys. Res.* **104D**, 4,189–4,212 (1999).
121. J. R. Bils, E. M. Thomson, M. A. Uman, and D. Mackerras, *J. Geophys. Res.* **93D**, 15,933–15,940 (1988).
122. R. S. Zuelsdorf et al., *Geophys. Res. Lett.* **25**, 481–484 (1998).
123. T. Ogawa, in H. Volland, ed., *Handbook of Atmospheric Electrodynamics*, vol. 1, CRC Press, Boca Raton, FL, 1995, pp. 93–136.
124. M. Brook and N. Kitagawa, *J. Geophys. Res.* **69**, 2,431–2,434 (1964).
125. C. D. Weidman, E. P. Krider, and M. A. Uman, *Geophys. Res. Lett.* **8**, 931–934 (1981).
126. P. Richard, A. Delannoy, G. Labaune, and P. Laroche, *J. Geophys. Res.* **91D**, 1,248–1,260 (1986).
127. V. A. Rakov and M. A. Uman, *J. Geophys. Res.* **95D**, 5,447–5,453 (1990).
128. M. Ishii, K. Shimizu, J. Hojo, and K. Shinjo, *Proc. 24th Int. Lightning Protection Conf.*, Birmhingham, UK, 1998, vol. 1, pp. 11–16.
129. D. M. Mach and W. D. Rust, *J. Geophys. Res.* **98D**, 2,635–2,638 (1993).
130. V. P. Idone and R. E. Orville, *J. Geophys. Res.* **87C**, 4,903–4,915 (1982).
131. K. Berger, R. B. Anderson, and H. Kroniger, *Electra* **80**, 23–37 (1975).
132. R. E. Orville, *J. Geophys. Res.* **96D**, 17,135–17,142 (1991).
133. V. P. Idone et al., *J. Geophys. Res.* **98D**, 18,323–18,332 (1993).
134. G. I. Serhan, M. A. Uman, D. G. Childers, and Y. T. Lin, *Radio Sci.* **15**, 1,089–1,094 (1980).
135. C. D. Weidman and E. P. Krider, *Radio Sci.* **21**, 964–970 (1986).
136. J. E. Nanevicz, E. F. Vance, and J. M. Hamm, in R. L. Gardner, ed., *Lightning Electromagnetics*, Hemisphere, NY, 1990, pp. 191–210.
137. C. T. Rhodes et al., *J. Geophys. Res.* **99D**, 13,059–13,082 (1994).
138. D. Levine and E. P. Krider, *Geophys. Res. Lett.* **4**, 13–16 (1977).
139. E. M. Thomson et al., *J. Geophys. Res.* **89D**, 4,910–4,916 (1984).
140. V. A. Rakov and M. A. Uman, *J. Geophys. Res.* **95D**, 5,455–5,470 (1990).
141. E. M. Thomson, *J. Geophys. Res.* **85C**, 1,050–1,056 (1980).
142. K. Berger, in R. H. Golde, ed., *Lightning*, vol. 1, Academic Press, NY, 1977, pp. 119–190.
143. V. Mazur and L. Ruhnke, *J. Geophys. Res.* **98D**, 12,913–12,930 (1993).
144. Y. T. Lin et al., *J. Geophys. Res.* **84C**, 6,307–6,324 (1979).
145. W. D. Rust, W. L. Taylor, D. R. MacGorman, and R. T. Arnold, *Bull. Amer. Meteorol. Soc.* **62**, 1,286–1,293 (1981).
146. W. D. Rust, W. L. Taylor, D. R. MacGorman, and R. T. Arnold, *Geophys. Res. Lett.* **8**, 791–794 (1981).
147. N. D. Murray, R. E. Orville, and G. R. Huffines, *Geophys. Res. Lett.* **27**, 2,249–2,252 (2000).
148. T. B. McDonald, M. A. Uman, J. A. Tiller, and W. H. Beasley, *J. Geophys. Res.* **84C**, 1,727–1,734 (1979).
149. M. J. Master, M. A. Uman, W. H. Beasley, and M. Darveniza, *IEEE Trans. PAS* **PAS-103**, 2,519–2,529 (1984).
150. M. Füllekrug, *Proc. 11th Int. Conf. Atmos. Electr.*, National Aeronautics and Space Administration, CP-1999-209261, Guntersville, AL, 1999, pp. 709–711.
151. J. Preta, M. A. Uman, and D. G. Childers, *Radio Sci.* **20**, 143–145 (1985).
152. R. M. Passi and R. E. López, *J. Geophys. Res.* **94D**, 13,319–13,328 (1989).
153. W. P. Roeder et al., *Proc. 11th Int. Conf. Atmos. Electr.*, National Aeronautics and Space Administration, CP-1999-209261, Guntersville, AL, 1999, pp. 238–241.
154. J. M. Cook, *Preprints, 9th Symp. Meteorol. Obs. Instrum.*, American Meteorological Society, Boston, 1995, pp. 110–112.

155. ASOS Automated Surface Observing System web page, National Weather Service, NOAA, Silver Spring, Maryland, 2000, *http://www.nws.noaa.gov/asos/*.

156. D. S. Wilks, *Statistical Methods in the Atmospheric Sciences*, Academic Press, San Diego, California, 1995.

157. Local Area Products web page, Global Atmospherics, Inc., Tucson, Arizona, 2000, *http://www.glatmos.com/products/local/localarea.html*.

158. ALDARS: Automated Lightning Detection And Reporting System web page, National Weather Service, NOAA, Pueblo, Colorado, 2000, *http://www.crh.noaa.gov/pub/ltg/Aldars.html*.

159. K. A. Kraus, T. A. Seliga, and J. R. Kranz, *Preprints, 16th Int. Conf. Interactive Inf. Process. Syst. (IIPS) Meteorol. Oceanic Hydrol.*, American Meteorological Society, Boston, 2000, pp. 106–109.

160. ASOS Version 2.60 Installation Status web page, National Weather Service, NOAA, Silver Spring, Maryland, 2000, *http://www.nws.noaa.gov/aomc/vers260.htm*.

161. H. W. Shirer and H. N. Shirer, Great Plains-1 Lightning Locator web page, Penn State University, University Park, Pennsylvania, 1998, *http://bub2.met.psu.edu/default.htm*.

162. T. Shütte, O. Salka, and S. Israelsson, *J. Climatology Appl. Meteorol.* **26**, 457–463 (1987).

163. H. J. Christian, *Proc. 11th Int. Conf. Atmos. Electr.*, National Aeronautics and Space Administration, CP-1999-209261, Guntersville, AL, 1999, pp. 715–718.

164. S. J. Goodman and H. J. Christian, in R. J. Gurney, J. L. Foster, and C. L. Parkinson, eds., *Atlas of Satellite Observations Related to Global Change*, Cambridge University Press, NY, 1993, pp. 191–219.

165. B. N. Turman, *J. Geophys. Res.* **83C**, 5,019–5,024 (1978).

166. Operational Linescan System (OLS) web page, Marshall Space Flight Center, NASA, Huntsville, Alabama, 1999, *http://thunder.msfc.nasa.gov/ols/*.

167. Space Research and Observations web page, Marshall Space Flight Center, NASA, Huntsville, Alabama, 1999, *http://thunder.msfc.nasa.gov/research.html*.

168. H. J. Christian et al., *Proc. 11th Int. Conf. Atmos. Electr.*, National Aeronautics and Space Administration, CP-1999-209261, Guntersville, AL, 1999, pp. 726–729.

169. Optical Transient Detector (OTD) web page, Marshall Space Flight Center, NASA, Huntsville, Alabama, 1999, *http://thunder.msfc.nasa.gov/otd/*.

170. L. W. Thomson and E. P. Krider, *J. Atmos. Sci.* **39**, 2,051–2,065 (1982).

171. H. J. Christian, R. J. Blakeslee, S. J. Goodman, and D. M. Mach, Algorithm Theoretical Basis Document (ATBD) for the Lightning Imaging Sensor (LIS), Marshall Space Flight Center, NASA, Huntsville, Alabama, 2000, *http://eospso.gsfc.nasa.gov/atbd/listables.html.*

172. OTD, LIS, and LDAR Data Browse Gallery web page, Marshall Space Flight Center, NASA, Huntsville, Alabama, 2000, *http://thunder.msfc.nasa.gov/data/*.

173. K. Driscoll, White Paper on Lightning Detection from Space, Marshall Space Flight Center, NASA, Huntsville, Alabama, 1999, *http://thunder.msfc.nasa.gov/bookshelf/docs/white_paper_driscoll.html*.

174. E. Williams et al., *Atmos. Res.* **51**, 245–265 (1999).

175. Lightning Imaging Sensor (LIS) web page, Marshall Space Flight Center, NASA, Huntsville, Alabama, 1999, *http://thunder.msfc.nasa.gov/lis/*.

176. T. Ushio et al., *Proc. 11th Inter. Conf. Atmos. Electr.*, National Aeronautics and Space Administration, CP-1999-209261, Guntersville, AL, 1999, pp. 738–741.

177. D. J. Boccippio, S. J. Goodman, and S. Heckman, *J. Appl. Meteo.* **39**, 2,231–2,248 (2000).

178. S. J. Goodman et al., *Geophys. Res. Lett.* **27**, 541–544 (2000).

179. LIS Ground Truth Browse Calendar web page, Marshall Space Flight Center, NASA, Huntsville, Alabama, 2000, *http://thunder.msfc.nasa.gov/lightning-cgi-bin/lisgt/lisgt_ref.pl*.

180. Lightning Mapping Sensor (LMS) web page, Marshall Space Flight Center, NASA, Huntsville, Alabama, 1999, *http://thunder.msfc.nasa.gov/lms/*.

181. D. M. Mach, D. R. MacGorman, W. D. Rust, and R. T. Arnold, *J. Atmos. Oceanic Technol.* **3**, 67–74 (1986).

182. R. E. Orville, R. W. Henderson, and R. B. Pyle, *Preprints, 16th Conf. Severe Local Storms Conf. Atmos. Electr.*, American Meteorological Society, Boston, 1990, pp. J27–J30.

183. W. A. Lyons, R. B. Bent, and W. F. Highlands, *Preprints, Int. Conf. Interactive Inf. Process. Sys. Meteorol. Oceanic Hydrol.*, American Meteorological Society, Boston, 1985, pp. 320–327.

184. IMPACT ESP sensor web page, Global Atmospherics, Inc., Tucson, Arizona, 2000, *http://www.glatmos.com/products/wide/impactESP.html*.

185. V. P. Idone et al., *J. Geophys. Res.* **103D**, 9,045–9,055 (1998).

186. V. P. Idone et al., *J. Geophys. Res.* **103D**, 9,057–9,069 (1998).

187. M. A. Uman, *J. Geophys. Res.* **90D**, 6,121–6,130 (1985).

188. M. Ference, H. B. Lemon, and R. J. Stephenson, *Analytical Experimental Physics*, 2nd rev. ed., University of Chicago Press, Chicago, 1956.

189. J. C. Willett, E. P. Krider, and C. Leteinturier, *J. Geophys. Res.* **103D**, 9,027–9,034 (1998).

190. R. E. Orville, *Mon. Weather Rev.* **119**, 573–577 (1991).

191. Lightning Climatology web page, in *Operational Uses of Lightning Data*, AWIPS Informational Series, National Weather Service Training Center, NOAA, Silver Spring, Maryland, *http://www.nwstc.noaa.gov/d.HMD/Lightning/climo.htm*.

192. G. N. Oetzel and E. T. Pierce, *Radio Sci.* **4**, 199–201 (1969).

193. H. R. Arnold and E. T. Pierce, *Radio Sci. J. Res.* **68D**, 771–776 (1964).

194. K. L. Zonge and W. H. Evans, *J. Geophys. Res.* **71**, 1,519–1,523 (1966).

195. E. T. Pierce, in F. Horner, ed., *Radio Noise of Terrestrial Origin*, Elsevier, Amsterdam, 1962, pp. 55–71.

196. T. O. Britt, C. L. Lennon, and L. M. Maier, *Lightning Detection and Ranging System*, NASA Technical Briefs Online, KSC-11785, Associated Business Publications, New York, New York, 1998; *http://www.nasatech.com/Briefs/Apr98/KSC11785.html*.

197. G. S. Forbes and S. G. Hoffert, in E. R. Hosler and G. Buckingham, eds., *1993 Research Reports, NASA/ASEE Summer Faculty Fellowship Program*, National Aeronautics and Space Administration, CR-199891, 1995, pp. 195–224.

198. D. E. Proctor, *J. Geophys. Res.* **76**, 1,478–1,489 (1971).

199. D. E. Proctor, *J. Geophys. Res.* **86C**, 4,041–4,071 (1981).

200. R. C. Murty and W. D. MacClement, *J. Appl. Meteorol.* **12**, 1,401–1,405 (1973).

201. W. L. Taylor, *J. Geophys. Res.* **83**, 3,575–3,583 (1978).
202. D. E. Proctor, R. Uytenbogaardt, and B. M. Meredith, *J. Geophys. Res.* **93D**, 12,683–12,727 (1988).
203. W. D. MacClement and R. C. Murty, *J. Appl. Meteorol.* **17**, 786–795 (1978).
204. GHRC User Services, *Lightning Detection and Ranging (LDAR) Dataset Summary*, Global Hydrology and Climate Center, Huntsville, Alabama, 2000; *http://ghrc.msfc.nasa.gov/uso/readme/ldar.html*.
205. L. M. Maier, E. P. Krider, and M. W. Maier, *Mon. Weather Rev.* **112**, 1,134–1,140 (1984).
206. D. J. Boccippio, S. Heckman, and S. J. Goodman, *Proc. 11th Inter. Conf. Atmos. Electr.*, National Aeronautics and Space Administration, CP-1999-209261, Guntersville, AL, 1999, pp. 254–257.
207. C. J. Neumann, *J. Applied Meteorol.* **10**, 921–936 (1971).
208. P. Richard and G. Auffray, *Radio Sci.* **20**, 171–192 (1985).
209. C. O. Hayenga, *J. Geophys. Res.* **89D**, 1,403–1,410 (1984).
210. V. Mazur, P. R. Krehbiel, and X. -M. Shao, *J. Geophys. Res.* **100D**, 25,731–25,753 (1995).
211. X. -M. Shao, P. R. Krehbiel, R. J. Thomas, and W. Rison, *J. Geophys. Res.* **100D**, 2,749–2,783 (1995).
212. J. W. Warwick, C. O. Hayenga, and J. W. Brosnahan, *J. Geophys. Res.* **84C**, 2,457–2,463 (1979).
213. C. Rhodes and P. R. Krehbiel, *Geophys. Res. Lett.* **16**, 1,169–1,172 (1989).
214. E. Defer, C. Thery, P. Blanchet, and P. Laroche, *Proc. 11th Int. Conf. Atmos. Electr.*, National Aeronautics and Space Administration, CP-1999-209261, Guntersville, AL, 1999, pp. 14–17.
215. X. -M. Shao, D. N. Holden, and C. T. Rhodes, *Geophys. Res. Lett.* **23**, 1,917–1,920 (1996).
216. A. Galván, V. Cooray, T. Götschl, and V. Scuka, *Proc. 11th Int. Conf. Atmos. Electr.*, National Aeronautics and Space Administration, CP-1999-209261, Guntersville, AL, 1999, pp. 162–165.
217. T. J. Tuomi, *Proc. 11th Int. Conf. Atmos. Electr.*, National Aeronautics and Space Administration, CP-1999-209261, Guntersville, AL, 1999, pp. 196–199.
218. S. Soula, G. Molinié, S. Chauzy, and N. Simond, *Proc. 11th Int. Conf. Atmos. Electr.*, National Aeronautics and Space Administration, CP-1999-209261, Guntersville, AL, 1999, pp. 384–387.
219. Y. Yair, O. Altaratz, and Z. Levin, *Proc. 11th Int. Conf. Atmos. Electr.*, National Aeronautics and Space Administration, CP-1999-209261, Guntersville, AL, 1999, pp. 460–463.

LIQUID CRYSTAL DISPLAY TECHNOLOGY

GREGORY P. CRAWFORD
MICHAEL J. ESCUTI
Brown University
Providence, RI

INTRODUCTION

As the display in most imaging systems is the final medium through which an image is rendered for manipulation and verification, an understanding of display technologies is essential to the imaging process. Because individuals working in the field of imaging science and technology may spend more time looking at a display screen than at anything else in their office or laboratory, it is imperative that it be comfortable to use and appropriate for the particular context.

Twenty years ago, system manufacturers often integrated the electronic display directly into the system to provide a complete package for a specified imaging application. Although this approach does not afford much flexibility, the display could be closely matched to a specific application because user requirements were well defined. This custom design approach enabled optimizing the graphics controller, system software, and user interface for the display, user, and application requirements. Despite the positive attributes of this "black-box" approach, such as high performance and superior application specific image quality, closed architecture platforms tend to be more expensive and suffer from incompatibility with peripheral add-ons and software packages not supported by the system manufacturer.

Today, the situation is dramatically different due to the continual evolution of the graphics controller interface. By mixing images with text and graphics, software developers require more from the display to support moving images without diminishing display performance for static images. The graphical capability of today's standard computer platforms has now made it unprofitable for vendors of imaging systems to develop their own displays for system-specific tasks. End users now typically purchase a computer platform, display, and a variety of other peripherals from multiple vendors and integrate them with ease (i.e., a plug-and-play philosophy). In such a marketplace, one must be well educated to match display technology to application needs. This article provides the reader with a fundamental knowledge of working principles of liquid crystal displays (LCDs), their capabilities, and their limitations.

ADDRESSING DISPLAYS

Before we delve into the operation of a LCD, it is important to understand how these displays are addressed and their impact on resolution, refresh rates, and image fidelity. Many treatises begin with material and device configurations, but we will first develop a basic understanding of electrical addressing schemes that apply to all LCDs. Our hope is that the reader will be better prepared to recognize the capabilities and limitations of the various display configurations presented afterward.

A LCD with high-information content (e.g., computer or television screen) consists of a two-dimensional array of pixels, where a pixel is defined as the smallest switching element of the array. If the two-dimensional array has a total of N rows and M columns ($N \times M$ pixels), then in principle, there can be $N \times M$ electrical connections to control each pixel independently. This is known as *direct addressing* and is practical only for very low-resolution displays. For medium and higher resolution displays,

addressing is accomplished through *passive-* and *active-matrix* techniques. Both of these approaches require only $N + M$ electrical connections, thereby greatly simplify the electronics, and make higher resolution possible.

Luminance–voltage plots for three hypothetical displays are depicted in Fig. 1. This display characteristic ultimately dictates the type of addressing that can be used to create images by using a LCD. Luminance is the physical measure of brightness of a display or any surface and most commonly has units of candelas per squared meter (cd/m^2), nits, or footlamberts (fL). The two measurable quantities from the luminance–voltage curve that have the greatest impact on display addressing are the threshold voltage V_{TH} (the voltage at which the luminance begins to increase) and a parameter Δ (the additional voltage beyond V_{TH} needed to cause the display to approach or reach its highest luminance). If a liquid crystal (LC) material does not start to respond to an electronic stimulus until it has reached a well-defined voltage, then it is said to have a threshold; otherwise, if the display material responds to all voltages, then it is said to be thresholdless (1).

For simple direct addressing schemes, like the seven-segment digit electrodes shown in Fig. 2, the threshold(less) nature of the material is irrelevant because the segmented electrodes (or pixels) are independent of each other. The appropriate combinations of segments are addressed by dedicated logic circuitry (i.e., every pixel is independently driven by its own external voltage source), and the screen refresh rate is only as long as needed for a single pixel to switch. Direct addressing is practical only for low-resolution displays (<50 pixels).

In a passive addressing scheme (also known as multiplexing), one substrate has row electrodes, and the other substrate has column electrodes, as shown in Fig. 3.

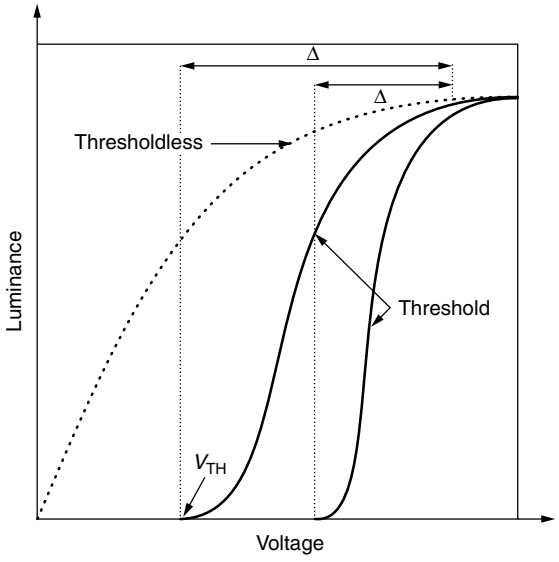


Figure 1. Luminance–voltage graph depicting the difference between materials that have well-defined thresholds and those materials that have no threshold (thresholdless). V_{TH} is defined as the voltage at which the luminance begins to increase, and a parameter Δ is defined as the additional voltage beyond V_{TH} needed to cause the display to approach or reach its highest luminance.

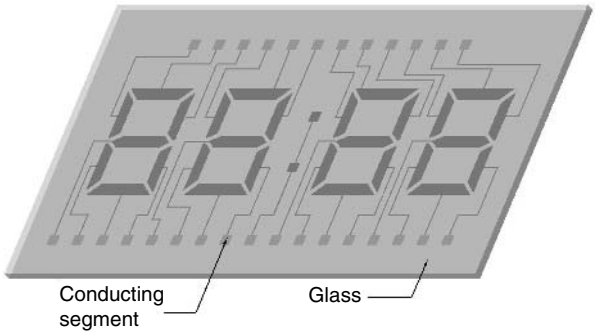


Figure 2. An example of direct addressing, where the segments are independently driven to create low-resolution information.

Every pixel is uniquely determined by the region of overlap of a row and a column electrode, making it possible to access $N \times M$ pixels with only $N + M$ connections. A display that has a passive matrix is driven one line at a time; that is, one row is selected for addressing, and all of the columns are addressed using voltages associated with the image for that row. At some time interval later, the next row receives an appropriate voltage pulse, and the columns again are addressed using the information required for that row. The net result is that a pixel is influenced only sufficiently to produce an optical effect when the time-averaged voltage [called the root-mean-square (rms)] across the row and column electrodes is beyond the threshold ($V_{ON} \geq V_{TH}$). All of the rows not being updated are driven by row voltages that will not affect the image information already present. The catch is that there will always be a voltage on the non selected rows, and therefore that voltage must satisfy the relationship $V_{OFF} \leq V_{TH}$ so as not to induce an optical change. Therefore, a well-defined threshold in the luminance–voltage characteristic of the LCD material is necessary to prevent the non selected rows from being addressed (i.e., cross talk). LC materials typically exhibit threshold behavior, which many displays to be multiplexed; display materials that are thresholdless, such as electrophoretics (2) and gyricon (3), cannot.

An expression can be derived that completely specifies the maximum number of rows N_{MAX} that can be addressed in terms of the voltage threshold and the parameter Δ, a measure of the nonlinearity of the LC material:

$$\frac{\Delta}{V_{TH}} \leq \frac{1}{\sqrt{N_{MAX}}}. \tag{1}$$

To maximize the number of addressable rows for higher resolution, one usually uses a material that has a very non-linear and steep luminance–voltage response, small Δ, rather than using materials that have a large V_{TH} (which increases power consumption).

Another important equation relates the rms voltages of the ON-state V_{ON}, and the OFF-state, V_{OFF}, to the maximum number of rows that is given by (for $N_{MAX} \gg 1$):

$$\frac{V_{ON}}{V_{OFF}} \cong 1 + \frac{1}{\sqrt{N_{MAX}}}, \tag{2}$$

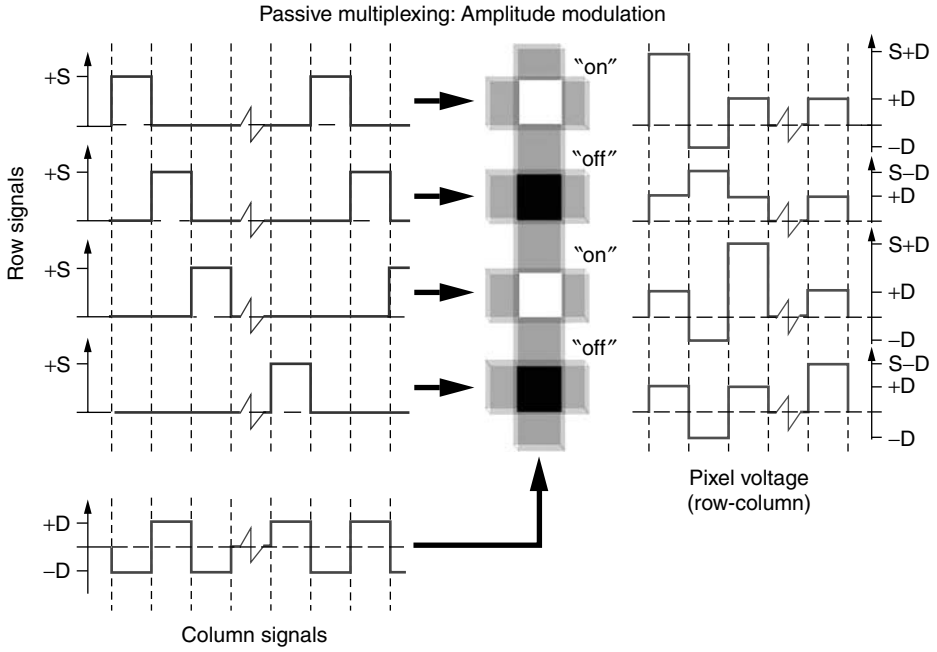


Figure 3. A simple example of multiplexing on an N-row matrix showing the row addressing waveforms, column addressing waveforms, and the corresponding pixel waveforms. This type of multiplexing, known as amplitude modulation, varies the amplitude of the pixel waveform to achieve the desired image. The normally black pixel is "ON" if the rms average pixel voltage exceeds V_{TH}. See color insert.

where V_{ON}/V_{OFF} is referred to as the selection ratio. When N_{MAX} is large, this ratio approaches unity, leading to poor contrast because the difference would decrease between the luminance in the ON- and OFF-states. Therefore, Eq. (2) dictates the optical contrast ratio (CR) of the display, which is defined as the ratio of the luminance in the ON-state and the luminance in the OFF-state ($CR = L_{ON}/L_{OFF}$). Equations (1) and (2) were first derived by Alt and Pleshko in 1974 and remain the fundamental expressions governing multiplexing (4). To increase resolution using passive addressing techniques, a technique known as "dual" scan can be implemented. In the dual-scan approach, two column drivers are used to address the upper $N/2$ rows and lower $N/2$ rows. Examination of Eq. (1) shows that the dual-scan approach improves the ratio Δ/V_{TH} by a factor of $\sqrt{2}$ (5).

Although some LC configurations that use the multiplexing technique have reasonable contrast for large values of N, other limitations enter into consideration such as response time (the time taken for a pixel to be switched to the fully ON-state plus the time needed to relax to a completely OFF-state), viewing angle (the maximum polar angle for which a display maintains reasonable contrast and minimal color degradation), and gray-scale issues (those that involve the operation of a pixel at a luminance which is intermediate between the ON- and OFF-states). The frame rate in a passively addressed display is severely limited by the time it takes for a single row to switch, as well as the number of rows in the display, because the frame rate must be greater or equal to the product of the two. In only a few exceptions, displays that use passive addressing will not support full video frame rates. However, this problem can be essentially eliminated using an active-matrix addressing scheme, where a nonlinear element is used for pixel isolation.

Active-matrix addressing is recognized by the display industry as the ultimate solution for high fidelity, high information content, full color, and significant gray-scale applications. Additionally, this addressing approach can be used with thresholdless and large Δ materials because a discrete nonlinear switch is integrated into each pixel structure. An active-matrix LCD incorporates a two-dimensional circuit array (or matrix) to provide the electrical addressing of individual pixels (6). This matrix incorporates an active device in each pixel, usually a thin-film-transistor (TFT), positioned at one of the corners. Almost all LCD pixels are essentially dielectric capacitors whose leakage is minimal when a charge is placed on the electrodes through the transistor. And due to the electrical isolation afforded by the transistor, the voltage on one pixel remains constant while other pixel elements are subsequently addressed; therefore, the Alt–Pleshko limitation expressed by Eq. (2) does not constrain the contrast ratio, as it does in passive-addressing schemes.

A schematic diagram of an active-matrix circuit is shown in Fig. 4, where each pixel element is defined by the overlap of row and column bus lines. The circuit diagram shows that each pixel has one TFT and a LC capacitor formed between a top conducting surface, (typically indium tin oxide (ITO)), and the active-matrix substrate. This approach is significantly more complex than passive addressing, as can be seen from the cross section of an active-matrix TFT display also shown in Fig. 4. The intricate underpinnings of active-matrix addressing and the complex processing required to create one are beyond our scope, but the basic operation can be understood as follows. The display is addressed one line at a time, as in passive addressing. When a row (called the scan or gate line) is addressed, a positive voltage pulse of duration T/N (where N is again the number of rows and T is the frame time) is applied to the line and turns on all transistors in the row. The transistors act as switches that transfer electrical charges to the LC cells from the columns (called data or source lines). When subsequent

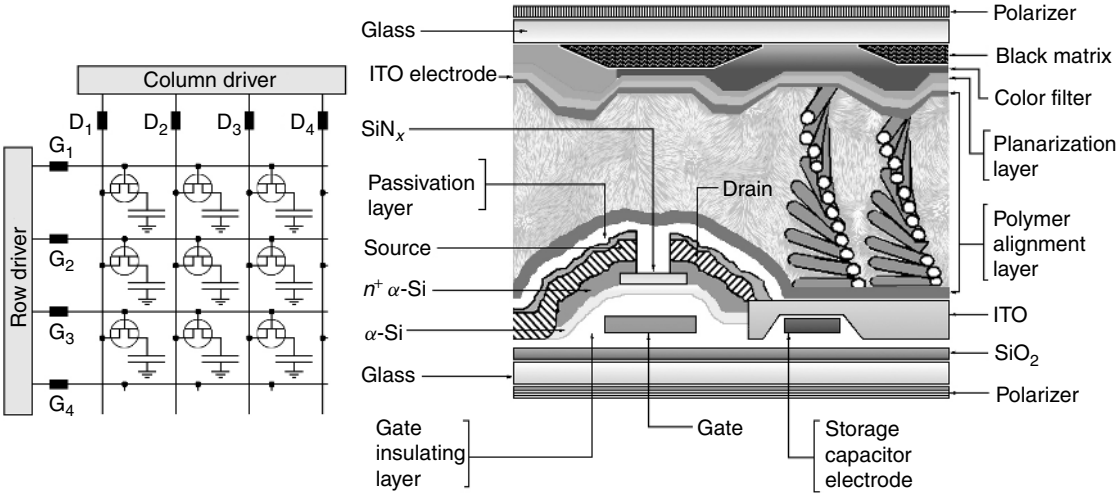


Figure 4. The driving circuit of an active-matrix thin-film transistor next to the cross section of an active-matrix pixel. As is apparent from the pixel structure becomes substantially more complex when active-matrix schemes are used. See color insert.

rows are addressed, a negative voltage is applied to the gate line that turns off all of the transistors in the row and holds the electrical charges in the LC capacitors for one frame until the line is addressed again. Alternating row-select voltages are required for most LCD materials, and the polarity of the data voltage is usually switched in alternate frames. The refresh rate in this addressing scheme is not limited by the number of rows but only by the response time of the LC.

A nonlinear pixel element can be implemented by a variety of approaches, but we will mention only the two most prominent active-matrix types: amorphous silicon (α-Si) and polycrystalline silicon (poly-Si). Both of these involve complex fabrication of thin films of silicon structures (<300 nm) on glass or quartz substrates. The details of active-matrix development are sufficiently complex to be far beyond the scope of this review, but we include a simple introduction to familiarize the reader with the basics.

Pioneering work on the active matrix began in the early 1970s (7), and the now conventional α-Si approach was first proposed in 1979 (8), where the deposited thin film of silicon has a small grain structure and is randomly configured (9). In this approach, inexpensive glass substrates are typically processed at $\sim$300 °C, using well established processes, including sputtering, photolithography (typically 5–8 photomasks), plasma-enhanced chemical vapor deposition (PECVD), wet chemical etch, and reactive ion etching (REI). The resulting electron mobility is low ($\sim$0.5 cm^2/V/s), but it is more than adequate to form a useful switch for the pixel (9). However, there are two prominent limitations. First, the aperture ratio (the transmissive area of the pixel divided by the total pixel area) of the display is decreased to 50–75% due to the use of an absorbing light shield to overcome the photoconductivity of the α-Si TFT. Second, because the supporting row- and column-drivers must be mounted on additional IC chips bonded to the display substrate, the connections (>4,000) increase fabrication complexity and decrease reliability. Nonetheless, α-Si active-matrix displays have been made in all sizes and are most popular as laptop monitors and other large-area LCDs. Current research issues include reducing the number of photomasks (10), increasing display resolution (11), and increasing the aperture ratio (12).

In contrast, a poly-Si substrate can yield TFTs whose electron mobility is much higher ($\sim$440 cm^2/V/s), but must be processed at significantly higher temperatures (6). Fabrication typically involves the same α-Si process described before on a more expensive quartz substrate. Additional processing at the higher temperatures leads to recrystallization in a furnace or by laser annealing (13,14). An intriguing approach uses a laser ablation/annealing approach where CMOS-TFTs are fabricated at high temperatures using a quartz substrate and are subsequently transferred to flexible plastic substrates without any noticeable deterioration in poly-Si performance (15). In all approaches, the silicon grain becomes larger and more uniform and allows electrons to flow much more freely. The greatest benefits of these poly-Si substrates are the ability to fabricate row- and column-drivers directly on the periphery of the glass substrates and the reduction of the TFT size to $\sim 5 \times 5$ μm. Additionally, the aperture ratio can be made substantially higher. Disadvantages include the high process temperature, increased fabrication complexity (higher accuracy required in photolithography and ion implantation), and the higher off-leakage current. Integration of driver electronics onto the substrate, increased display brightness, lower power consumption, and the ability to form smaller pixels at higher densities (>200 dpi) make these poly-Si displays particularly useful for microdisplays and medium size displays.

Display addressing directly impacts resolution and optical performance. Because most imaging applications require high resolution, active-matrix addressing is the most prominent addressing approach in the imaging field. Because of the complexity of the substrate, active-matrix

addressing always involves more cost, but it enables a high-resolution regime. A rule of thumb for LCD technology is that passive addressing can achieve $N_{MAX} \leq 400$, whereas active addressing can achieve $N_{MAX} \geq 1,000$.

PROPERTIES OF LIQUID CRYSTAL MATERIALS

To understand the nuts and bolts of a LCD, it is worthwhile to review briefly the material properties that make LCs functional for display applications. The liquid crystal phase is a state of matter that is intermediate between a solid crystal and an isotropic liquid; in fact, it has properties of both. The geometry of a LC molecule is highly anisotropic; these molecules can be four to six times longer than they are wide and are often modeled as rigid rods (16). The nematic phase is the simplest LC phase and is the most used in commercial display applications. This phase possesses only orientational order along the long axes of the elongated molecules, but no positional or bond orientational order, as in a conventional crystal. We discuss the ferroelectric smectic C* liquid crystal phase later, so we will limit our discussion here to the nematic phase.

The elastic properties of LCs are their most characteristic feature. At the display level, elastic theory is used to predict stable configurations and electric field-induced elastic deformations of the material that are responsible for the image. The elastic theory expression is often written in the following form:

$$f = \tfrac{1}{2}\{K_{11}(\nabla \cdot \boldsymbol{n})^2 + K_{22}(\boldsymbol{n} \cdot \nabla \times \boldsymbol{n})^2 + K_{33}(\boldsymbol{n} \times \nabla \times \boldsymbol{n})^2 - \varepsilon_0 \Delta\varepsilon(\boldsymbol{E} \cdot \boldsymbol{n})^2\}. \quad (3)$$

Here, f is the free energy density; $\boldsymbol{n}$ is the nematic director; $\boldsymbol{E}$ is the applied electric field; K_{11}, K_{22}, and K_{33} are known as the splay, twist, and bend elastic constants, respectively; and $\Delta\varepsilon$ is the dielectric anisotropy. The nematic director is denoted as $\boldsymbol{n}$ and represents the average orientational symmetry axis of an ensemble of LC molecules. The elastic constants are typically of the order of 10^{-11} N, the dielectric anisotropy $\Delta\varepsilon$ is typically $\sim$5–15 for most display materials, and $\boldsymbol{E}$ is typically <1 V/μm for conventional transmissive LCD technology. Display performance parameters can be predicted by minimizing Eq. (3) with respect to the boundary conditions imposed by the surfaces of the display substrates to obtain the configuration that has the lowest free energy. A number of mathematical representations can be used to calculate the device properties of LCDs (17). Equation (3) is used to predict the details of the field-aligned configuration of the LC material and its threshold voltage. The switching dynamics can be determined by balancing the elastic and electric field torques derived from Eq. (3) and the viscous torque of the material $-\gamma_1(d\theta/dt)$, where γ_1 is the rotational viscosity. The rotational viscosity γ_1 is typically from 1–2 poise for display materials. The switching time is highly dependent on the choice of LC material and configuration but is typically in the 1–30 ms range. The reader is referred to the following references to explore basic LC properties further (18–30).

The techniques involved in transforming the configurations derived from elastic theory and into the optical properties of LCDs have been treated extensively and will only be summarized here. In addition to their geometrical anisotropy, nematic LCs used in displays also possess optical anisotropy, or birefringence, that is responsible for the optical changes in contrast when one configuration is electrically switched to another configuration. The birefringence Δn for LCs is in the range of 0.05–0.3; $\Delta n \sim 0.1$ is often used for display applications because it leads to a high contrast design known as the first minimum or maximum (see later). The 2×2 Jones matrix technique (31) can be used to calculate the transmission of light through a LCD at normal incidence but neglects Fresnel diffraction and multiple reflection effects in thin layers. The more sophisticated 4×4 (32) method represents a complete solution of Maxwell's equations and is used for the general case of light transmission through a LCD at any angle. Nematic LC configurations pertinent to displays are most often modeled by this latter method by the display community (33).

LIQUID CRYSTAL DISPLAY CONFIGURATIONS

Having presented the basics of display addressing and the basic properties of nematic LCs, we can now introduce the various configurations relevant to displays. It would be impossible to describe, even briefly, every type of current LCD configuration, so we will discuss the operational principles of a few that are most relevant to imaging. Broadly speaking, LCDs can be classified as transmissive or reflective. A transmissive display is a light shutter that modulates a powerful backlight. Reflective displays take advantage of ambient lighting to reflect light back to the viewer effectively and therefore do not require a backlight. Transflective displays operate in both modes. Removing the backlight can result in significant power saving and therefore longer battery life. These are largely being developed for portable applications rather than high-resolution imaging. Therefore, we will focus on transmissive displays and end by presenting a promising low-power reflective display technology.

Twisted Nematic (TN) Liquid Crystal Displays

The most commercially successful LCD configuration, known as the twisted nematic (TN), uses cross-polarizers and a molecular orientation of the molecules whose long axis twists through a 90° angle between the orienting substrates. A cross-section of a TN-LCD, presented in Fig. 5, includes two glass substrates that have polarizers laminated on the outer surfaces and an ITO conducting layer on the inner surfaces covered by a polymer (e.g., polyimide). The polyimide layer is mechanically rubbed with a cloth to create microgrooves on the surface that uniformly align the long axis of the LC molecules at the surface (34). The alignment of the rub direction of the two substrates is placed parallel to the transmission axis of the respective laminated polarizer, but perpendicular to each other. The surface alignment mechanisms introduced by the polyimide then result in a LC layer that

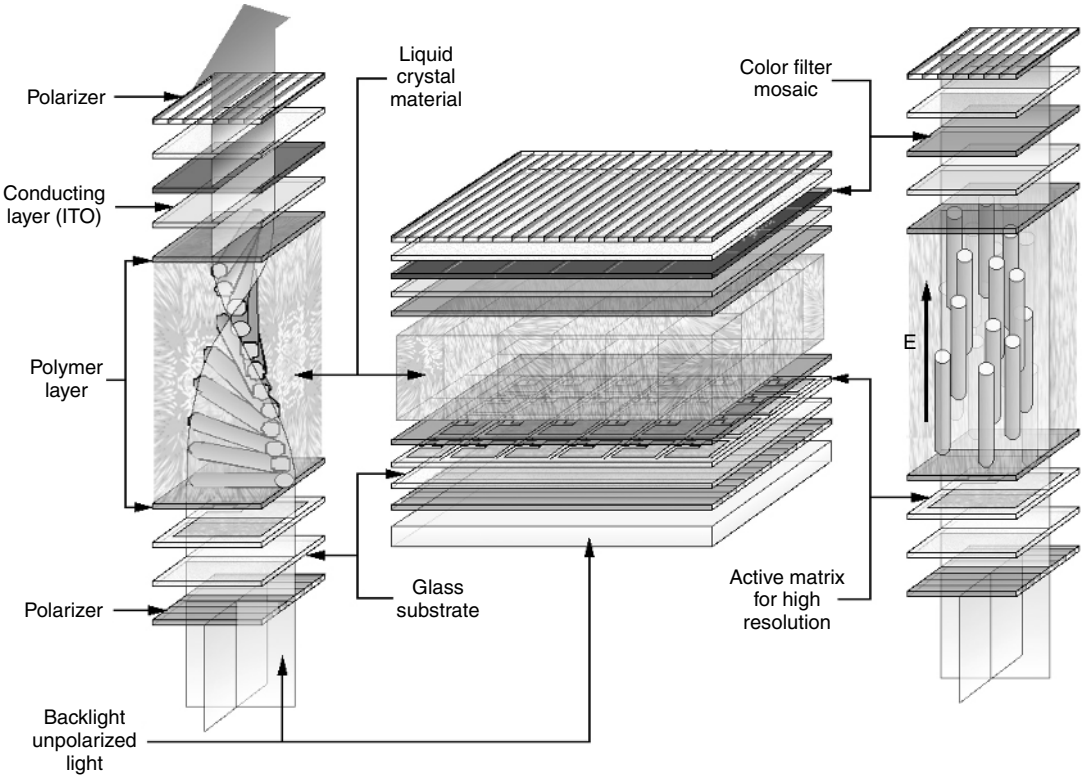


Figure 5. The operation of the twisted nematic (TN) configuration, which is currently the most common LCD mode. The center illustration shows the optical stack of all of the components of the TN display cell. The adjacent insets show the field-on and field-off states of the normally white (NW) configuration. Note that the cross-section is not drawn to scale. See color insert.

has a 90° twist sandwiched between crossed polarizers. A 5-μm cell gap is used for most TN displays. In the figure, the extraordinary axis (long axis) of the LC molecules at the surfaces is parallel to the transmission axes of the polarizers (called the e-mode). However, the o-mode, when the ordinary axis is parallel instead, is also used. These two configurations lead to slightly different transmission and viewing-angle characteristics (35). Finally, placing a red, green, and blue color filter array between the top substrate and ITO layer creates a full color display, known as spatial color synthesis.

The TN principle of operation is as follows. After the first polarizer, linearly polarized light will follow the LC twist and waveguide to an orthogonal linear polarization state. This process, sometimes called adiabatic following, enables light to escape through the second polarizer. This is valid if the twist angle (in this case $\pi/2$) is much smaller than the retardation of the nematic ($2\pi\Delta nd/\lambda$), known as the Mauguin condition. Using common LC values, birefringence $\Delta n = 0.10$, cell thickness $d = 5\ \mu m$, and wavelength $\lambda = 510$ nm, this ratio is $1 : \sim 4$. If this limit is not satisfied, then the light exiting the LC layer is elliptically polarized, and the output from the top polarizer will be reduced. This configuration is called normally white (NW), and describes the zero-voltage state as the transmissive one. This is most common in today's technology. The polarizers can also be arranged in a parallel configuration, so that the zero-voltage state is black and the voltage-on state is the transmissive one.

The transmission of a NW-mode TN-LCD can be derived by using the Jones formalism (35):

$$T = \frac{1}{2} - \frac{1}{2}\frac{\sin^2\left[\frac{\pi}{2}\sqrt{1+u^2}\right]}{1+u^2}. \tag{4}$$

Here, $u = 2d\Delta n/\lambda$, and it is assumed that the director n at the display substrates is parallel to the transmission axis of the polarizers. The 1/2 term indicates that the maximum transmission, as shown in Fig. 6, can be ideally 50% of the input light due to the loss in the first polarizer. Note that the maximum transmission occurs when the argument

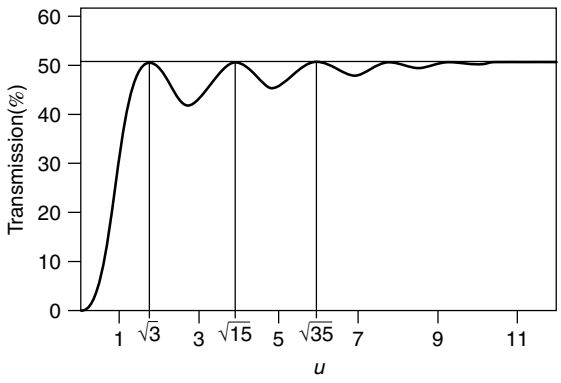


Figure 6. The transmission of the NW-TN cell in the field-off state as a function of the Mauguin parameter.

of $\sin^2[\ldots]$ is an integral multiple of π, corresponding to $u = \sqrt{3}, \sqrt{15}, \sqrt{35}$, etc. The first maximum that occurs at $\sqrt{3}$ corresponds to the maximum brightness and contrast ratio. Optimizing around other maximums is also done (at $\sqrt{15}$, the second maximum, for example), but this is not nearly as common. Nematic domains can sometimes appear in the TN mode, which degrade the contrast but can be overcome by using a polyimide alignment layer which induces an out-of-plane pretilt (typically in the range of $2-10^\circ$).

When a voltage is applied to the pixel, an electric field is created perpendicular to the substrates. Because the dielectric anisotropy of the LC is positive ($\Delta\varepsilon > 0$), the molecules tend to align parallel to the field. Equation (5) can be used to predict the actual threshold voltage of the twisted nematic configuration, which is the point where the molecules just begin to realign. The threshold voltage for a TN display is given by the following expression:

$$V_{\mathrm{TH}} = \pi\sqrt{\frac{K_{11}}{\varepsilon_o \Delta\varepsilon}}\left[1 + \left(\frac{K_{33} - 2K_{22}}{4K_{11}}\right)\right]^{1/2} \tag{5}$$

Using typical values $K_{11} = 10^{-11}$ N, $K_{22} = 5.4 \times 10^{-12}$ N, $K_{33} = 15.9 \times 10^{-12}$ N, and $\Delta\varepsilon = 10.5$, then $V_{\mathrm{TH}} = 1.1$ V, a common threshold for most nematic mixtures. Note that this threshold is the voltage where the LC starts to align due to the applied voltage and does not say anything about the director profile. Above V_{TH}, the broken symmetry of the twist due to the out-of-plane reorientation of the nematic molecules align perpendicular to the substrates and light passes through teh LC layer without any change in polarization. In the NW configuration, the output polarizer then absorbs the light. Grayscale levels are attainable when intermediate voltages are used. In summary, the TN-LCD therefore simply modulates the intensity of a powerful backlight by acting on the polarization state of the incident light.

Notice that the display shown in Fig. 5 is addressed by an active matrix, which is common because the luminance–voltage curve for the twisted nematic LCD is not very steep and is not conducive to multiplexing schemes (36,37); therefore an active matrix is exclusively used to address the twisted nematic for high-end, high-resolution imaging. For example, if we consider Eq. (1) and substitute typical values for a TN material ($\Delta = 0.7$ and $V_{\mathrm{TH}} = 1.8$ V), the maximum number of rows that can be addressed is approximately six; therefore only very low resolution displays of this mode are possible using multiplexing schemes.

The switching time (typically in the range of 10–30 ms) of the TN configuration is proportional to the viscosity and the square of the cell gap and therefore is very sensitive to the cell gap. Although thinner cells enable faster switching, they often compromise the Mauguin condition, which can reduce brightness and the contrast ratio. The art of display design consists, in large part, of balancing these parameters for maximum benefits.

Supertwisted Nematic (STN) Liquid Crystal Displays

While the TN-TFT display configuration has become the standard in applications that require high pixel densities and gray-scale resolution, the supertwisted nematic (STN) configuration (38,39) has been very successful in the low to medium resolution realm. The reasons for this primarily surround the ability of this display mode to be multiplexed (passive-matrix addressing). As discussed earlier, an $N \times M$-pixel display requires $N + M$ electrical connections using this passive scheme (the same number as for active-matrix addressing) but does not require a TFT matrix. And even though there are some optical performance trade-offs involving contrast and switching times, multiplexing simplifies the manufacturing process tremendously and enables the fabrication of inexpensive medium-resolution displays. In this mode, as in the TN, color is achieved through spatial color synthesis.

An expanded view of the basic STN display that has a twist of 270° is shown in Fig. 7. As for the TN, the substrates are coated by a polyimide layer that uses orthogonal rubbing for alignment. Notice that the illustrated twist in the field-off state is greater than the 90° that would be the minimum energy configuration resulting from surface alignment alone; a regular nematic in this configuration would immediately relax to the TN mode. The twist angle in the STN is maintained at >90° using materials with an inherently twisted structure, usually a chiral nematic, which exhibits intrinsic handedness that manifests as a macroscopic twist. This structure is identical to the cholesteric phase described later and can be characterized by the pitch p_0 and the sense (left-or right-handedness) of the twist. When an ordinary nematic is doped by using a chiral nematic, adjusting the concentration ratio can modify the pitch.

The primary electro-optic benefit of the supertwist is an increase in the steepness of the transmission curve (a decrease in Δ). Using typical values for the STN ($\Delta = 0.1$ and $V_{\mathrm{TH}} = 2.0$ V), the maximum number of addressable rows is ~400; therefore, medium resolution is possible. Because of the larger twist angle, this mode does not meet the Mauguin condition and does not exhibit the waveguiding property of the TN. The transmission equation for this mode is much more complex, but an illustration of midlayer tilt can give some insight. As is apparent from Fig. 8, the reorientation of the midlayer, which is responsible for the optical switching in twist mode cells, is much more sudden for higher twist angles and leads to a much sharper voltage threshold. Additionally, the two polarizers in the STN mode are generally not crossed or parallel, nor is the actual twist of the LC necessarily a multiple of 90°. These angles are now included as parameters over which optimization for a particular application must be performed (38).

The trade-offs of this mode include slightly higher drive voltages, increased sensitivity to the pretilt angle, color leakage in both the on- and off-pixel states, and reduced contrast. Although the analysis is somewhat more involved [the twist term in Eq. (3) becomes $K_{22}(\boldsymbol{n} \cdot \nabla \times \boldsymbol{n} + p_0)^2$, and now includes the pitch of the twist p_0), the voltage threshold can be found through elastic theory considerations (for a zero pretilt):

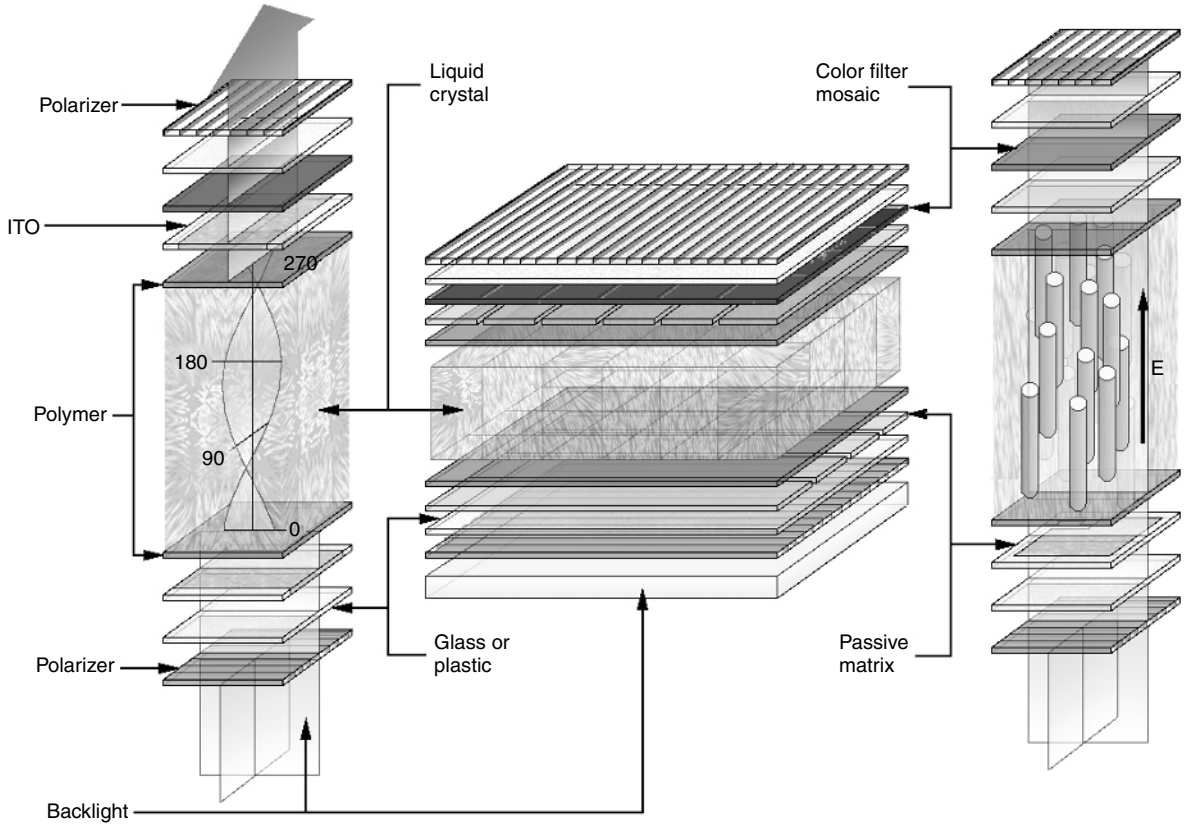


Figure 7. The operation of a supertwisted nematic (STN) configuration. The center illustration shows the optical stack of all of the components of the supertwisted nematic display configuration that is multiplexed. The adjacent insets show the field-on and field-off states of the NW configuration. Note that the cross-section is not drawn to scale. See color insert.

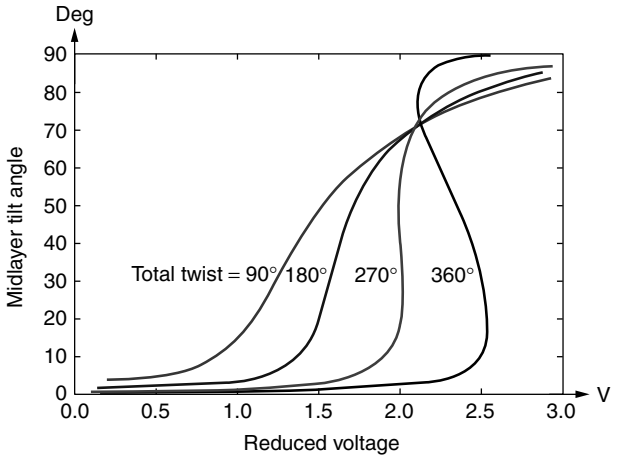


Figure 8. Example of the way midlayer tilt of a LC display changes as a function of twist. The increasingly nonlinear voltage response of twists greater than 90° are the reason that higher degrees of multiplexing are possible in the STN mode. The reduced voltage is defined as the ratio of the drive voltage and the voltage threshold of the 90° twist.

$$V_{\mathrm{TH}} = \pi\sqrt{\frac{K_{11}}{\varepsilon_o \Delta\varepsilon}}\left[1 + \frac{K_{33} - 2K_{22}}{K_{11}}\left(\frac{\phi}{\pi}\right)^2 + 4\frac{K_{22}}{K_{11}}\frac{d}{p_0}\frac{\phi}{\pi}\right]^{1/2} \tag{6}$$

The extra term arises from the presence of the inherent twisting strength and slightly increases the threshold [note that this reduces to Eq. (4) when $p_0 \Rightarrow \infty$ and $\Phi \Rightarrow \pi/2$]. Using typical values $K_{11} = 10^{-11}$N, $K_{22} = 5.4 \times 10^{-12}$N, $K_{33} = 15.9 \times 10^{-12}$N, $\Delta\varepsilon = 10.5$, $\Phi = 270°$, and $d/p_0 = 0.75$, then $V_{\mathrm{TH}} = 2.2$ V, higher than the TN.

A result of the larger twist angles is the appearance of striped textures that can occur within pixels and destroy contrast and color performance. These defects, as for the TN domain defects described before, can be overcome by an appropriate pretilt angle, that can range from 4–8°. Two factors negatively impact contrast in a STN. First, as previously mentioned, the smaller Δ parameter of a multiplexed display leads to a lower selection ratio [Eq. (2)]. Second, the lack of adiabatic waveguiding in this mode leads to a moderate sensitivity to wavelength. The chromatic anisotropy of the transmission of a NW cell leads to a yellowish-green appearance — an ON-state that is not really white. Furthermore, the substantial leakage of long and short wavelengths in the NB cell results in dark pixels that are not actually black. However, a solution to this leakage is to use an additional passive STN layer that has a reversed twist (40).

The switching times of the STN mode are proportional to the viscosity of the LC and to the square of the cell gap and are longer than those of a comparable TN cell. This is due to the typically larger viscosity and additional chiral twist.

In-Plane Switching Liquid Crystal Displays

The in-plane switching (IPS) mode (41,42) is another LCD mode that has been increasingly successful in large area desktop applications due to the inherently wide field of view (>100°). As will be discussed later, both the TN and STN mode are limited in their viewing-angle performance primarily because of out-of-plane reorientation of LC molecules in response to an electric field. In the IPS configuration, this out-of-plane tilt is avoided by using an in-plane electric field, generated by a patterned electrode structure. Spatial color synthesis is also used to generate full color displays.

Figure 9 shows the operation of a normally black (NB) IPS display using crossed polarizers. The interdigitated ITO fingers of the electrode structure lie entirely on the bottom substrate. Without an electric field, the parallel-aligned polyimide-coated substrates lead to uniform alignment of the nematic through the depth of the cell. Because the transmission axis of the input polarizer is aligned parallel to the rub direction, as shown in Fig. 9, no birefringence is apparent, and all of the light is absorbed by the second polarizer. However, when the in-plane electric field is applied 45° to the rub direction, a twist configuration appears that changes the polarization state of the incident light. Except for a small pretilt designed to minimize defects, there is no out-of-plane tilt. Loosely speaking, when the field is strong enough to reorient most of the LC molecules, this layer can be approximately modeled as a birefringent waveplate, and the normalized transmission through the cell in Fig. 9 is the following:

$$T = \frac{1}{2}\sin^2\left(\frac{\pi \Delta n d}{\lambda}\right) \tag{7}$$

At low fields, a simple transmission expression is not available, and analysis must be done numerically. Because there is no splay or bend in this mode, the energy density involves only the twist and electric field terms. A voltage threshold can be derived as

$$V_{\mathrm{TH}} = \frac{\pi l}{d}\sqrt{\frac{K_{22}}{\varepsilon_o |\Delta\varepsilon|}} \tag{8}$$

where l is the distance between the interdigitated electrodes, d is the cell thickness, and $\Delta\varepsilon$ can be either negative or positive. Using typical values $K_{22} = 5.0 \times 10^{-12}$N, $\Delta\varepsilon = 10$, and $l/d = 4$, then $V_{\mathrm{TH}} = 3$ V, slightly higher than both the TN and the STN.

Unlike the TN configuration, the electro-optic response in this mode is very sensitive to the cell gap and the electrode spacing. And although this mode has a voltage threshold, an active matrix must be used for addressing due to the large Δ parameter.

Although the wide viewing angle is the acclaimed feature of the IPS mode, several issues are currently

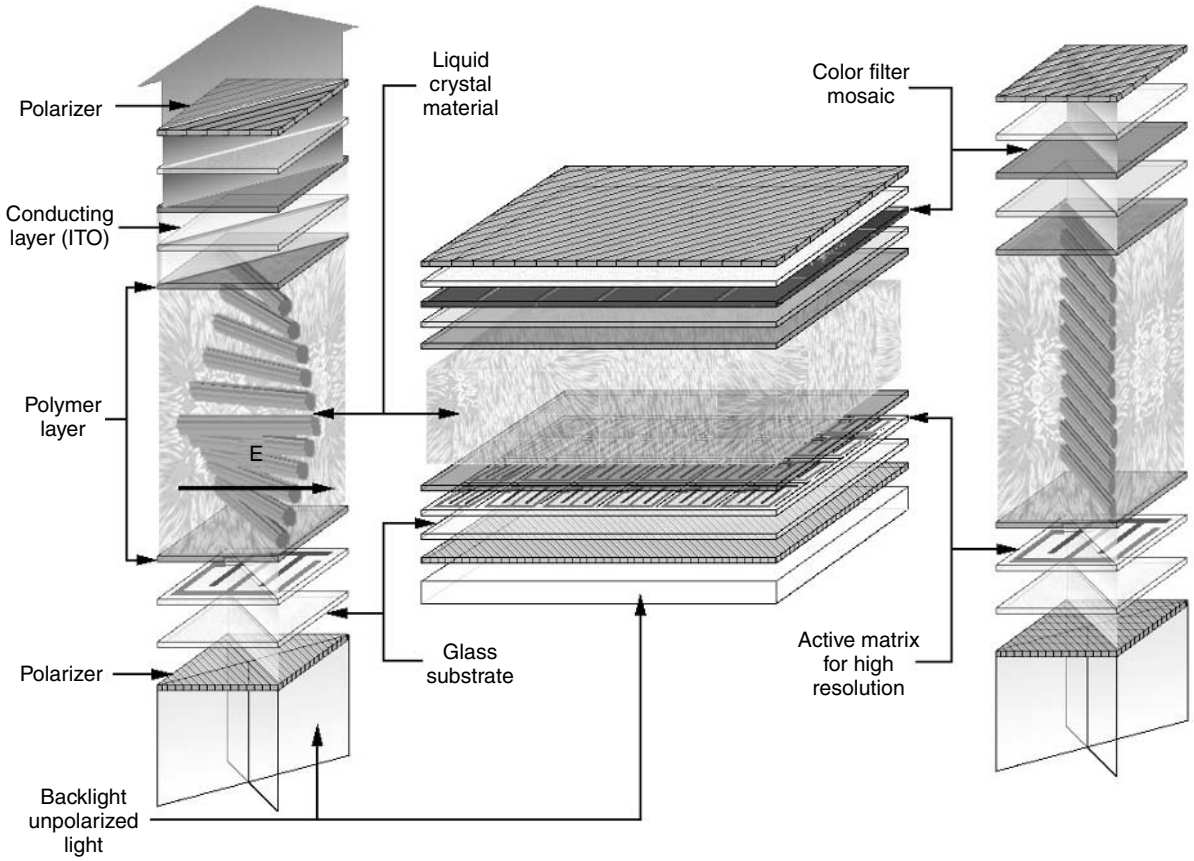


Figure 9. The operation of a normally black (NB) in-plane switching (IPS) configuration. The IPS mode is very different from the TN and STN technology in that the electric field is in the plane of the substrates. Note that the cross-section is not drawn to scale. See color insert.

being investigated. First, and most critical, the switching time continues to be somewhat longer than for the TN, making it difficult for this display to support true video refresh rates (43). Second, the pixel architecture generally leads to a comparably smaller pixel aperture, and the LC configuration leads to an ideal transmission less than 50%. Both of these contribute to decreased overall transmission and can be overcome by a brighter backlight, but this leads to greater power requirements; however, this is not much of a problem for the desktop market. Third, the drive electronics for this mode typically demand more power, adding further to power consumption. Finally, the transmission of the IPS-LCD cell does not respond uniformly to all wavelengths, as can be seen from Eq. (9). In spite of these challenges, the IPS mode continues being used in large area desktop display.

Ferroelectric Liquid Crystals

Ferroelectric liquid crystals (FLCs) are an important addition to the previous discussion of nematic LCDs because extremely fast switching and bistable displays are possible. Our treatment here is introductory, so we suggest further reading on FLCs and chiral liquid crystals in these thorough sources (44,45). Most broadly, both nematic and ferroelectric molecules exhibit shape anisotropy. However, FLCs are distinguished by inherent chirality and the presence of a permanent dipole oriented perpendicularly to the long axis of the molecule. On the macroscopic level, this transverse dipole leads to the smectic C* phase of the bulk LC, as seen in Fig. 10a. The LC molecules become organized into layers within which no 2-D positional order exists, while orientational order is maintained at some characteristic angles (θ, ϕ). The polar tilt angle θ is consistent throughout each layer and can be as large as $\sim 45^\circ$ and as small as a few degrees (46,47). Furthermore, though the azimuthal angle ϕ is approximately uniform throughout each layer, the symmetry axis from one layer to the next rotates about the normal vector by a small $d\phi$. The net polarization of the helix is zero due to averaging of the in-plane polarization of each smectic layer. However, when the helix is unwound

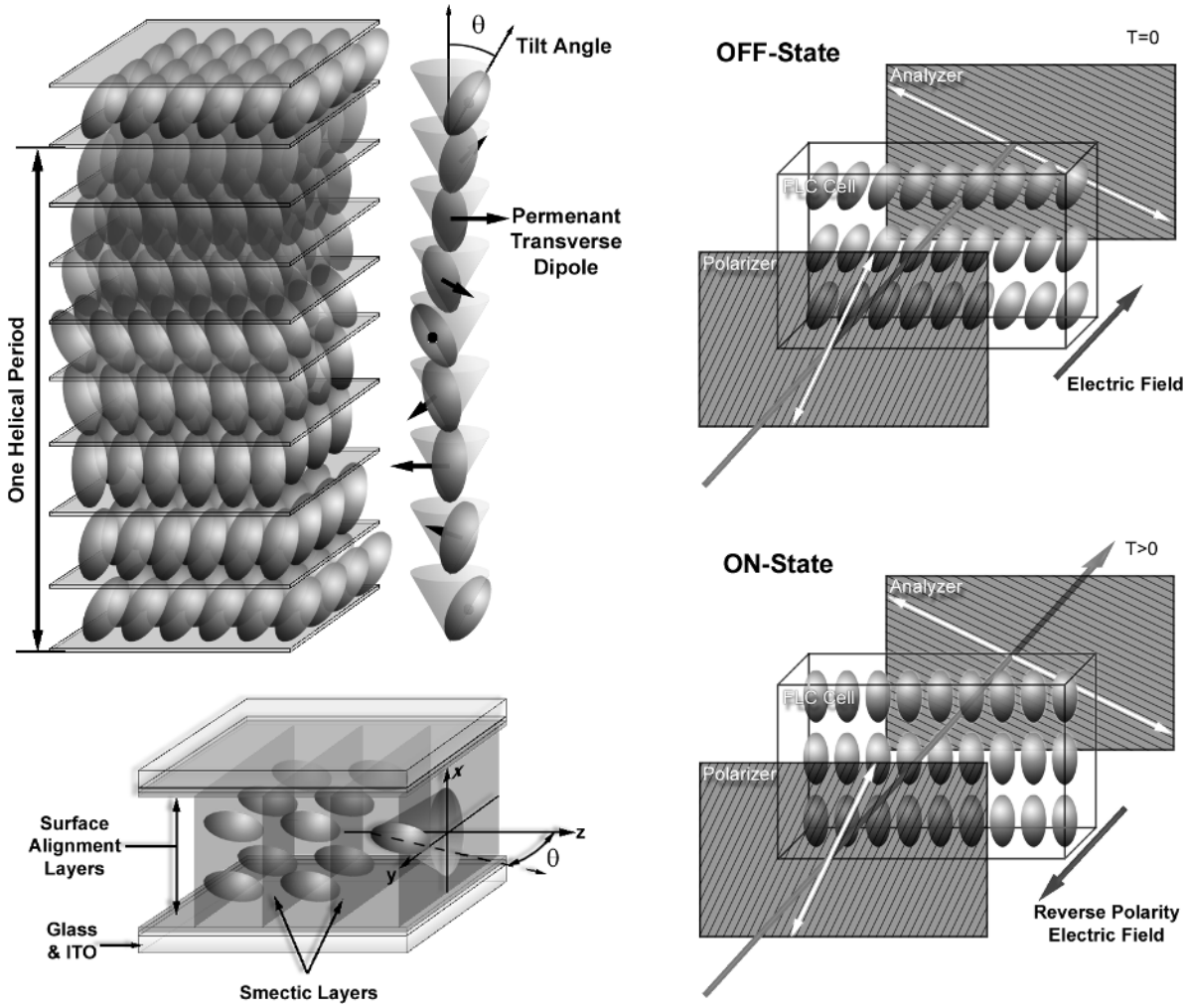


Figure 10. Ferroelectric liquid crystals (FLC): (**a**) the structure of the smectic C* phase; (**b**) the surface-stabilized (SSFLC) mode, where the helix is unwound and FLC molecules align to form a uniaxial birefringent layer; (**c**) bistable switching states in the SSFLC controlled by the polarity of the electric field. See color insert.

to form the ferroelectric phase (by the influence of fields or surfaces), a net spontaneous polarization exists, typically in the range of $P_S \sim 1-200$ nC/cm^2. It is important to note that both the spontaneous polarization and the tilt angle are highly temperature dependent, usually $\propto \sqrt{T_C - T}$, where T_C is the Curie temperature (the transition into the higher temperature phase, usually smectic C* $\rightarrow$ smectic A) (48).

The unique ordering and electro-optic response of FLCs can be used to produce a fast switching and bistable display through the Clark–Lagerwall effect (49). This effect is seen when a FLC is confined between substrates whose surface alignment is parallel and have a cell gap much less than the helical pitch, as shown in Fig. 10b. Known as surface-stabilized FLCs (SSFLCs) (50), the helix is completely unwound by the substrate anchoring influence, the FLC symmetry axis lies along the substrate alignment direction, and the smectic layers are perpendicular to the substrates (called the quasi-bookshelf texture). When a voltage is applied, the permanent dipoles uniformly align parallel to the electric field because rotation of the FLC is constrained to the azimuthal cone. In this configuration, the mesogenic cores (the LC body) align parallel to the plane of the substrates and form a uniform birefringent medium. However, when the polarity of the electric field is reversed, the FLCs rotate to the opposite side of the cone. The molecular ordering will persist "indefinitely" in both of these cases, even when the electric field is removed. This bistable memory is a much sought after property because it can result in low-power operation. However, this bistability is particularly sensitive to temperature and physical shock (such as deformations of the substrates), and defects are a substantial device issue.

A display can be formed when this cell is placed between crossed polarizers where one axis of symmetry is parallel to a polarizing axis. Both the ON- and OFF-states are illustrated in Fig. 10c. Linearly polarized light leaving the first polarizer encounters no birefringence in the FLC layer and remains linearly polarized; the second polarizer then absorbs all of the light. However when a voltage of opposite polarity is applied, the FLCs rotate by 2θ. In this case, the birefringence of the FLC layer leads to a phase shift in the linearly polarized light that enters the layer, and the transmission of the LCD can be modeled using a waveplate model (51), where θ is the FLC tilt angle and T_0 is the unpolarized intensity entering the first polarizer:

$$T = T_0 \frac{1}{2} \sin^2(4\theta) \sin^2\left(\frac{\pi \Delta n d}{\lambda}\right) \tag{9}$$

Maximum contrast occurs when $\theta = 22.5^\circ$ and $2d\Delta n/\lambda = 1$, and usually requires very small cell gaps ($d \sim 2$ μm) for visible wavelengths. As a result of the FLC reorientation only along the azimuthal cone, switching times are substantially less than those in the TN configuration; the azimuthal viscosity of a FLC is usually substantially less than the viscosity of the rotational viscosity of a nematic. And contrary to most nematic LCD modes, the electric field directly influences both the rise and fall times in the FLC which are inversely proportional to the spontaneous polarization and the electric field. Both of these factors lead to switching times in the range of $\sim 10-200$ μs.

The voltage threshold of the SSFLC mode involves both the spontaneous polarization and the nonpolar anchoring energy W_d (usually in the strong-anchoring regime):

$$V_{TH} \simeq 4 \frac{W_d}{P_S} \tag{10}$$

For example, $W_d = 10^{-4}$ J/m^2 and $P_S = 20$ nC/cm^2 leads to $V_{TH} \sim 2$ V, well within the requirements for display systems. The limitations of these devices include the onset of defects due to temperature or shock, the difficulty of producing gray-scale images, and the challenge of maintaining small cell gaps in displays of any substantial size. Nonetheless, FLCs can be found in many applications including field-sequential color displays (52) and small flexible displays (53).

A second electro-optic effect used as a display mode is known as the deformed helix ferroelectric (DHF) (54). In this case, the helical pitch is much smaller than the cell gap and shorter than the wavelength of visible light. As a result, the helical axis lies along the plane of the substrates perpendicularly to the electrodes. An externally applied field distorts the helix and results in the rotation of the optical axis (the average direction of the molecules) away from the helix. Although this mode is not bistable and requires an active matrix, it does enable gray-scale imaging with very fast switching times (55).

Reflective Displays from Cholesteric Liquid Crystals

There is a color-reflective bistable LC-based display that is on the verge of significant commercialization. The properties of a cholesteric LC material allow it to form two stable textures that persist, even when the drive electronics are inactive: a reflective planar texture that has a helical twist whose pitch p can be tuned to reject a portion of visible light or focal conic textures (scrambled helices) that are relatively transparent. Figure 11 shows the basic structure of a cholesteric display that is backed by a black substrate. In the planar texture case (Fig. 11a), the periodicity of the helices enables them to Bragg reflect a narrow range of colors, whereas all of the others pass through and are absorbed by a black background. The viewer sees a brilliant color reflection whose bandwidth is in the perfect planar texture of ~ 100 nm, governed by $\Delta\lambda = p\Delta n$. Ideally this reflection peak can only be at 50% efficiency because cholesteric displays reflect either the right-handed component or left-handed component of circular polarized light, depending on the intrinsic twist of the material itself (56).

Upon the application of an applied voltage ($\sim 10-15$ V), the planar structure transforms into the focal conic that is nearly transparent to all wavelengths in the visible, as shown in Fig. 11b. The viewer sees the black background, thereby creating an optical contrast between reflecting color pixels and black pixels. In this state, the voltage can be removed, and the focal conic state will remain indefinitely, so there is bistable memory between the reflecting planar state and the transparent focal conic state. To revert to the planar reflecting texture from the focal conic state, the pixel must go through the highly aligned state (also known as the homeotropic

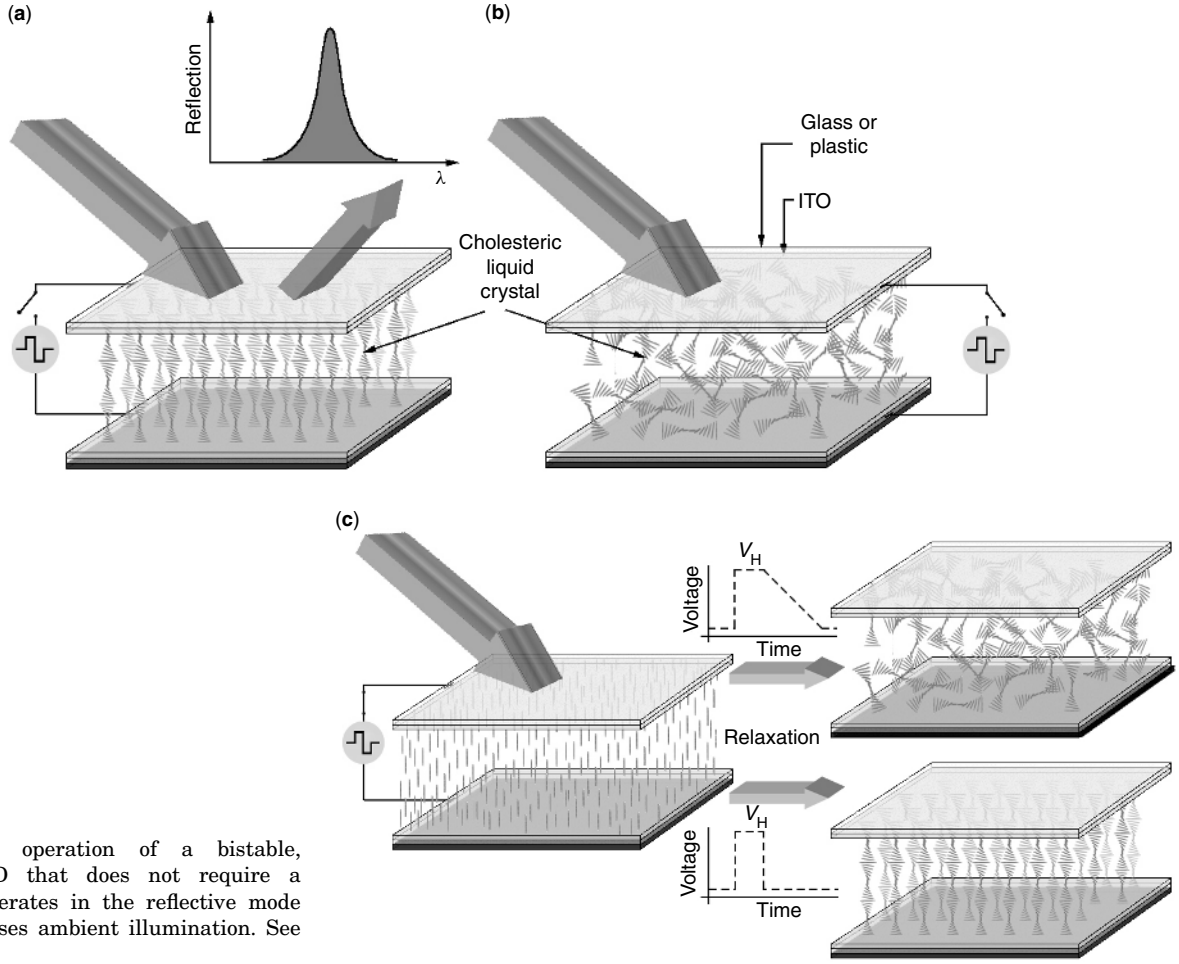


Figure 11. The operation of a bistable, cholesteric LCD that does not require a backlight; it operates in the reflective mode and therefore uses ambient illumination. See color insert.

state), as shown in Fig. 11c. The transformation requires 30–35 V. An abrupt turn-off of the voltage after the aligned state results in the planar texture. The Bragg-type reflection of cholesterics is far from Lambertian-like but has a more specular (mirror-like) nature; to smear out the reflection across a broader viewing angle, unique alignment techniques that slightly "fracture" the planar texture are employed to spread out the Bragg reflection across a broader viewing angle at the expense of on-axis reflection. Grayscale is achieved in cholesteric technology by controlling the focal conic domains using different levels of voltage.

Because these devices are transparent, vertical integration is possible, as shown in Fig. 12, to create a true color addition scheme. Although stacking can create a complicated drive circuitry, it does preserve resolution and brightness because the pixels are vertically integrated rather than spatially arranged across the substrate plane. The photopic white reflectance of the vertical stack is >40%. The dynamic response times of cholesteric materials are of the order of 30–40 ms. By implementing unique addressing schemes, video is possible using cholesteric LC technology. The most attractive feature of the cholesteric display is that its reflectance–voltage curve has a well-defined threshold, which enables the use of inexpensive passive-addressing schemes, even for high resolution (57). A full color VGA image from this reflective display

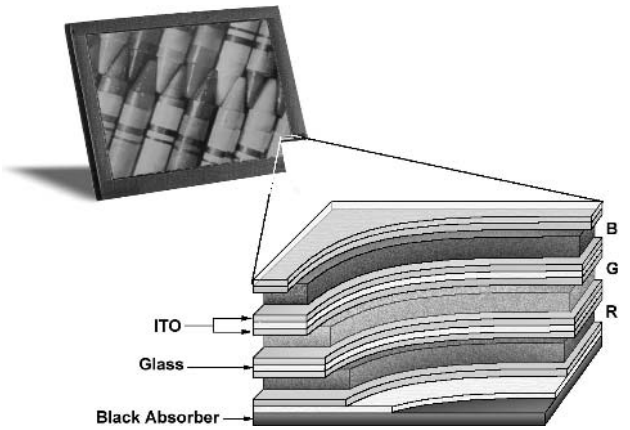


Figure 12. For cholesteric reflective displays, color is created by stacking red, green, and blue panels, thereby preserving brightness and resolution. This is in contrast to the spatial patterning of red, green, and blue color filter arrays used in TN, STN, and IPS technology (Photograph courtesy of J. W. Doane of Kent Displays). See color insert.

configuration is shown in Fig. 12. Cholesteric LC materials are being developed for document viewers, electronic newspapers and books, and information signs — portable applications where bistability is extremely beneficial.

DISPLAY PERFORMANCE PARAMETERS

Operation of important display technologies has been discussed before, so we will now briefly summarize three important parameters that are commonly used for display characterization.

Contrast Ratio

The optical contrast ratio (CR) is a simple performance measure that captures how clearly an image can be seen (58,59). It is defined as the ratio of the luminance of the ON-state and the luminance of the OFF-state: $CR = L_{ON}/L_{OFF}$. A high CR demands high transmission in the bright state and is particularly sensitive to the dark-state brightness, which means that very high ratios are possible, even from a dim display. Unfortunately, the unbounded nature of this equation often makes display comparisons difficult, and consequently, it is not uncommon to find completely different image quality on two competing displays that have the same CR.

The contrast of the TN and IPS modes is superior to the STN due to their addressing scheme and the respective nematic configuration. The multiplexing technique used by the latter leads to a slight increase in the dark-state luminance. The color leakage inherent in the STN mode further contributes to the contrast ($<15:1$). Both TN and IPS displays typically exhibit similar contrast ($>100:1$) due to good dark states when viewed on-axis, but the TN will often appear brighter.

Viewing Angle

The field of view is one of the most critical performance parameters of a high information content LCD panel, primarily due to the size of modern flat-panel displays (typically $>13''$). A viewing cone of at least 30° is needed just to view the extreme corners if the observer is seated 20″ away. Most LCD modes maintain excellent on-axis characteristics but exhibit poor contrast and grayscale inversions when viewed off-axis. This can be seen clearly in the isotransmittance curves for a NW-TN cell shown in Fig. 13. These problems arise primarily from a combination of the off-axis leakage of the polarizers and the out-of-plane tilt of the nematic molecules in the center of the cell (60). This results in an angularly dependent birefringence and in elliptical, rather than linear, off-axis polarization states incident on the second polarizer. A tremendous amount of effort has gone into improving this for the TN-TFT-LCD, and a variety of approaches have been successful: compensation films (20,35,38,61), multidomain approaches (62,63), and novel nematic configurations [vertical alignment (64), optically compensated bend (67)]. As previously mentioned, the LC molecules in the IPS mode remain substantially in the plane, and only the leakage from the polarizers is common.

Color

Unlike the previous two display measures, the color performance of a display necessarily involves the physiology of the eye. Because of this, two realms exist for characterizing displays: radiometric and photometric. Radiometry involves measuring actual optical power in terms of radiant flux (watts), combined with units of area, distance, solid angle, and time. However, because the human visual system does not have a flat spectral response, photometry was developed. In this case, radiometric data undergo a linear transformation to take into account the average photo-optic response of the eye and are expressed in terms of luminous flux (lumens). Depending

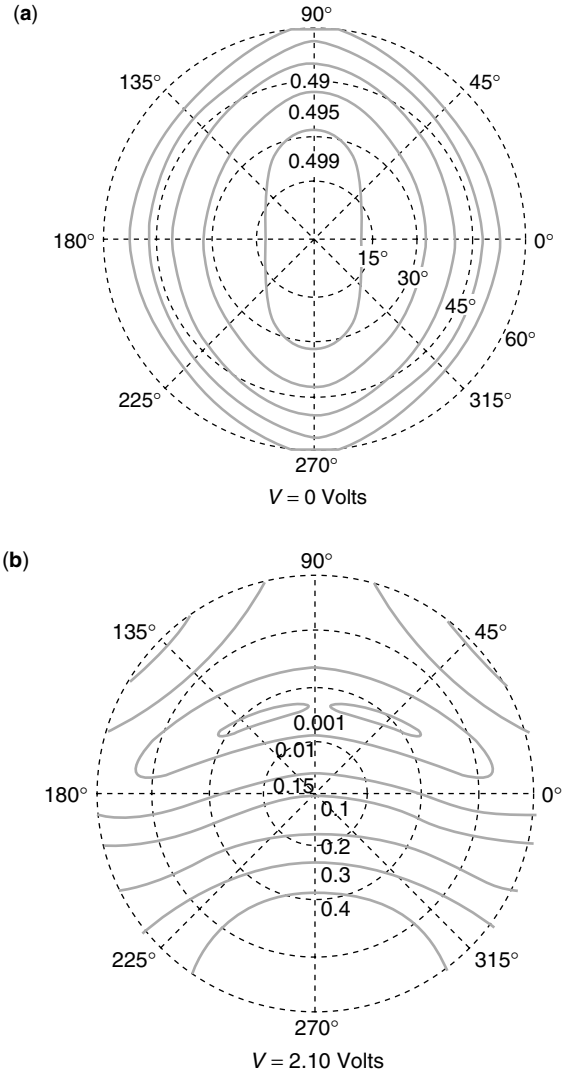


Figure 13. Examples of isotransmittance curves of an uncompensated NW-TN cell for the (**a**) bright state and (**b**) 10% gray-level state. These plots show constant-transmittance contours on a polar plot that has a 360° azimuthal span and 60° polar range measured from the vector normal to the display cell. Notice only a small anisotropy of the contours in the bright state which indicates that this display would appear just as bright from all polar angles as large as 60°. However, when the gray-scale voltage is applied, the variation in transmission over the azimuthal angles and the gray-scale inversions at 90° are apparent and would result in a very poor image. Also note that at best only 50% of the unpolarized backlight can be transmitted.

on the context, units of radiance (watts/steridian $\times$ m^2) and luminance (candelas/m^2, lumens/steridian $\times$ m^2, nits, and footlamberts) are commonly used for radiometric and photopic measures, respectively. One particularly useful aspect of photometry is the development of a color space wherein any optical power spectrum can be mapped onto a two-parameter chromaticity space that accounts for the photopic response of the eye. We refer the reader to the extensive treatment of display color performance in (58,63).

The properties of the human visual system allow numerous possibilities of color synthesis with LCDs. Full-color commercial displays almost always use a spatial synthesis approach, in which a broad color gamut can be perceived using three spatially distinct, independently addressed subpixels of differing primary colors (usually red, green, and blue). Because these subpixels usually use absorbing color filters, one consequence of this scheme is a reduction in the ideal brightness of a display by at least two-thirds in most situations. In addition to this color-addition method, a pixel can also be formed using three primary-color layers that are vertically stacked. This enables maximum ideal throughput of the backlight and highest pixel resolution, but drawbacks include the necessity of three LC layers, their associated addressing matrices, and additional substrates, all of which dramatically increase manufacturing complexity. Nonetheless, this approach is ideal for reflective displays to maximize efficiency. In transmissive displays, a color-subtraction scheme using dichroic dyes can also be devised using the vertically stacked pixel. In another possibility, a sequence of primary colors flashed quickly will be integrated by the eye and will be perceived as a different color. This is known as field sequential color synthesis (64,65).

SUMMARY

The development of LCDs is continuing at a feverish pace because of the market need for high-fidelity displays in numerous imaging applications. Perhaps many innovative and profitable careers are supported by the modern maxim: "You can image all content on some LCDs and some content on all LCDs, but you cannot image all content on all LCDs"! Since imaging applications are ever-changing and display technology is continually evolving, we have chosen in this article to target the fundamental aspects of the technology in order to provide the reader with a flexible, yet foundational, understanding.

ABBREVIATIONS AND ACRONYMS

α-Si	amorphous silicon
CR	contrast ratio
DHF	deformed helix ferroelectric
FLC	ferroelectric liquid crystal
ITO	indium tin oxide
IPS	in-plane switching
LC	liquid crystal
LCD	liquid crystal display
NB	normally black
NW	normally white
PECVD	plasma-enhanced chemical vapor deposition
Poly-Si	polycrystalline silicon
REI	reactive ion etching
STN	supertwisted nematic
SSFLC	surface-stabilized ferroelectric liquid crystal
TFT	thin-film-transistor
TN	twisted nematic

BIBLIOGRAPHY

1. G. P. Crawford, *IEEE Spectrum* **37**, 40–46 (1992).
2. B. Comiskey, J. D. Albert, H. Yoshizawa, and J. Jacobson, *Nature* **394**, 253–255 (1997).
3. N. K. Sheridon, *J. of Soc. for Information Display* **7**, 141–144 (1999).
4. P. M. Alt and P. Pleshko, *IEEE Trans. Elec. Dev.* **ED-21**, 146–155 (1974).
5. T. Scheffer, in P. Collings and J. Patel, eds., *Handbook of Liquid Crystal Research*, Oxford University Press, New York, 1997, pp. 445–471.
6. S. Kobayashi, H. Hiro, and Y. Tanaka, in P. Collings and J. Patel, eds., Oxford University Press, New York, 1997, pp. 415–444.
7. T. P. Brody, J. A. Asars, and G. D. Dixon, *IEEE Trans. Elec. Dev.* **ED-20**, 995–1101 (1973).
8. P. G. Le Comber, *Electron. Lett.* **15**, 179–181 (1979).
9. V. G. Chigrinov, *Liquid Crystal Devices*, Artech House, NY, 1999, pp. 238–246.
10. D. E. Mentley and J. A. Castellano, *Liquid Crystal Display Manufacturing*, Stanford Resources, Inc., San Jose, 1994.
11. C. W. Kim et al., *SID Digest* **31**, 1,006–1,009 (2000).
12. H. Kinoshita et al., *SID Digest* **30**, 736–739 (1999).
13. S. Nakabu et al., *SID Digest* **30**, 732–735 (1999).
14. T. Sameshima, M. Hara, and S. Usui, *Jpn. J. Appl. Phys.* **28**, 2,131–2,133 (1989).
15. A. T. Voutsas, D. Zahorski, and S. Janicot, *SID Digest* **30**, 290–293 (1999).
16. S. Utsunomiya, S. Inoue, and T. Shimoda, *SID Digest* **31**, 916–919 (2000).
17. R. A. Pelcovits, in P. Collings and J. Patel, eds., *Handbook of Liquid Crystal Research*, Oxford University Press, New York, 1997, pp. 71–95.
18. L. M. Blinoff and V. G. Chigrinov, *Electrooptic Effects in Liquid Crystal Materials*, Springer, New York, 1996.
19. P. G. de Gennes and J. Prost, *The Physics of Liquid Crystals*, Oxford Science Publications, Oxford Press, New York, 1993.
20. S. Chandraskehar, *Liquid Crystals*, Cambridge University Press, Cambridge, 1994.
21. I. C. Khoo and F. Simoni, *Physics of Liquid Crystalline Materials*, Gordon and Breach Science Publishers, Philadelphia, 1991.
22. S. A. Pikin, *Structural Transformations in Liquid Crystals*, Gordon and Breach Science Publishers, Philadelphia, 1991.
23. H. Stegemeyer, *Liquid Crystals*, Springer, New York, 1994.
24. A. A. Sonin, *The Surface Physics of Liquid Crystals*, Gordon and Breach Science Publishers, Philadelphia, 1995.
25. D. Demus et al., *Physical Properties of Liquid Crystals*, Wiley-VCH, Weinheim, 1999.

26. P. J. Collings and M. Hird, *Introduction to Liquid Crystals*, Taylor and Francis, London, 1997.
27. S. Elston and R. Sambles, *The Optics of Thermotropic Liquid Crystals*, Taylor and Francis, London, 1998.
28. R. C. Jones, *J. Opt. Soc. A* **31**, 488–499 (1941).
29. S. Teitler and J. Henvis, *J. Opt. Soc. Am.* **60**, 830–840 (1970).
30. D. W. Berreman, *Phil. Trans. R. Soc.* **309**, 203–216 (1983).
31. M. Schadt, *Annu. Rev. Mater. Sci.* **27**, 305–379 (1997).
32. P. Yeh and C. Gu, *Optics of Liquid Crystal Displays*, Wiley Interscience, New York, 1999.
33. J. A. Castellano, *Handbook of Display Technology*, Academic Press Inc., San Diego, 1992.
34. D. E. Mentley and J. A. Castellano, *Flat Information Displays Market and Technology Trends*, Stanford Resources Inc., San Jose, 1993,1994.
35. T. Scheffer and J. Nehring, in B. Bahdur, ed., *Liquid Crystals: Applications and Uses*, World Scientific, Singapore, 1993.
36. T. J. Scheffer and J. Nehring, *J. Appl. Phys.* **58**, 3,022–3,031 (1985).
37. I. C. Khoo and S. T. Wu, *Optics and Non-linear Optics of Liquid Crystals*, World Scientific, Singapore, 1993.
38. M. Oh-e, M. Ohta, S. Aratani, and K. Kondo, *Digest Asia Display '95* 577–580 (1995).
39. H. Wakemoto et al., *SID Digest* **28**, 929–932 (1997).
40. M. Hasegawa, *SID Digest* **28**, 699–702 (1997).
41. I. Musevic, R. Blinc, and B. Zeks, *The Physics of Ferroelectric and Antiferroelectrc Liquid Crystals*, World Scientific, Singapore, 2000.
42. H. -S. Kitzerow and C. Bahr, eds., *Chirality in Liquid Crystals*, Springer, New York, 2001.
43. J. S. Patel and J. W. Goodby, *J. Appl. Phys.* **59**, 2,355–2,360 (1986).
44. T. Geelhaar, *Ferroelectrics* **84**, 167–181 (1988).
45. A. W. Hall, J. Hollingshurst, and J. W. Goodby, in P. Collings and J. Patel, eds., Handbook of Liquid Crystal Research, Oxford University Press, New York, 1997, pp. 41–70.
46. N. A. Clark and S. T. Lagerwall, *Appl. Phys. Lett.* **36**, 899–901 (1980).
47. N. A. Clark and S. T. Lagerwall, *Ferroelectrics* **59**, 25–67 (1984).
48. J. Z. Xue, M. A. Handschy, and N. A. Clark, *Liq. Cryst.* **2**, 707–716 (1987).
49. T. Yoshihara, T. Makino, and H. Inoue, *SID Digest* **31**, 1,176–1,179 (2000).
50. M. Muecke et al., *SID Digest* **31**, 1,126–1,129 (2000).
51. A. G. H. Verhulst, G. Cnossen, J. Funfschilling, and M. Schadt, *International Display Research Conference* **94**, 377–380 (1994).
52. J. Funfschilling and M. Schadt, *J. Appl. Phys.* **66**, 3,877–3,882 (1989).
53. D. K. Yang, L. C. Chien, and Y. K. Fung, in G. P. Crawford and S. Zumer, eds., *Liquid Crystals in Complex Geometrics*, Taylor Francis, London, 1996, pp. 103–142.
54. X. Y. Huang, N. Miller, and J. W. Doane, *SID Digest* **28**, 899–902 (1997).
55. P. A. Keller, *Electronic Display Measurement*, John Wiley & Sons, Inc., New York, NY, 1997.
56. L. W. MacDonald and A. C. Lowe, eds., *Display Systems Design and Applications*, John Wiley & Sons, Inc., New York, NY, 1997, chap. 14–17.
57. P. J. Bos and K. Werner, *Information Display* **13**, 26–30 (1997).
58. H. Mori et al., *Jpn. J. Appl. Phys.* **36**, 143–147 (1997).
59. M. S. Nam et al., *SID Digest* **28**, 933–936 (1997).
60. J. Chen et al., *Appl. Phys. Lett.* **67**, 1,990–1,992 (1995).
61. K. Ohmuro, S. Kataoka, T. Sasaki, and Y. Koike, *SID Digest* **28**, 845–48 (1997).
62. T. Uchida and T. Miyashita, *IDW'95 Digest* 39–42 (1995).
63. R. G. Kuehni, *Color: An Introduction to Practice and Principles*, John Wiley & Sons, Inc., New York, NY, 1997.
64. T. Yoshihara, T. Makino, and H. Inoue, *SID Digest* **31**, 1,176–1,179 (2000).
65. T. R. H. Wheeler and M. G. Clark, in H. Widdel and D. L. Post, eds., *Color in Electronic Displays*, Plenum Press, NY, 1992, pp. 221–281.

M

MAGNETIC FIELD IMAGING

K. N. HENRICHSEN
CERN
Geneva, Switzerland

INTRODUCTION

Magnetic field mapping is the production of maps or images of magnetic fields in space. Magnetic field maps are needed for designing and optimizing of magnets used in particle accelerators, spectrometers (mass, nuclear magnetic resonance, and electron paramagnetic resonance), and magnetic resonance imaging systems. Magnetic field maps are also used in geologic exploration where the variations in the magnitude and direction of the earth's magnetic field are indicative of subsurface features and objects. Field mapping relies on various methods of measuring the magnetic field, generally one point at a time. These measurement methods are the main focus of this article.

It is curious to note that most measurement methods have remained virtually unchanged for a very long period, but the equipment has been subject to continual development. In the following, only the more commonly used methods will be discussed. These methods are complementary and a wide variety of the equipment is readily available from industry. For the many other existing measurement methods, a more complete discussion can be found in two classical bibliographical reviews (1,2). An interesting description of early measurement methods can be found in (3). Much of the following material was presented at the CERN Accelerator School on Measurement and Alignment of Accelerator and Detector Magnets (4). Those proceedings contain a recent compendium of articles in this field and form a complement to the classical reviews.

MEASUREMENT METHODS

Before computers became common tools, electromagnets were designed by using analytical calculations or by measuring representative voltage maps in electrolytic tanks and resistive sheets. Magnetic measurements on the final magnets and even on intermediate magnet models were imperative at that time.

Nowadays, it has become possible to calculate the strength and quality of magnetic fields with impressive accuracy. However, the best and most direct way to verify that the expected field quality has been reached is magnetic measurements on the finished magnet. It is also the most efficient way of verifying the quality of series produced electromagnets to monitor tooling wear during production.

Choice of Measurement Method

The choice of measurement method depends on several factors. The field strength, homogeneity, variation in time, and the required accuracy all need to be considered. The number of magnets to be measured can also determine the method and equipment to be deployed. As a guide, Fig. 1 shows the accuracy that can be obtained in an absolute measurement as a function of the field level, using commercially available equipment. An order of magnitude may be gained by improving the methods in the laboratory.

Magnetic Resonance Techniques

The nuclear magnetic resonance technique is considered the primary standard for calibration. It is frequently used for calibration purposes and also for high precision field mapping. The method was first used in 1938 (5,6) for measuring the nuclear magnetic moment in molecular beams. A few years later two independent research teams observed the phenomenon in solids (7–9). Since then, the method has become the most important way of measuring

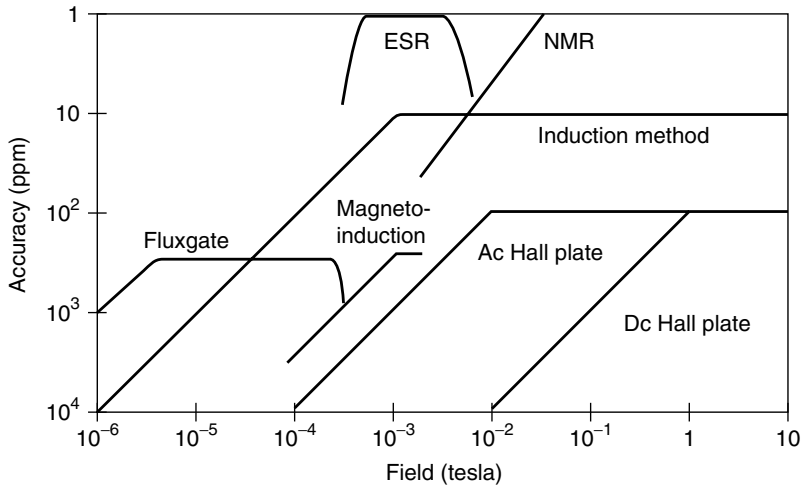


Figure 1. Measurement methods: Accuracies and ranges.

magnetic fields with very high precision. Because it is based on an easy and precise frequency measurement, it is independent of temperature variations. Commercially available instruments measure fields in the range from 0.011 T up to 13 T at an accuracy better than 10 ppm. Commercial units are also available for measuring weaker magnetic fields, such as the earth's magnetic field (30 to 70 μT), but at lower accuracy.

In practice, a sample of water is placed inside an excitation coil, powered from a radio-frequency oscillator. The precession frequency of the nuclei in the sample is measured either as nuclear induction (coupling into a detecting coil) or as resonance absorption (10). The measured frequency is directly proportional to the strength of the magnetic field whose coefficients are 42.57640 MHz/T for protons and 6.53569 MHz/T for deuterons. The magnetic field is modulated with a low-frequency signal to determine the resonance frequency (11).

The advantages of the method are its very high accuracy, its linearity, and the static operation of the system. The main disadvantage is the need for a rather homogeneous field to obtain a sufficiently coherent signal. A small compensation coil that is formed on a flexible printed circuit board and provides a field gradient may be placed around the probe when used in a slightly inhomogeneous field. A correction of the order of 0.2 T/m may be obtained (11). The limited sensitivity and dynamic range also set limits to this method's suitability. It is, however possible to use several probes with multiplexing equipment, if a measurement range of more than half a decade is needed.

Pulsed NMR measurements have been practiced for various purposes (12,13), even at cryogenic temperatures (14). But equipment for this type of measurement is not yet commercially available.

Finally, it should be mentioned that a rather exotic method of NMR measurement using water flowing in a small tube has given remarkably good results in low fields (15–17). It fills the gap in the measurement range up to 11 mT, for which NMR equipment is not yet commercially available. In addition, it provides a method of measurement in strong ionizing radiation such as in particle accelerators. It was tested for measurements in the bending magnets installed in the CERN Large Electron Positron collider (LEP). A resolution of 0.0001 mT was reached in the range from the remanent field of 0.5 mT up to the maximum field of 112.5 mT, and corresponding reproducibility was observed (18). The remarkable sensitivity and resolution of this measurement method makes it suitable for absolute measurements in low fields. In fact, it was even possible to detect the earth's magnetic field outside the magnet, corresponding to an excitation frequency of about 2 kHz. However, the operation of this type of equipment is rather complicated due to the relatively long time delays in the measurement process.

Electron spin resonance (ESR) (19–22) is a related and very precise method for measuring weak fields. It is now commercially available in the range from 0.55–3.2 mT, has a reproducibility of 1 ppm, and is a promising tool in geology applications.

Magnetic resonance imaging (MRI) has been proposed for accelerator magnet measurements (23). It is a very promising technique, which has proven its quality in other applications. However, the related signal processing requires powerful computing facilities, which were not so readily available in the past.

The Fluxmeter Method

This method is based on the induction law. The change of flux in a measurement coil will induce a voltage across the coil terminals. It is the oldest of the currently used methods for magnetic measurements, but it can be very precise (24). It was used by Wilhelm Weber in the middle of the last century (25) when he studied the variations in the strength and direction of the earth's magnetic field. Nowadays, it has become the most important measurement method for particle accelerator magnets. It is also the most precise method for determining the direction of magnetic flux lines; this is of particular importance in accelerator magnets. The coil geometry is often chosen to suit a particular measurement. One striking example is the Fluxball (26) whose complex construction made it possible to perform point measurements in inhomogeneous fields.

Measurements are performed either by using fixed coils in a dynamic magnetic field or by moving the coils in a static field. The coil movement may be rotation through a given angle, continuous rotation, or simply movement from one position to another. Very high resolution may be reached in field mapping by using this method (27).

Very high resolution may also be reached in differential fluxmeter measurements using a pair of search coils connected in opposition, where one coil moves and the other is fixed, thus compensating for fluctuations in the magnet excitation current and providing a much higher sensitivity when examining field quality. The same principle is applied in harmonic coil measurements, but both coils move. A wide variety of coil configurations is used, ranging from the simple flip-coil to the complex harmonic coil systems used in fields of cylindrical symmetry.

Induction Coils

The coil method is particularly suited for measurements with long coils in particle accelerator magnets (28,29), where the precise measurement of the field integral along the particle trajectory is the main concern. Long rectangular coils were usually employed and are still used in magnets that have a wide horizontal aperture and limited gap height. In this case, the geometry of the coil is chosen to link with selected field components (30). The search coil is usually wound on a core made from a mechanically stable material to ensure a constant coil area, and the wire is carefully glued to the core. Special glass or ceramics that have low thermal dilatation are often used as core materials. During coil winding, the wire must be stretched so that its residual elasticity assures well-defined geometry and mechanical stability of the coil.

Continuously rotating coils that have commutating polarity were already employed in 1880 (3). The harmonic

coil method has now become very popular for use in circular cylindrical magnets, in particular, superconducting beam transport magnets. The coil support is usually a rotating cylinder. This method has been developed since 1954 (31,32). The induced signal from the rotating coil was often transmitted through slip rings to a frequency selective amplifier (frequency analyzer), thus providing analog harmonic analysis.

The principle of a very simple harmonic coil measurement is illustrated in Fig. 2. The radial coil extends through the length of the magnet and is rotated around the axis of the magnet. As the coil rotates, it cuts the radial flux lines. Numerous flux measurements are made between predefined angles. This permits precise and simultaneous determination of the strength, quality, and geometry of the magnetic field. A Fourier analysis of the measured flux distribution results in a precise description of the field parameters in terms of the harmonic coefficients:

$$B_r(r, \varphi) = B_o \sum_{n=1}^{\infty} \left(\frac{r}{r_o}\right)^{n-1} (b_n \cos n\varphi + a_n \sin n\varphi)$$

where B_o is the amplitude of the main harmonic and r_o is a reference radius. b_n and a_n are the harmonic coefficients. In this notation b_1 will describe the normal dipole coefficient, b_2 the normal quadrupole coefficient, etc. The corresponding skew field components are described by the coefficients a_1, a_2 etc.

Due to the advent of modern digital integrators and angular encoders, harmonic coil measurements have improved considerably and are now considered the best choice for most types of particle accelerator magnets, in particular those designed with cylindrical symmetry (33). In practice, the coil is rotated one full turn in each angular direction, and the electronic integrator is triggered at the defined angles by an angular encoder connected to the axis of the coil. To speed up the calculation of the Fourier series, it is an advantage to choose n equally spaced measurement points, where n is a power of 2 (e.g., 512). A compensating coil, connected in series and rotated with the main coil, may be used to suppress the main field component and thus increase the sensitivity of the system for measuring field quality. Dynamic fields are measured with a static coil linking to selected harmonics (34). The harmonic coil measurement principle and its related equipment are described in detail in (35). A thorough description of the general theory, including detailed error analysis, can be found in (36). The practical use of the harmonic coil method for large-scale measurements in superconducting magnets is described in (37,38) and more recent developments are in (39–43)

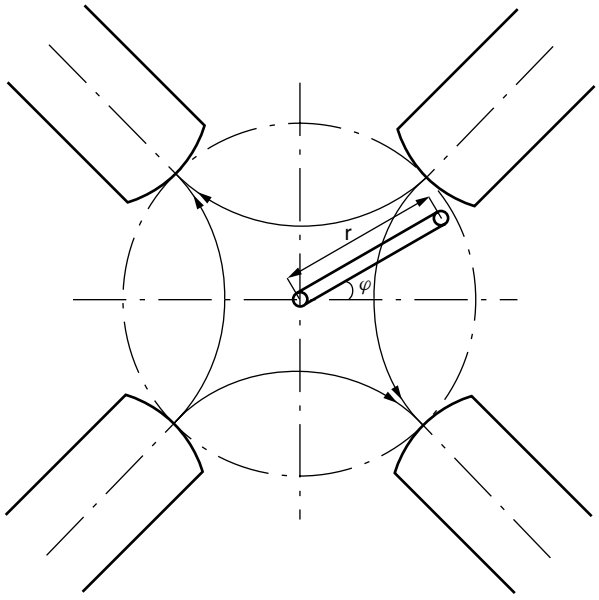


Figure 2. Harmonic coil measurement.

Another induction measurement method consists of moving a stretched wire in a magnetic field, thus integrating the flux cut by the wire (44). It is also possible to measure the flux change while varying the field and keeping the wire in a fixed position. Tungsten is often selected as a wire material, if the wire cannot be placed in a vertical position. The accuracy is determined by the mechanical positioning of the wire. Sensitivity is limited but can be improved by using a multiwire array. This method is well suited to geometric measurements, to the absolute calibration of quadrupole fields, and in particular to measurements in strong magnets that have very small apertures.

The choice of geometry and method depends on the useful aperture of the magnet. The sensitivity of the fluxmeter method depends on the coil surface and on the quality of the integrator. The coil–integrator assembly can be calibrated to an accuracy of a few tens of ppm in a homogeneous magnetic field by reference to a nuclear magnetic resonance probe, but care must be taken not to introduce thermal voltages in the related cables and connectors. Induced erratic signals from wire loops exposed to magnetic flux changes must also be avoided. One must measure the equivalent surface of the search coil and also its median plane which often differs from its geometric plane due to winding imperfections. In long measurement coils, it is important to ensure very tight tolerances on the width of the coil. If the field varies strongly over the length of the coil, it may be necessary to examine the variation of the effective width.

The main advantage of search coil techniques is the possibility of very flexible coil design. The high stability of the effective coil surface is another asset. The linearity and the wide dynamic range also play important roles. The technique can be easily adapted to measurements at cryogenic temperatures. After calibration of the coils at liquid nitrogen temperature, only a minor correction has to be applied for use at lower temperatures. On the other hand, the need for relatively large induction coils and their related mechanical apparatus, which is often complex, may be a disadvantage. Finally, measurements with moving coils are relatively slow.

Flux Measurement

Induction coils were originally used with ballistic galvanometers and later with more elaborate fluxmeters (45). The coil method was improved considerably by the

development of photoelectric fluxmeters (46) which were used for a long time. The measurement accuracy was further improved by the introduction of the classic electronic integrator, the Miller integrator. It remained necessary, however, to employ difference techniques for measurements of high precision (47). Later, the advent of digital voltmeters made fast absolute measurements possible, and the Miller integrator has become the most popular fluxmeter. Due to the development of solid-state dc amplifiers, this integrator has become inexpensive and is often used in multicoil systems.

Figure 3 shows an example of such an integrator. It is based on a dc amplifier that has a very low input voltage offset and a very high open-loop gain. The thermal variation of the integrating capacitor (C) is the most critical problem. Therefore, integrating components are mounted in a temperature-controlled oven. Another problem is the decay of the output signal through the capacitor and the resetting relay. So, careful protection and shielding of these components is essential to reduce the voltages across the critical surface resistances.

The dielectric absorption of the integrating capacitor sets a limit to the integrator precision. A suitable integrating resistor is much easier to find. Most metal-film resistors have stabilities and temperature characteristics that match those of the capacitor. The sensitivity of the integrator is limited by the dc offset and the low-frequency input noise of the amplifier. A typical value is 0.5 μV, which must be multiplied by the measurement time to express the sensitivity in terms of flux. Thermally induced voltages may cause a problem, so care must be taken in choosing of cables and connectors. In tests at CERN, the overall stability of the integrator time constant proved to be better than 50 ppm during a period of three months. A few electronic fluxmeters have been developed by industry and are commercially available.

In more recent years, a new type of digital fluxmeter has been developed, which is based on a high-quality dc amplifier connected to a voltage-to-frequency converter (VFC) and a counter. The version shown in Fig. 4 was developed at CERN and is now commercially available. The input of the VFC is given an offset of 5 V to provide a true bipolar measurement. This offset is balanced by a 500-kHz signal which is subtracted from the output of the VFC. Two counters are used to measure with continuously moving coils and to provide instant readings of the integrator. One of the counters can then be read and reset while the other is active. In this way, no cumulative errors will build up. The linearity of this fluxmeter is 50 ppm. Its sensitivity is limited by the input amplifier, as in the case of an analog amplifier.

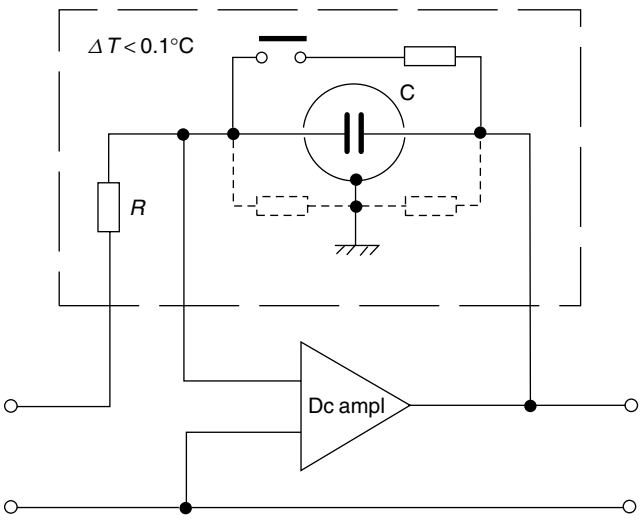


Figure 3. Analog integrator.

This system is well adapted to digital control but imposes limits on the rate of change of the flux because the input signal must never exceed the voltage level of the VFC. To obtain reasonable resolution, the minimum integration period over the full measurement range must be of the order of one second.

The Hall Generator Method

In 1879, E.H. Hall discovered that a very thin metal strip that is immersed in a transverse magnetic field and carries a current develops a voltage mutually at right angles to the current and field that opposed the Lorentz force on the electrons (48). In 1910, the first magnetic measurements were performed using this effect (49). It was, however, only around 1950 that suitable semiconductor materials were developed (50–52) and since then the method has been used extensively. It is a simple and fast measurement method, that provides relatively good accuracy, and therefore it is the most commonly used method in large-scale field mapping (53–55). The accuracy can be improved at the expense of measurement speed.

Hall Probe Measurements

The Hall generator provides an instant measurement, uses very simple electronic measurement equipment, and offers a compact probe that is suitable for point measurements. A large selection of this type of gaussmeter is now commercially available. The probes can be mounted on relatively light positioning gear (55). Considerable measurement time may be gained by mounting Hall generators in modular multiprobe arrays and applying multiplexed voltage measurement (56). Simultaneous measurements in two or three dimensions may also be carried out by using suitable probe arrays (57,58). The wide dynamic range and the possibility of static operation are other attractive features.

However, several factors set limits on the accuracy obtainable. The most serious limitation is the temperature coefficient of the Hall voltage. Temperature stabilization is usually employed to overcome this problem (59), but it increases the size of the probe assembly. The temperature coefficient may also be taken into account in the probe calibration by monitoring the temperature during measurements (60). It also depends, however, on the level of the magnetic field (60), so relatively complex calibration tables are needed. Another complication can be that of the planar Hall effect (61), which makes measuring a weak field component normal to the plane of the Hall generator problematic if a strong field component is parallel to this

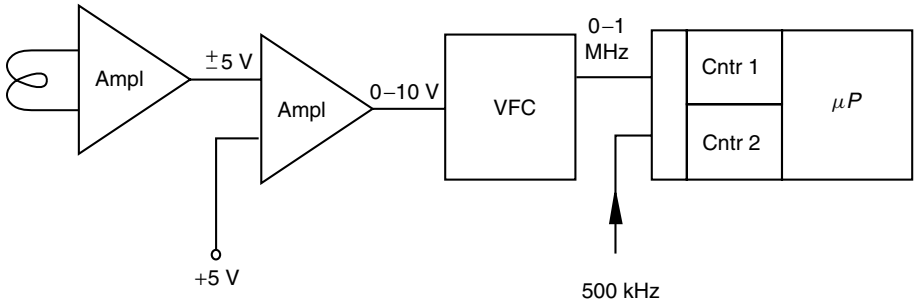


Figure 4. Digital integrator.

plane. This effect limits the use in fields of unknown geometry and in particular its use for determining of field geometry.

Last but not least is the problem of the nonlinearity of the calibration curve because the Hall coefficient is a function of the field level. The Hall generator of the cruciform type (62) has better linearity and a smaller active surface than the classical rectangular generator. Therefore, its magnetic center is better defined, so it is particularly well suited for measurements in strongly inhomogeneous fields. Special types that have smaller temperature dependence are available on the market, but they have lower sensitivity.

The measurement of the Hall voltage sets a limit of about 20 μT on the sensitivity and resolution of the measurement, if conventional dc excitation is applied to the probe. This is caused mainly by thermally induced voltages in cables and connectors. The sensitivity can be improved considerably by applying ac excitation (63,64). Good accuracy at low fields can then be achieved by employing synchronous detection techniques for measuring of the Hall voltage (65).

Special Hall generators for use at cryogenic temperatures are also commercially available. Although they have very low temperature coefficients, they unfortunately reveal an additional problem at low temperatures. The so-called "Shubnikov–de Haas effect" (66,67) shows up as a field-dependent oscillatory effect of the Hall coefficient which may amount to about 1% in high fields, depending on the type of semiconductor used for the Hall generator. This adds a serious complication to calibration. The problem may be solved by locating the Hall generator in a heated anticryostat (68). The complications related to the planar Hall effect are less important at cryogenic temperatures and are discussed in detail in (69). Altogether, the Hall generator has proved very useful for measurements at low temperature (70).

Calibration

Hall generators are usually calibrated in a magnet in which the field is measured simultaneously by the nuclear magnetic resonance technique. The calibration curve is most commonly represented as a polynomial of relatively high order (7 or 9) fitted to a sufficiently large number of calibration points. This representation has the advantage of a simple computation of magnetic induction from a relatively small table of coefficients.

A physically better representation is a piecewise cubic interpolation through a sufficient number of calibration points, which were measured with high precision. This can be done as a simple Lagrange interpolation or even better with a cubic spline function. The advantage of the spline function comes from its minimum curvature and its "best approximation" properties (71). The function adjusts itself easily to nonanalytic functions and is very well suited to interpolation from tables of experimental data. The function is defined as a piecewise polynomial of the third degree that passes through the calibration points so that the derivative of the function is continuous at these points. Very efficient algorithms can be found in the literature (72). The calculation of the polynomial coefficients may be somewhat time-consuming but need only be done once at calibration time. The coefficients (typically about 60 for the bipolar calibration of a cruciform Hall generator) can be easily stored in a microprocessor (59,65), and the subsequent field calculations are very fast. The quality of the calibration function can be verified from field values measured between the calibration points. A well-designed Hall-probe assembly can be calibrated to long-term accuracy of 100 ppm. The stability may be considerably improved by powering the Hall generator permanently and by keeping its temperature constant (56).

Fluxgate Magnetometer

The fluxgate magnetometer (73) is based on a thin linear ferromagnetic core on which detection and excitation coils are wound. The measurement principle is illustrated in Fig. 5. In its basic version, it consists of three coils

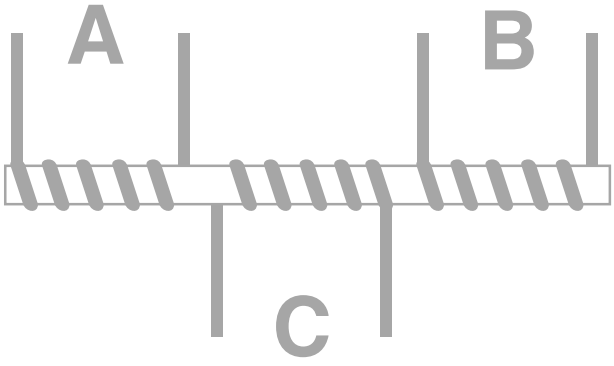


Figure 5. Fluxgate magnetometer.

wound around a ferromagnetic core: an ac excitation winding A, a detection winding B that indicates the zero field condition, and a dc bias coil C that creates and maintains the zero field. In practice, the coils are wound coaxially in successive layers. The core is made from a fine wire of Mumetal or a similar material that has an almost rectangular hysteresis curve. The method was introduced in the 1930s and was also named "peaking strip." It is restricted to use with low fields but has the advantage of offering a linear measurement and is well suited for static operation. As a directional device of very high sensitivity, it is suitable for studying weak stray fields around magnets and mapping the earth's magnetic field. Much more complex coil configurations are wound for precision measurements and where the measured field should not be distorted by the probe. The most interesting application is now in space research; important developments of this technique have taken place over the last decades (74–76). The use of modern materials for magnetic cores has improved the sensitivity to about 20 pT and can assure a wide dynamic range. The upper limit of the measurement range is usually of the order of a few tens of mT, but it can be extended by applying water cooling to the bias coil. Fluxgate magnetometers that have a typical range of 1 mT and a resolution of 1 nT are commercially available from several sources. They have many other practical applications, for example, in navigation equipment.

Magnetoresistivity Effect

Magnetoresistivity was discovered by W. Thomson in 1856 (77). It was exploited quite early, and a commercial instrument already existed at the end of the last century. Technical problems were, however, significant (78). Dependence on temperature and mechanical stress, combined with difficulties of manufacture and problems with electrical connections, caused a general lack of reliability in this measurement method. Similarly to the Hall generator, it was only when semiconductor materials became available that the method turned into a success. Then, inexpensive magnetoresistors came on the market and were also used for magnetic measurements (79). A more recent application for field monitoring was implemented in one of the large LEP spectrometers at CERN (80).

Visual Field Mapping

The best known visual field mapper is made by spreading iron powder on a horizontal surface placed near a magnetic source, thus providing a simple picture of the distribution of flux lines. Another very classical way of observing flux-line patterns is to place a free-moving compass needle at different points in the volume to be examined and note the direction of the needle. This compass method was applied, long before the discovery of electromagnetism, to studies of the variations in the direction of the earth's magnetic field. Another visual effect may be obtained by observing the light transmission through a colloidal suspension of diamagnetic particles subject to the field (81,82).

Faraday Effect

The magneto-optical rotation of the plane of polarization of polarized light (Faraday effect) is a classical method for visualizing magnetic fields. A transparent container filled with a polarizing liquid and placed inside the magnet gap may reveal, for example, the field pattern in a quadrupole by observation through polarization filters placed at each end of the magnet. The rotation of the plane is proportional to the field strength and the length of the polarizing medium and may give a certain indication of the field geometry. This measurement principle has proved useful for measuring transient magnetic fields (83,84). It is less convincing when applied to the precise determination of magnet geometry, even though modern image processing techniques might improve the method substantially.

Floating Wire Method

Floating wire measurements were quite popular in the past (85). If a current-carrying conductor is stretched in a magnetic field, it will curve subject to the electromagnetic force and describe the path of a charged particle whose momentum corresponds to the current and the mechanical tension in the wire. A flexible annealed aluminium wire was used to reduce the effects of stiffness and gravity. This method has now been entirely replaced by precise field mapping and simulation of particle trajectories by computer programs.

Measurements Based on Particle Beam Observation

A method for precisely measuring the beam position with respect to the magnetic center of quadrupole magnets installed in particle accelerators has been developed during the last decade (86,87). The procedure consists of modulating the field strength in individual lattice quadrupoles while observing the resulting beam orbit oscillations. Local dc orbit distortions are applied in the search for the magnetic center. This so-called K-modulation provides perfect knowledge of the location of the particle beam with respect to the center of a quadrupole. In addition, it may provide other very useful observations for operating and adjusting of the accelerator (88). This is obviously of particular importance for superconducting accelerators (89). It is very difficult to provide a superconducting quadrupole magnet that has a direct optical reference to its magnetic center, so errors caused by changes of temperature profiles and other phenomena may build up as time passes.

The method may be further improved by synchronous detection of the oscillation, so that its phase can be identified. The sensitivity of the detection is impressive. Experience from LEP (90) showed that an absolute accuracy of 0.05 mm could be obtained in both the vertical and horizontal planes. Furthermore, it was observed that modulation of the quadrupole field by about 300 ppm could be clearly detected, which means that measurements may be carried out on colliding beams while particle physics experiments are taking place. This measurement method also played an important role for adjusting the so-called Final Focus Test Beam at Stanford Linear Accelerator Center (SLAC) (91,92).

Magnetoinductive Technology

Magnetic field sensors have been developed based on a change in inductance (L) caused by an applied magnetic field (93). These sensors, referred to as magnetoinductive sensors, contain an alloy whose permeability changes linearly over the sensors' useful range in an applied magnetic field. When the alloy is incorporated into an inductor, the inductance will change as the applied field changes. Magnetoinductive sensors contain a resonant LC circuit. As the applied field changes, so does the resonant frequency of the circuit. These devices have a dynamic range of $\pm$ 1,000 μT and an accuracy of $\pm$ 0.4 μT.

CONCLUDING REMARKS

Proven measurement methods and powerful equipment are readily available for most of the measurement tasks related to beam-guiding magnets as well as for spectrometer magnets. Therefore, it is prudent to examine existing possibilities carefully before launching the development of a more exotic measurement method. Many unnecessary costs and unpleasant surprises can be avoided by choosing commercially available instruments. The measurement methods described are complementary, and a combination of two or more of them will certainly meet most requirements.

In the field of new technologies, two methods merit consideration. Magnet resonance imaging is a promising technique, that could find lasting application. The use of superconducting quantum interference devices (SQUIDS) might also in the long run become an interesting alternative as an absolute standard and for measuring of weak fields (94,95). The complexity of these methods still prevents current laboratory use.

BIBLIOGRAPHY

1. J. L. Symonds, *Rep. Progr. Phys.* **18**, 83–126 (1955).
2. C. Germain, *Nucl. Instr. Meth.* **21**, 17–46 (1963).
3. L. W. McKeehan, *J. Opt. Soc. Am.* **19**, 213–242 (1929).
4. S. Turner, ed., *Proceedings, CERN Accelerator School, Measurement and Alignment of Accelerator and Detector Magnets*, CERN 98-05, Geneva, Switzerland, 1998.
5. J. J. Rabi, J. R. Zacharias, S. Millman, and P. Kusch, *Phys. Rev.* **53**, 318 (1938).
6. J. J. Rabi, S. Millman, P. Kusch, and J. R. Zacharias, *Phys. Rev.* **55**, 526–535 (1939).
7. E. M. Purcell, H. C. Torrey, and R. V. Pound, *Phys. Rev.* **69**, 37–38 (1946).
8. F. Bloch, W. W. Hansen, and M. Packard, *Phys. Rev.* **69**, 127 (1946).
9. F. Bloch, W. W. Hansen, and M. Packard, *Phys. Rev.* **70**, 474–485 (1946).
10. N. Bloembergen, E. M. Purcell, and R. V. Pound, *Phys. Rev.* **73**, 679–712 (1948).
11. K. Borer and G. Fremont, *CERN* **77**, 19–42 (1977).
12. W. G. Clark, T. Hijmans, and W. H. Wong, *J. Appl. Phys.* **63**, 4,185–4,186 (1988).
13. R. Prigl et al., *Nucl. Instr. Meth.* **A374**, 118–126 (1996).
14. W. G. Clark, J. M. Moore, and W. H. Wong, *Proc. of the 2nd Int. Ind. Symp. on the Supercollider*, Miami Beach, 1990, pp. 405–414.
15. G. Suryan, *Proc. Indian Acad. Sci.* **A33**, 107–111 (1951).
16. C. Sherman, *Rev. Sci. Instr.* **30**, 568–575 (1959).
17. J. M. Pendlebury et al., *Rev. Sci. Instr.* **50**, 535–540 (1979).
18. L. Jansak et al., *Proceedings of the Int. conf. on Measurement, Measurement 97*, Smolenice, Slovakia, May 1997, pp. 328–331.
19. D. Bourdel, J. Pescia, and P. Lopez, *Rev. Phys. Appl.* **5**, 187–190 (1970).
20. F. Hartmann, *IEEE Trans. Magn.* **MAG-8**(1), 66–75 (1972).
21. D. Duret et al., *Rev. Sci. Instr.* **62**(3), 685–694 (1991).
22. N. Kernevez, D. Duret, M. Moussavi, and J. -M. Leger, *IEEE Trans. Magn.* **28**(5), 3,054–3,059 (1992).
23. D. A. Gross, *Proc. of the ICFA Workshop on Superconducting Magnets and Cryogenics*, Brookhaven National Lab., Upton, May 1986, pp. 309–311.
24. J. H. Coupland, T. C. Randle, and M. J. Watson, *IEEE Trans. Magn.* **MAG-17**, 1,851–1,854 (1981).
25. W. Weber, *Ann. Physik* **2**, 209–247 (1853).
26. W. F. Brown and J. H. Sweer, *Rev. Sci. Instr.* **16**, 276–279 (1945).
27. A. Daël et al., *Int. J. Mod. Phys. A* **2B**(HEACC'92), 650–652 (1993).
28. E. A. Finlay, J. F. Fowler, and J. F. Smee, *J. Sci. Instr.* **27**, 264–270 (1950).
29. B. C. Brown, *Proc. of the ICFA Workshop on Superconducting Magnets and Cryogenics*, Brookhaven National Lab., Upton, May 1986, pp. 297–301.
30. B. de Raad, Thesis, Delft, 1958, pp. 55–67.
31. W. C. Elmore and M. W. Garrett, *Rev. Sci. Instr.* **25**, 480–485 (1954).
32. I. E. Dayton, F. C. Shoemaker, and R. F. Mozley, *Rev. Sci. Instr.* **25**, 485–489 (1954).
33. C. Wyss, *Proc. 5th Int. Conf. on Magnet Technology (MT-5)*, Frascati, Italy, 1975, pp. 231–236.
34. G. H. Morgan, *Proc. 4th Int. Conf. on Magnet Technology*, Brookhaven National Lab., Upton, 1972, pp. 787–790.
35. L. Walckiers, CERN Accelerator School, Montreux, Switzerland, *CERN* **92-05**, 138–166 (1992).
36. W. G. Davies, *Nucl. Instr. Meth.* **A311**, 399–436 (1992).
37. H. Brück, R. Meinke, and P. Schmüser, *Kerntechnik* **56**, 248–256 (1991).
38. P. Schmüser, CERN Accelerator School, Montreux, Switzerland, *CERN* **92-05**, 240–273 (1992).
39. M. I. Green, R. Sponsel, and C. Sylvester, *Proc. of the 5th Int. Ind. Symp. on the Supercollider*, San Francisco, 1993, pp. 711–714.
40. R. Thomas et al., *Proc. of the 5th Int. Ind. Symp. on the Supercollider*, San Francisco, 1993, pp. 715–718.
41. J. Billan et al., *IEEE Trans. Magn.* **30**(4) (MT-13), 2,658–2,661 (1994).
42. J. Buckley, D. Richter, L. Walckiers, and R. Wolf, *IEEE Trans. Appl. Superconducte* **5**(2) (ASC'94), 1,024–1,027 (1995).
43. J. Billan, S. De Panfilis, D. Giloteaux, and O. Pagano, *IEEE Trans. Magn.* **32**(4) (MT-14), 3,073–3,076 (1996).
44. D. Zangrando and R. P. Walker, *Nucl. Instr. Meth.* **A376**, 275–282 (1996).
45. M. E. Grassot, *J. Phys.* **4**, 696–700 (1904).

46. R. F. Edgar, *Trans. Am. Inst. Elect. Eng.* **56**, 805–809 (1937).
47. G. K. Green, R. R. Kassner, W. H. Moore, and L. W. Smith, *Rev. Sci. Instr.* **24**, 743–754 (1953).
48. E. H. Hall, *Amer. J. Math.* **2**, 287–292 (1879).
49. W. Peukert, *Elektrotechn. Zeitschr.* **25**, 636–637 (1910).
50. G. L. Pearson, *Rev. Sci. Instr.* **19**, 263–265 (1948).
51. H. Welker, *Z. Naturforschung* **7a**, 744–749 (1952).
52. H. Welker, *Elektrotechn. Zeitschr.* **76**, 513–517 (1955).
53. E. Acerbi et al., *IEEE Trans. Magn.* **MAG-17**(MT-7), 1,610–1,613 (1981).
54. C. Bazin et al., *IEEE Trans. Magn.* **MAG-17**, 1,840–1,842 (1981).
55. D. Swoboda, *IEEE Trans. Magn.* **MAG-17**, 2,125–2,128 (1981).
56. M. Tkatchenko, Private communication.
57. S. Kawahito, S. O. Choi, M. Ishida, and T. Nakamura, *Sensors and Actuators A* **40**, 141–146 (1994).
58. J. Kvitkovic and M. Majoros, *6th European Mag. Mater. Appl. Conf.*, Vienna, Austria, 1995.
59. K. Brand and G. Brun, *CERN* **79**, 02–24 (1979).
60. M. W. Poole and R. P. Walker, *IEEE Trans. on Magn.* **MAG-17**, 2,129–2,132 (1981).
61. C. Goldberg and R. E. Davis, *Phys. Rev.* **94**, 1,121–1,125 (1954).
62. J. Hauesler and H. J. Lippmann, *Solid State Electron.* **11**, 173–182 (1968).
63. J. J. Donoghue and W. P. Eatherly, *Rev. Sci. Instr.* **22**, 513–516 (1951).
64. C. D. Cox, *J. Sci. Instr.* **41**, 695–691 (1964).
65. K. R. Dickson and P. Galbraith, *CERN* **85**, 13–42 (1985).
66. J. Babiskin, *Phys. Rev.* **107**, 981–992 (1957).
67. H. P. R. Frederikse and W. R. Hosler, *Phys. Rev.* **110**, 880–883 (1958).
68. M. Polak, *Rev. Sci. Instr.* **44**, 1,794–1,795 (1973).
69. M. Polak and I. Hlasnik, *Solid State Electron.* **13**, 219–227 (1970).
70. J. Kvitkovic and M. Polak, *Eur. Conf. on Appl. Superconductivity (EUCAS)*, Gottingen, Germany, 1993, pp. 1,629–1,632.
71. J. L. Walsh, J. H. Ahlberg, and E. N. Nilson, *J. Math. Mech.* **11**, 225–234 (1962).
72. A. Ralston and H. Wilf, eds., *Mathematical Methods for Digital Computers*, vol. 2, John Wiley & Sons, Inc., New York, NY, 1967, pp. 156–168.
73. J. M. Kelly, *Rev. Sci. Instr.* **22**, 256–258 (1951).
74. D. I. Gordon and R. E. Brown, *IEEE Trans. Magn.* **MAG-8**, 76–82 (1972).
75. F. Primdahl, *J. Phys. E: Sci. Instr.* **12**, 241–253 (1979).
76. O. V. Nielsen, T. Johansson, J. M. Knudsen, and F. Primdahl, *J. Geophys. Res.* **97**, 1,037–1,044 (1992).
77. W. Thomson, *Philos. Trans.* **146**, 736–751 (1856).
78. P. Kapitza, *Proc. R. Soc. A* **119**, 358 (1928).
79. E. Welch and P. R. Mace, *Proc. 3rd Int. Conf. on Magn. Technol.*, Hamburg, Germany, 1970, pp. 1,377–1,391.
80. G. Brouwer et al., *Nucl. Instr. Meth.* **A313**, 50–62 (1992).
81. J. K. Cobb and J. J.Muray, *Nucl. Instr. Meth.* **46**, 99–105 (1967).
82. D. Trbojevic et al., *1995 Particle Accelerator Conf., Dallas*, (May 1995), pp. 2,099–2,021.
83. J. Malecki, M. Surma, and J. Gibalewicz, *Acta Phys. Polon.* **16**, 151–156 (1957).
84. J. L. Robertson, D. T. Burns, and D. Hardie, *Nucl. Instr. Meth.* **203**, 87–92 (1982).
85. L. G. Ratner and R. J. Lari, *Proc. Int. Symp. on Magn. Technol.*, Stanford, 1965, pp. 497–504.
86. D. Rice et al., *IEEE Trans. Nucl. Sci.* **NS-30**, 2,190–2,192 (1983).
87. P. Rojsel, *Nucl. Instr. Meth.* **A343**, 371–382 (1994).
88. R. Brinkmann and M. Boge, *4th Euro. Particle Accelerator Conf.*, London, 1994, pp. 938–940.
89. J. Deregel et al., *CERN LHC Project Report 4*, (March 1996).
90. I. Barnett et al., *6th Beam Instrumentation Workshop*, Vancouver, Canada, 1994.
91. F. Bulos et al., *1991 Particle Accelerator Conf.*, San Francisco, (May 1991), pp. 3,216–3,218.
92. P. Tenenbaum et al., *1995 Particle Accelerator Conf.*, Dallas, (May 1995), pp. 2,096–2,098.
93. *Digital Compass and Magnetometer having a Sensor Coil Wound on a High Permeability Isotropic Core*, US Pat. 4,851,775, N. H. Kim and T. Hawks.
94. G. L. Romani, *Proc. 9th Int. Conf. on Magn. Technol. (MT-9)*, Zurich, Switzerland, 1985, pp. 236–242.
95. D. Drung, *Eur. Conf. Appl. Superconductivity (EUCAS)*, Gottingen, Germany, 1993, pp. 1,287–1,294.

MAGNETIC RESONANCE IMAGING

ROBERT W. PROST
Medical College of Wisconsin
Milwaukee, WI

INTRODUCTION

Imaging of materials that contain nuclei of nonzero spin values has become a valuable tool in science, medicine, and industry. The principal use of MRI is to create an image of the distribution of certain physical properties of a substance, whether in the fluid-filled rock surrounding an oil well (1–4), soil (5), food-processing (6–11), industrial process control (12) or a human subject (13,14). These physical properties imaged include the chemistry of the material (15), its fluidity (9), its interactions with other materials in the molecular-level neighborhood (8), and even the flow of liquids within the object (16). Magnetic resonance phenomena occur on a timescale that is short relative to many chemical reactions and can often be used to investigate the kinetics of chemical reactions (17,18).

Many different elements are amenable to investigation by magnetic resonance, but the sensitivity of MR is poor. Typically, only one out of 100,000 nuclei can be detected in a magnetic field of 15,000 Gauss or 1.5 Tesla (19). This is analogous to detecting a single voice in the crowd at a University of Michigan football game (capacity 109,000)! As a result of this insensitivity, magnetic resonance is not suited to detecting of chemical moities of less than millimolar quantities. For the same reason, gases are largely unsuitable to investigation by magnetic resonance, except for hyperpolarized gases ^{3}He and ^{129}Xe (20,21).

All solid or liquid objects, that bear suitable nuclei can be imaged. However, spins within solid materials

lose their signal on a very short timescale. This brevity of signal lifetime makes imaging quite difficult, yet not impossible (22).

Examples of nuclei that bear a magnetic moment, or nonzero spin include hydrogen, helium-3, xenon-129, phosphorus, lithium, fluorine, carbon-13 and sodium-23.

An important and recent use of the MR phenomenon is in medical imaging. As of 1999, more than 4,500 MRI systems were installed in the United States (23). The reason for the success of MRI in medical imaging is fourfold. The first is the abundance of hydrogen in the human body. Hydrogen nuclei in the brain are about 70 Molar in concentration (24) which is much higher than any other material that can be detected by NMR. The second is that the hydrogen is distributed in the organs in a way which allows using the MR signal to create images. The third is that imaging of hydrogen produces exquisitely detailed images of internal organ structure. The fourth is the functional data that can be extracted. These data include molecular diffusion (25), distribution of chemical moieties (26), flow (27), flow in the capillary bed (28), oxygen utilization (29,30), magnetization transfer rates (31,32), and disruption of the blood–brain barrier (33).

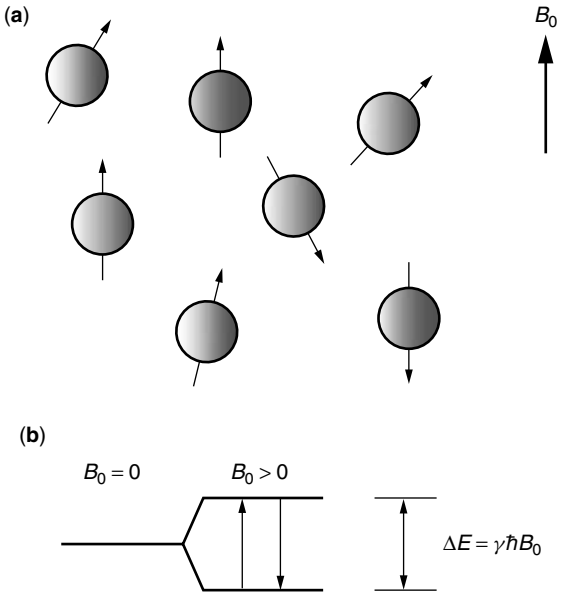


Figure 1. (**a**) Cartoon of nuclei showing alignment parallel and antiparallel to the main magnetic field B_0. (**b**) Quantum mechanical description of spin system in the degenerate ($B = 0$) and nondegenerate ($B > 0$) state, where the energy level difference between states is proportional to B_0.

BASIC MRI PHYSICS

A constant magnetic field on a group or ensemble of spins is creates a slight polarization where the ratio of spins antiparallel to B_0 to the spins parallel to B_0 is given by (34)

$$\frac{N_-}{N_+} = e^{(-\lambda \hbar H_0/kT)}, \tag{1}$$

where N_- is the number of spins antiparallel to B_0,
N_+ is the number of spins parallel to B_0,
γ is the gyromagnetic ratio of the nucleus,
$\hbar$ is the Schroedinger's constant,
k is Boltzmann's constant, and
T is the sample temperature.

This effect can be visualized as a tendency to align parallel or antiparallel to B_0. A classical conception suggests that this is caused by the torque exerted on the intrinsic magnetic field of the spin by the constant magnetic field B_0 (Fig. 1a). The polarization proportional to B_0 is very slight and accounts for only one out of 100,000 spins for hydrogen nuclei in a 1.5 Tesla magnet at room temperature. It is only this miniscule fraction of the spin ensemble, from which the MR signal can be created. The reason that the polarization is so slight is that the magnetic field competes with the randomizing effect of the background thermal vibrations. Lowering the sample temperature can dramatically increase the polarization of the spin ensemble. Usually, this is not an option in intact biological systems.

The correct description of the magnetic resonance phenomenon requires extensive use of quantum mechanics. The quantum mechanical description of the effect of B_0 is to split the energy levels of the spin system (Fig. 1b) (34). All further processes can be mathematically described by the evolution of the spin states (35). However, the quantum mechanical description does not provide an intuitive feeling for magnetic resonance phenomena. In 1946, Felix Bloch developed a phenomenological description of magnetic resonance, which visualizes the sum of the spins as a rotating vector (Fig. 2) that precesses around the direction of B_0 (36). The rate of the precession f is proportional to B_0, as described by the Larmor equation:

$$f = \gamma B_0. \tag{2}$$

The constant of proportionality γ is known as the gyromagnetic ratio. The value of γ is a property of the magnetic moment of the nucleus and thus is different

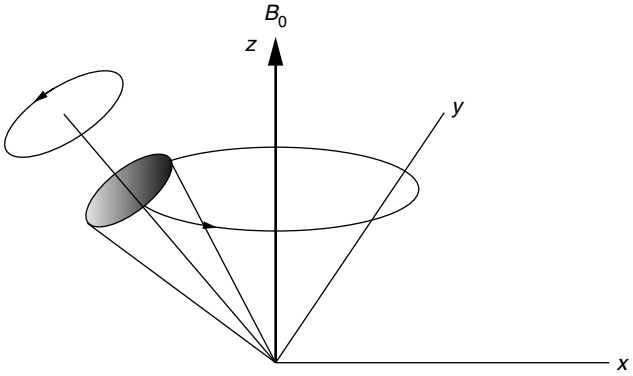


Figure 2. In the phenomenological model proposed by Bloch, the spin precesses about the B_0 vector, and the smaller circle illustrates the rotation of the sphere of charge which is the spin. The rotation gives rise to the intrinsic magnetic moment of the spin. The coordinate system used in this diagram is arbitrary but will be used throughout this article for consistency.

for all nuclei. The gyromagnetic ratio of hydrogen is 42.5759 MHz/Tesla. Therefore, if the main magnetic field is 1.5 Tesla, the frequency of precession is 63.86 MHz. This frequency is the resonance condition. Nuclei that have zero spin, such as carbon-12, have a $\gamma = 0$ and thus do not exhibit the resonance condition.

Excitation and Reception

To produce a detectable signal, the spin vector M_0 must be rotated at least partially into the transverse plane (Fig. 3), so that it creates an induced signal that can be detected by an antenna. To simplify the explanation of the way this is done, consider the rotating frame of reference. If one imagines the observer rotating around the z axis at the Larmor frequency, then the precession of the aligned spins becomes stationary (Fig. 4). The rotating frame of reference will be used for all further description of spin motion in this article. In the rotating frame, the B_0 field appears to have vanished because all of the precession is provided by the rotation of the frame. This ability to "ignore" B_0 in the rotating frame is one of the principal values of this approach. To effect the rotation of M_0, a second magnetic field, denoted B_1, is created in the transverse plane (37) in the rotating frame. Because the B_1 field has to rotate at the Larmor frequency, the B_1 field is created by a radio-frequency pulse whose frequency is at or near the resonant frequency $f = \gamma B_0$. The B_1 field is also known as an RF pulse because the duration of the pulse is usually quite limited, typically 1–10 ms. In the rotating frame, it is easy to see why this B_1 field will rotate the spins into the transverse plane. Because the B_0 field has vanished in the rotating frame, the spins simply precess about B_1. After a short time, they have rotated perpendicularly to their original direction, at which time the B_1 can be switched off. The B_1 field is created by transmitting the radio-frequency pulse into an antenna. Antennas used in radio communication are designed to radiate power by creating electromagnetic waves. The antennas used in magnetic resonance are designed to produce a pure magnetic field (38). This is important because the nuclear spins do not respond to an electric field such as that produced by a dipole communication antenna. In many instances, the antenna used for excitation is also used to receive the MR signal emitted by the sample. For historical reasons, MR antennas are referred to as coils. A coil positioned so that its normal vector is perpendicular to the B_0 field, as shown in Fig. 5 has a current induced in it by the transverse component of the magnetization vector M_{xy} (39). The graph in Fig. 6a illustrates the received signal when the resonant frequency of the spins ($= \gamma B_0$) is set the same as the operating frequency of the transceiver. If the operating frequency of the transceiver is changed by 100 Hz, the resulting signal will be as shown in Fig. 6b. Importantly, the same effect can be created by changing the B_0 field by (100 Hz/γ) gauss. It can be seen in this example that setting the frequency of operation to that of the spin system allows observing the spins in the "rotating frame of reference" where the magnetization vector no longer precesses about B_0. Slight frequency changes then

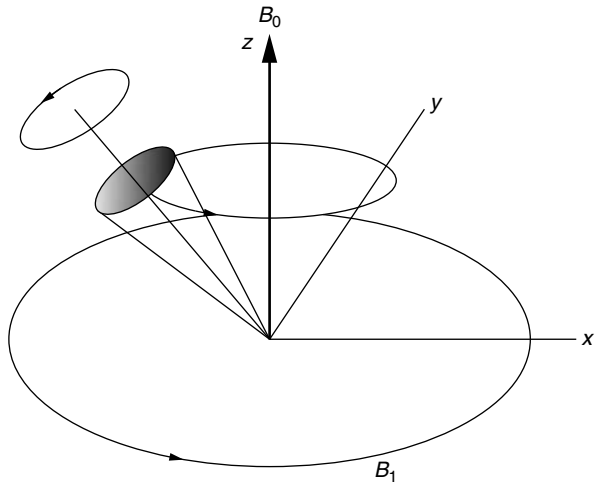


Figure 3. The B_1 (RF transmission) field in the laboratory frame of reference.

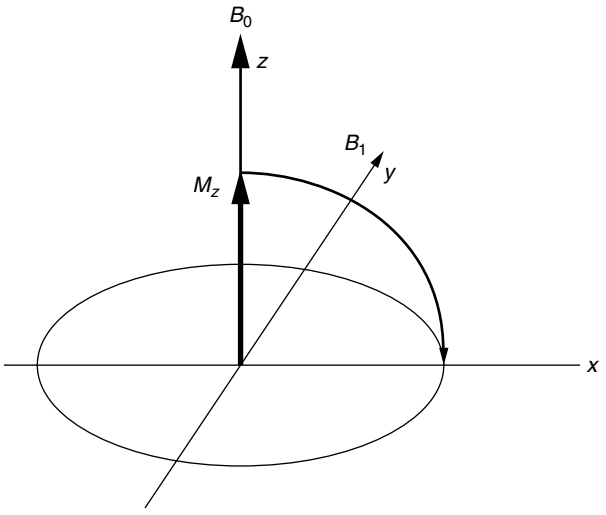


Figure 4. Effect of the B_1 field on the magnetic moment (M_z) in the rotating frame of reference. The torque provided by the B_1 field rotates the M_z vector into the transverse (x, y) plane where it becomes M_{xy}.

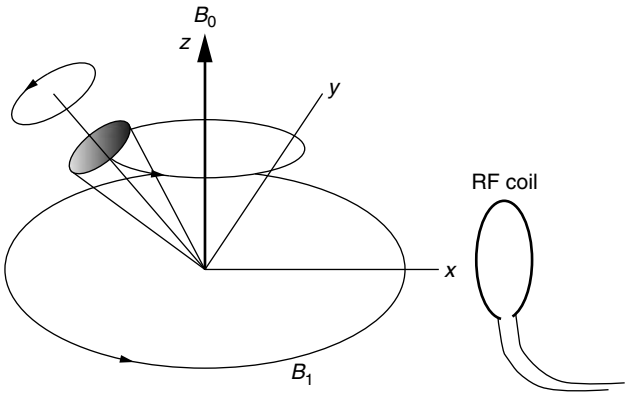


Figure 5. Spin system in the rotating frame of reference showing the induction of a signal in an RF coil whose normal vector is in the transverse plane.

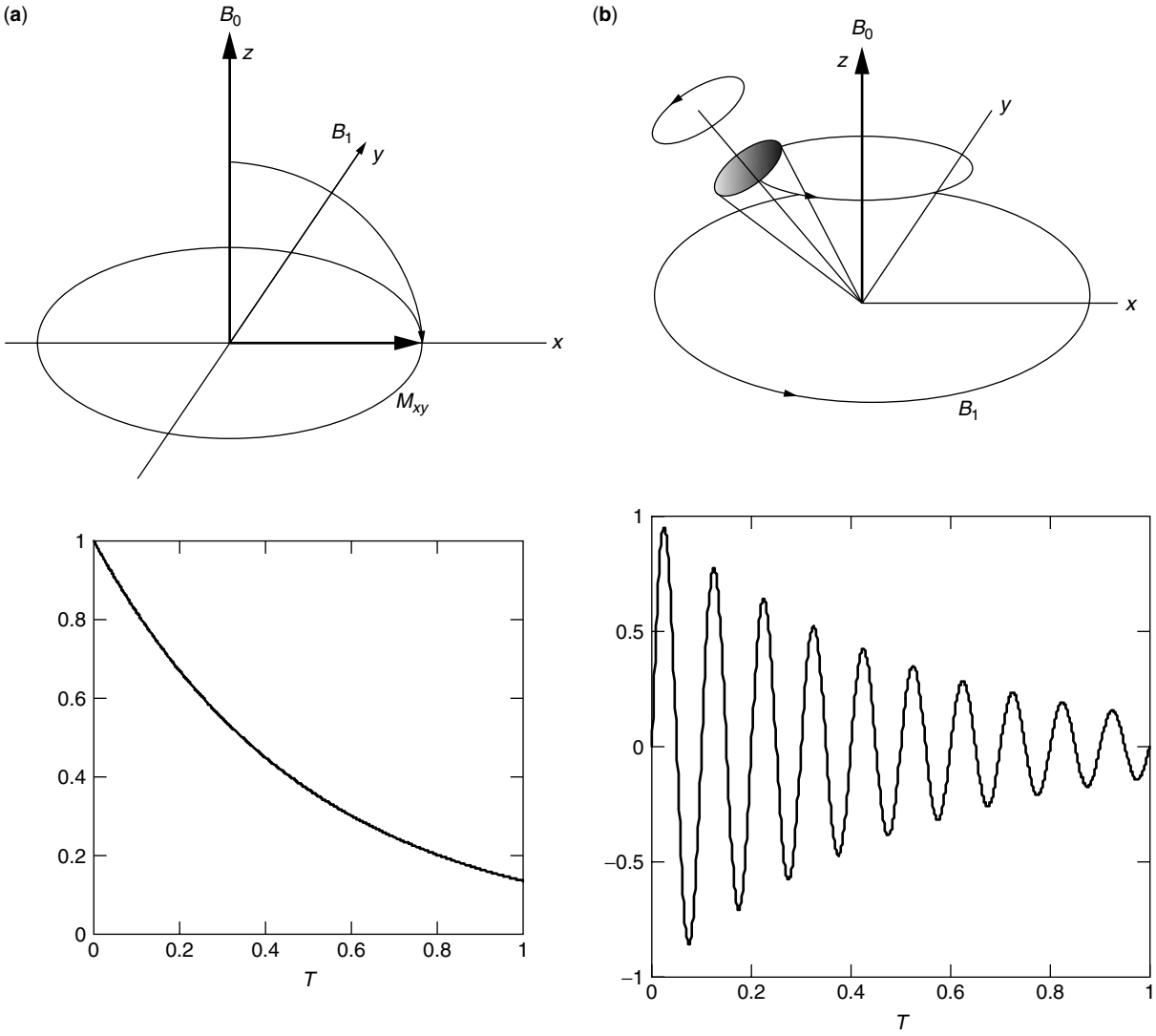


Figure 6. Transverse magnetization decreases with time in (**a**) the rotating frame at resonance and (**b**) the rotating frame 100 Hz off-resonance.

make the magnetization vector appear to move forward or back. Through spins are not always on resonance, it is instructive and makes understanding the spin simple system within the framework of the phenomenological Bloch model.

The power of the excitatory signal may be of the order of 10,000 watts when the MR of the body is at 1.5 Tesla, but the received signal is tiny, of the order of 10^{-12} watts. A signal this small is easily overwhelmed by stray fields from broadcast television, radio, and computer sources. Therefore, the MR procedure is always conducted within an electrically shielded enclosure like a Faraday cage.

The Free Induction Decay

The signal shown in Fig. 6, usually referred to as a free induction decay signal (FID), is the signal produced by the transverse component of the spins, M_{xy}, immediately after excitation. The frequency of precession for a spin is determined only by its local magnetic field. That field is the sum of the imposed B_0 field and typically far smaller fields. The smaller fields are the result of the nonuniformity of the B_0 field and other objects that distort the main field. The field can be distorted by magnetic objects, be they ferromagnetic, paramagnetic, or diamagnetic. The field can also be distorted by surfaces between regions of strongly different magnetic susceptibility such as that between the sample and the surrounding air. The main magnetic fields used in MR must be highly spatially homogeneous. A typical MR magnet used in clinical imaging may have an inhomogeneity of less than one part per million (ppm) across a 20-cm diameter spherical volume. The 1 ppm inhomogeneity translates to a difference of 0.015 Gauss on a 1.5 T magnet. In a group of spins within a sample, some spins will precess over a range of frequencies. In the rotating frame then, some spins will appear to precess in a forward direction, and others appear to precess backward. On a timescale from microseconds to seconds, the spins will spread out to form a uniform disk in the transverse plane, as in Fig. 7.

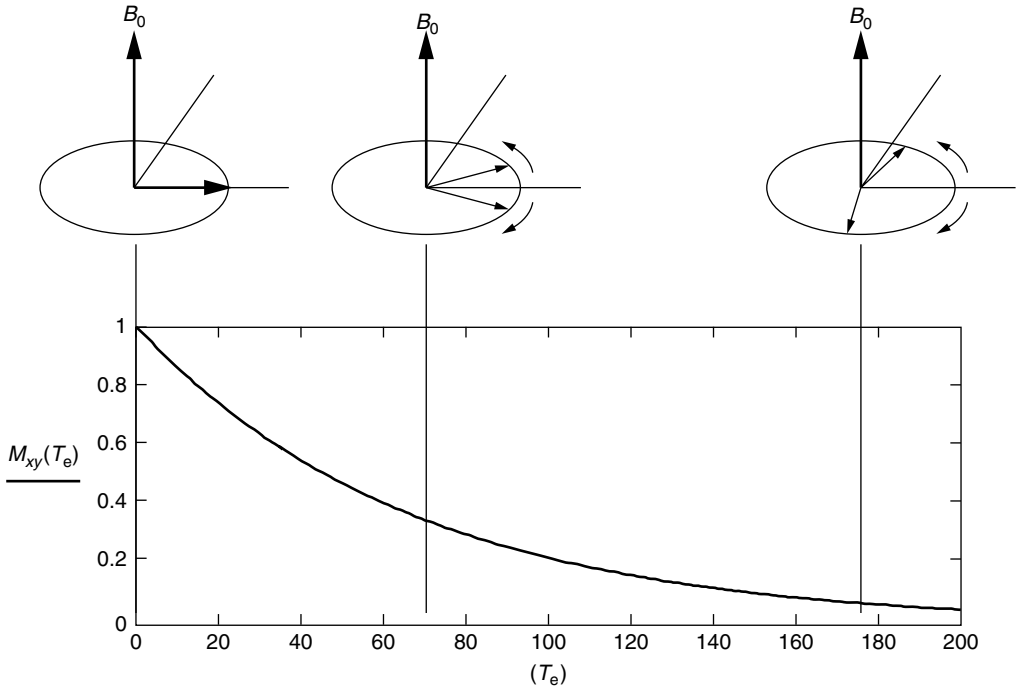


Figure 7. M_{xy} versus time for a group of spins on resonance. Increasing time allows the spins to disperse, where some spins precess faster, and others precess slower causing the sum vector M_{xy} to decrease exponentially with time. This is known as transverse relaxation.

When the spins are uniformly spread in the transverse plane, no signal is induced in the coil because there is no change in the net magnetic vector M_{xy} with time.

Echo Formation

The Gradient Echo. The FID is not the only method to receive a signal from the sample. The spin system may be manipulated to change the time at which M_{xy} is maximum. The signal thus formed is referred to as an echo, similar to the echo received in a radar system (40). For a gradient echo, the B_1 pulse tips the M_0 into the transverse plane as for the FID, but then a gradient magnetic field is applied that produces a strong dispersion of precessional frequencies in the sample, as in Fig. 8. So dispersed, no net M_{xy} vector remains, and the signal vanishes. The dispersion is a linear function of the distance from the origin in the direction of the gradient. By later reversing the sign of the magnetic gradient field which was applied, coherence is restored (41). At long sequence repetion times T_r, the amplitude of the signal is governed by a characteristic decay rate of the spin system called $T2^*$. At short T_r times, where the gradient echo is often used, image contrast can become a mixture of $T1$ and $T2^*$ contrast. At short T_r times, significant M_{xy} may remain that will then interact with the following B_1 pulse and generate a signal that interferes with the desired signal causing image ghosting. To discard this signal, a spoiling gradient may be inserted at the end of the pulse sequence. More effectively, the phase of the B_1 pulse can be varied. Alternatively, the interfering signal can be collected instead of the gradient echo. This signal, formed by the interaction of successive B_1 pulses is strongly $T2$ weighted. The family of pulse sequences based on this method is called steady-state free precession (SSFP) (42) or fast imaging with steady precession (FISP) (43). The available signal-to-noise of the technique is low due to the low flip angle of the B_1 pulses used.

The Spin Echo. Inhomogeneities in the magnetic field are not the only source of decay in the FID or gradient echo signal. Spins that are close to one another can exchange energy and lose coherence. Additionally, spins that move between the time the system is excited and when it is sampled will experience different local magnetic fields either from local inhomogeneities or imposed gradient magnetic fields. This is a diffusional motion. If one imposes another B_1 pulse (Fig. 9a) at a point halfway between the first B_1 pulse and the echo time, the spin system can be refocused (44,45). The second B_1 pulse is a 180° pulse and is twice as strong as the first pulse, a 90° pulse. The action of the 180° pulse can be thought of as follows. Imagine that the disk of spins that are spreading out is flipped over (Fig. 9b). The spins, still traveling in the same direction, are now heading back toward the point in the transverse plane where they were an instant after the initial 90° pulse. At a time equal to that between the 90° pulse and the 180° pulse, the spins will again reach the point of coherence (Fig. 9c). The signal which was lost due to magnetic field inhomogeneity is regained because spins retarded or advanced by local errors in the magnetic field all arrive together. The only signal lost in the spin-echo method comes from those spins that move microscopically (diffusion) or macroscopically(gross motion of the sample). The method of the spin echo was first described by Erwin Hahn in 1950 (46).

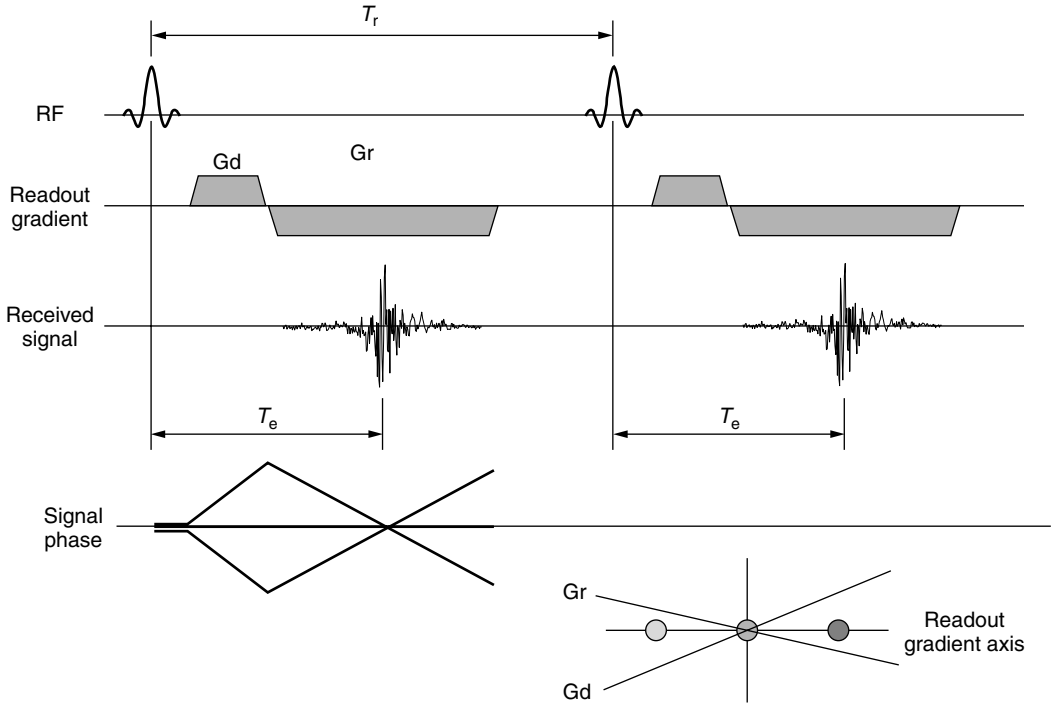


Figure 8. Echo formation in a gradient echo sequence. The effect of the frequency encoding gradient (Gd) is to add signal phase to the spins proportional to the distance from the isocenter in the direction of the gradient. The readout gradient (Gr) rewinds the phase, causing the echo to peak at the center of the readout time, T_e. The signal phase for three spins spaced out in the direction of the readout gradient (represented by red, green, and blue dots) is also plotted.

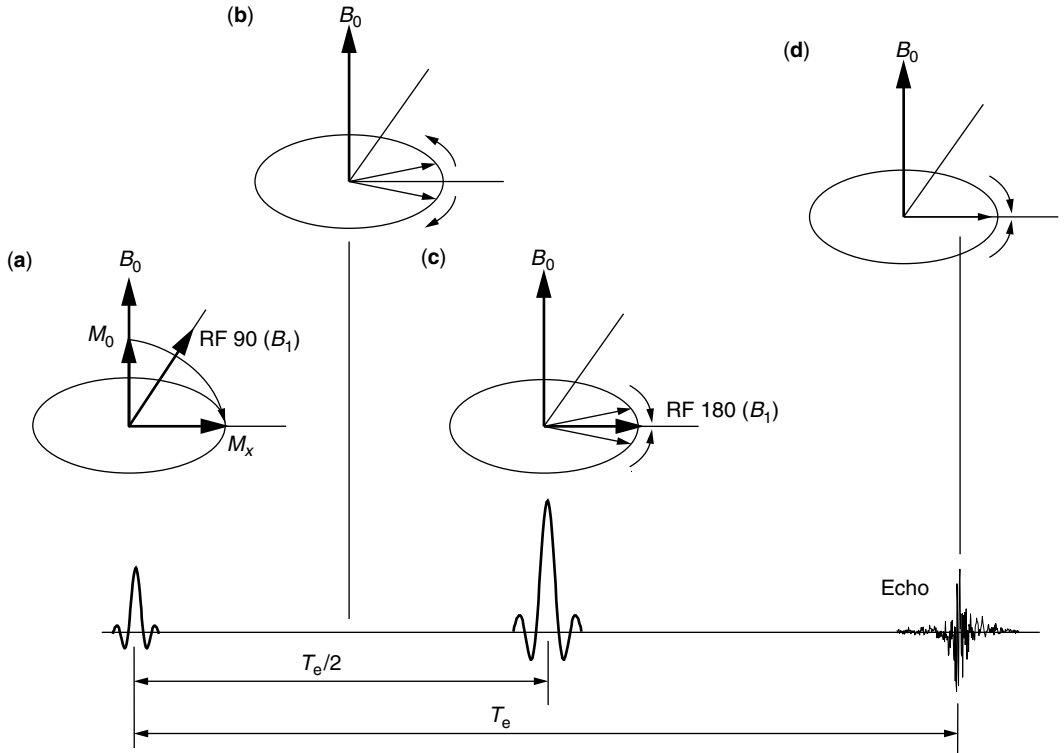


Figure 9. Echo formation in a spin-echo sequence. (**a**) Initial RF90 pulse rotates M_z into M_x. (**b**) Spins lose transverse coherence. (**c**) RF180 pulse reverses the direction of the diffusing spins. (**d**) Center of received signal coherence occurs at time T_e.

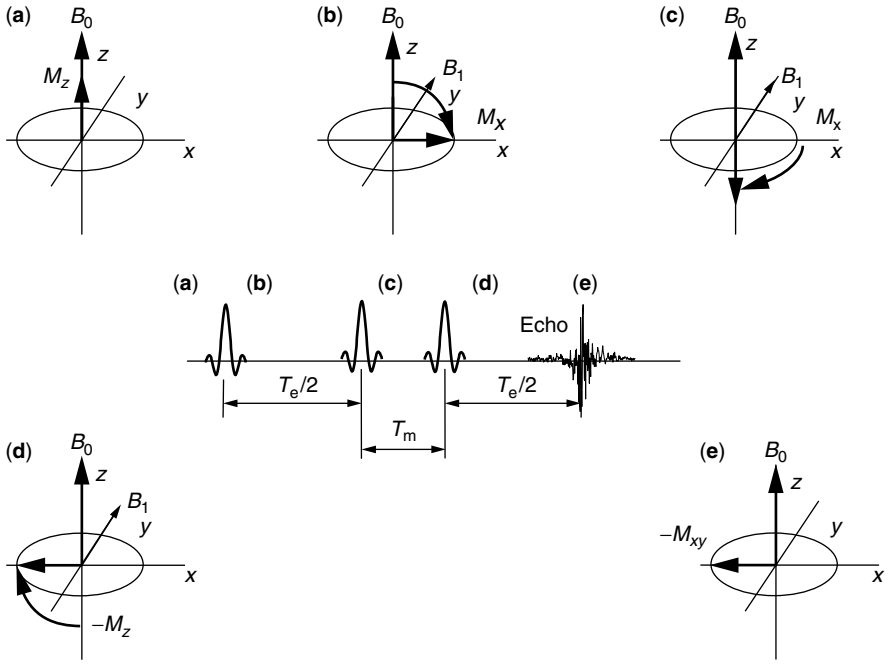


Figure 10. Echo formation in a Hahn echo or stimulated echo sequence. (**a**) Spins are initially aligned along B_0. (**b**) The first RF90 pulse rotates the spin vector to become M_x (**c**) A second RF90 pulse rotates M_x along $-z$ to become $-M_z$. The interval shown between (**c**) and (**d**) is longitudinal storage time and does not count against echo time because there is no transverse component to dephase. (**d**) The third and final RF90 pulse rotates $-M_z$ to become $-M_x$ where it can be read out.

The Stimulated Echo. An RF180 pulse can be used to refocus the spins, and as Hahn showed in his 1950 paper on the spin echo, a train of 90° RF pulses can produce a similar effect. If the RF 180 pulse in the spin-echo sequence is replaced by two RF90 pulses, the result will be as shown in Fig. 10. The M_z magnetization is first converted to the transverse plane. The second RF90 pulse converts the M_{xy} magnetization to $-M_z$, that is, the magnetization is now directed along $-z$. The third RF90 pulse then rotates the $-M_z$ magnetization back into the M_{xy} plane where it can be detected. The value of this sequence is that no $T2$ decay effect occurs in the period between the second and third RF pulses because there is no transverse component. This period can then be used to alter the spin system in some manner. One example of this is to add a water suppression pulse in this time when stimulated echoes are used for proton spectroscopy (47).

MR Parameters

Magnetic resonance is an extremely useful tool despite its limited sensitivity. The great utility of MR is derived from the many contrast mechanisms that are contained in the MR signal. All other modalities in diagnostic medical imaging have a two-dimensional image space. For example, in computed tomography (CT), the image dimensionality is space versus atomic density. In ultrasound, it is space versus acoustic reflectivity. In nuclear medicine, it is space versus density of radionucleide decays. In MR however, the contrast space is vast. A partial listing gives space versus spin density, $T1$, $T2$, $T2^*$, flow, diffusion, perfusion, magnetization transfer, and metabolite density. The challenge in using MR is selecting spin system preparative methods that present advantageous contrast. Sometimes these are combinations of the various basic contrasts, and at other times, they may be a single, pure contrast.

Spin Density. Spin density is the most like CT of all the MR contrast parameters. The spin density is simply the number of spins in the sample that can be detected. The observed spin density in medical imaging is always less than the actual spin density due to the fact that many spins are bound and lose signal before they can be observed.

T1 Relaxation. After the spins are tipped into the transverse plane, the spin vector along B_0 (called M_z) is depleted. If the spins are not tipped into the transverse plane again, the M_0 signal will regrow exponentially with a time constant $T1$:

$$M_z = M_0 \cdot (1 - e^{T_r/T1}), \tag{3}$$

where T_r is the time between repetitions of the excitation and $T1$ is the longitudinal relaxation rate. The M_z component, also called longitudinal magnetization, is the source of the signal for the experiment. If $M_z = 0$, the next repetition of the experiment will yield no signal because there will be no M_z to rotate into the transverse plane. The M_z vector is 99% of its equilibrium value M_0 when the T_r time is five times $T1$.

The mechanism for relaxation is the interaction between the excited spins and the stationary spins, referred to as the lattice. The amount of interaction with the lattice is the determining factor in $T1$. In biological systems, water molecules come into frequent contact with the lattice and thus have relatively short $T1$ times (<1 second). The $T1$ times for biological tissues tend to increase as the inverse square of the field strength (48). For simple isotropic motion, $T1$ is given by

$$\frac{1}{T1} \approx B\left\{\frac{\tau_b}{1+\omega^2\tau_b^2} + \frac{4\tau_b}{1+4\omega^2\tau_b^2}\right\} \tag{4}$$

where B is a measure of the interactive strength between the spin and the lattice,
ω is the resonant frequency, given by the Larmor equation, and
τ_b is the correlation time between the spin and the lattice.

In a gas, there is no lattice, so polarized gas molecules have $T1$ times of the order of days. In solids, the spins are relatively stationary and so do not experience frequent changes in contact with the lattice. Therefore, the $T1$ times in solids are also long. The $T1$ time of a sample increases as B_0 increases.

***T2** Relaxation.** The rate of decay of the M_{xy} signal after the initial excitation is $T2^*$ and can be appreciated in the graph of Fig. 7. This decay rate governs the amplitude of the signal in gradient echo experiments, despite the fact that the signal is sampled as a delayed echo, not an FID. The signal from the sample is a function of the delay from excitation, known as echo time, or T_e:

$$S(T_e) = S(0) \cdot e^{-T_e/T2^*} \tag{5}$$

The $T2^*$ time constant is a function of both the reversible losses due to magnetic field inhomogeneity and irreversible losses from diffusional motion $T2$ (49):

$$\frac{1}{T2^*} = \frac{1}{T2} + \gamma \Delta B_0 \tag{6}$$

Here the magnetic field inhomogeneity contribution is given by the term $\gamma \Delta B_0$.

T2 Relaxation

By adding the 180° B_1 pulse as previously noted, the reversible portion of the $T2^*$ decay can be removed by restoring it to the signal. Then, the signal decreases with time as

$$S(T_e) = S(0) \cdot e^{-T_e/T2} \tag{7}$$

Note that $T2^*$ is always less than or equal to $T2$.

Localization Principles and Techniques

Localizing the origin of the MR signal is essential in any system where information is desired from less than the entire object. In a high resolution NMR spectrometer, such as is used in chemistry labs, the sample is confined to a small tube and is nominally homogenous. The magnetic resonance of larger objects such as animals or humans requires that the signal be acquired from a subregion of the object. Localization techniques presently used fall into two categories. These are localization by B_1 field and by magnetic field gradient. In the first case, an RF coil smaller than the sample is used to acquire a signal only from the portion of the object close to the coil (50,51). This has been used in MR spectroscopy, but it is not a method for making MR images.

The use of magnetic field gradients permits localization in three dimensions. The gradient is created by large diameter coils that are wound as Maxwell pairs (52,53). By passing a current through the coil pair, a magnetic field, directed along the main magnetic field, is created whose intensity is a function of position. The coil geometry is such that the physical center of the coil experiences a field intensity of zero, and the intensity of the field away from the coil center is a linear function of distance along one axis. The direction of the field created by the gradient coil is parallel to the main field, as shown in Fig. 11. Only alterations of the main magnetic field result in a change in the resonant frequency of the spin at that location. All three gradients, X, Y, and Z produce changes in the main field which vary in intensity in the x, y or z directions, respectively.

The first method used for localization was designated the sensitive point method (54). This was accomplished by simultaneously applying gradients in three orthogonal dimensions. By adjusting the strength of the gradients, only one small volume in space remains on resonance. After collecting the signal from the spins at that location, the gradients are adjusted to shift the resonant point, and the signal from the newly resonant spins is collected. The speed with which the voxel can be shifted is limited by the $T1$ times of the sample. A single image that has a resolution of 256×256 and a T_r time of 1 second requires 65,536 seconds (more than 18 hours) to acquire.

A second method, called line scanning, uses two gradients to create a right rectangular prism of resonant spins (55,56). Spatial localization along the length of the prism was accomplished by frequency encoding where the origin of the signal was determined by the resonant

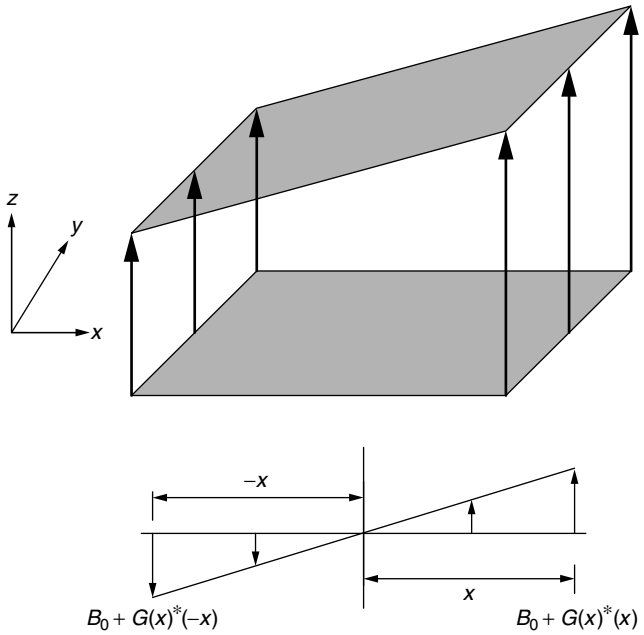


Figure 11. Magnetic field gradient imposed in the x direction. Notice that the resonant frequency of the spin changes linearly with the position in the x direction. A common misconception is that the gradient produces a field that is pointed in the direction of the gradient. This is not correct. The gradient field direction is along z, regardless of the direction of the gradient. Only modifications of the B_0 field (in the z direction) affect spin precession.

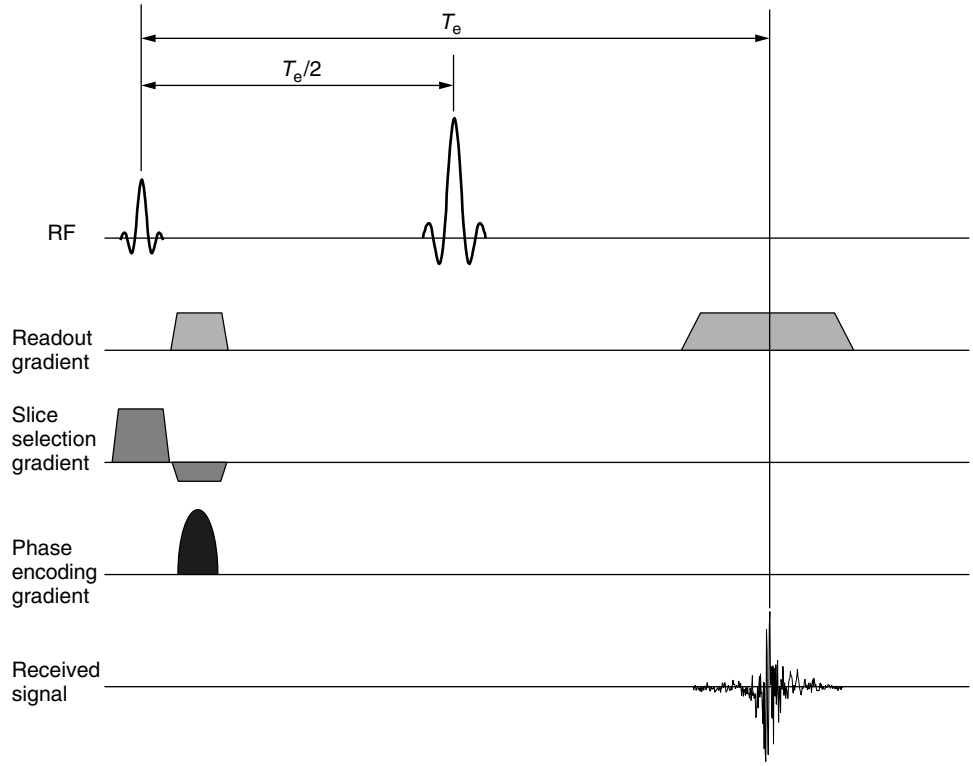


Figure 12. A complete spin-echo sequence that uses localization gradients to image a single slice. Only a single value of a phase encoding gradient is shown. To form an image, the sequence must be repeated using a unique value of phase encoding for each line of data in the two-dimensional k-space matrix.

frequency of the spin. The gain in efficiency for the line scanning method as opposed to the sensitive point method is due to the ability to collect data in a single signal from all points along the line. Signal locations were then distinguished by using the Fourier transform.

The third method, which is preponderant, is the "spin-warp" (57) or Fourier transform imaging method. This method enjoys significant improvements in efficiency over the point and line scan methods due to the way in which the image is sampled. In contrast to the sensitive point method where a single data point samples only the spins at that physical location, data points in Fourier transform imaging sample a phase and frequency for the entire slice. In the two-dimensional version, a slice is first excited by a shaped RF pulse which is transmitted while a gradient is imposed, this is known as the slice selection gradient (Fig. 12). If the slice selection gradient were missing, the entire object would be excited. To image a slice, a B_1 pulse is transmitted during the time that a gradient is applied. The bandwidth of the B_1 pulse is limited typically by using a modulation envelope of the form $\sin(x)/x$ (this is called a "sinc" function, and it contains only a limited band of frequencies). Therefore, only the spins within a slab will be excited. The slab thickness is a function of both the amplitude of the slice gradient and the bandwidth of the B_1 pulse, either of which may be changed to change the slab thickness. In two-dimensional imaging, no further localization is done across the slice. In the sinc envelope RF pulse, phase dispersion is created by the slice selection process that causes a complete loss of signal. Therefore, a slice refocusing gradient is applied that cancels the phase accumulated across the slice. The position of the slice in the direction of the slice selection gradient can be moved by changing the frequency of the B_1 pulse. The slice selection direction is entirely arbitrary (58). The normal vector of the slice may be aligned along any of the principal coordinate axes, X, Y, or Z. Additionally, the slice can be oriented obliquely by sharing the slice selection pulse between any or all of the gradients. Multiple slices can than be acquired in the same sequence. This is important because typically a number of slices must be acquired to image the entire object (59).

Signals from a group of spins distributed along the frequency encoding gradient can be distinguished by their resonant frequencies. As for line scanning, the individual resonant frequencies can be extracted from a single signal by using the Fourier transform. To localize in the second transverse dimension, the phases of the spins are changed as a function of the distance from the center. This is accomplished by temporarily applying a gradient that briefly changes the frequency of the spins and then returns them to resonance. The result is that the phase of the spin is a function of the distance from the isocenter in the direction of the phase encoding axis.

If a thicker slab is excited, as shown in Fig. 13, an additional direction of localization can be added to the sequence. In the third dimension, an additional phase encoding waveform is used. Reconstruction then creates slices from the slab encoding direction.

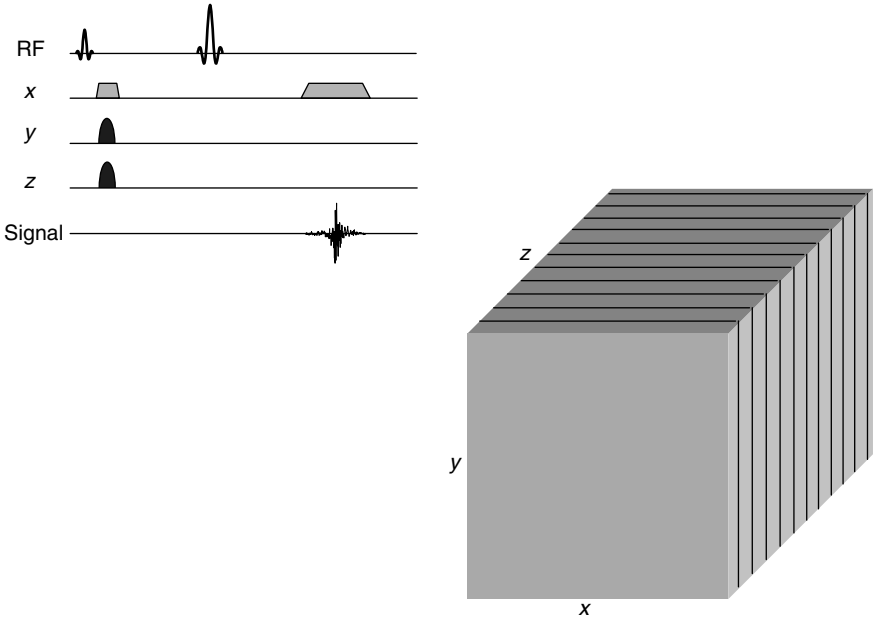


Figure 13. A three-dimensional spin-echo sequence where phase encoding is substituted for the slice selection gradient. Multiple slices are then made, as shown, by using one more Fourier transform on the collected three dimensional k-space data set.

A fourth method is similar to computed tomography. If the spin-warp method is used without the phase encoding gradient, a projection across the image is created. If the readout gradient is shared between the two axes which are not the slice selection axis, a group of projections from multiple angles around the object is acquired. An image is then created using the same methods as in computed tomography (60,61).

Imaging Mathematics. The amplitudes of the gradients used in phase and frequency encoding can be calculated by the need to prevent aliasing of the image. Aliasing results from the sampled nature of the data in both phase and frequency encoding directions. An image that is undersampled in the phase encoding direction would appear as in Fig. 14. To prevent aliasing, sampling step size must be small enough to satisfy Nyquist's criterion (62). In the frequency encoding direction, the step size is

$$\Delta k_{\mathrm{f}} = \frac{\gamma}{2\pi} G_{\mathrm{f}} \Delta\, t, \tag{8}$$

and the field of view (FOV) in the frequency encoding direction is

$$\mathrm{FOV}_{\mathrm{f}} = \frac{1}{\Delta k_{\mathrm{f}}} \tag{9}$$

In the phase encoding direction, the step size is

$$\Delta k_{\mathrm{p}} = \frac{\gamma}{2\pi} G_{\mathrm{p}} t_{\mathrm{p}}, \tag{10}$$

and the FOV in the phase encoding direction is

$$\mathrm{FOV}_{\mathrm{p}} = \frac{1}{\Delta k_{\mathrm{p}}}, \tag{11}$$

where Δk is the step size, measured in cm^{-1}. Image resolution is determined by the maximum value of k.

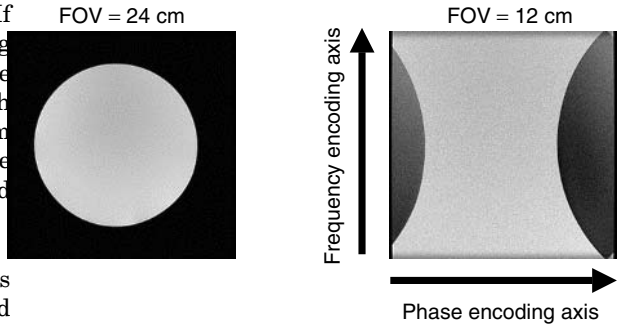


Figure 14. Aliasing in a single dimension. From the left image, the field of view for the right image was halved. The frequency encoding direction does not show any aliasing due to the use of a hypersharp digital filter. The phase encoding direction shows aliasing due to insufficient sampling.

This can be understood intuitively because the higher spatial frequencies are sampled farther from the origin. The physical field of view (FOV) of the image is determined by the density of samples. Because MR data are sampled data, they follow Nyquist's criterion and therefore must be sampled to a density sufficient to prevent aliasing. The requirements of k-space coverage and density of sampling have a strong influence on the trade-offs made in MRI, and these will be further discussed in the section on trade-offs. The 2D acquisition then lasts

$$T = T_{\mathrm{r}} \times N_{\mathrm{ph}} \times N_{\mathrm{EX}}. \tag{12}$$

The 3D acquisition lasts

$$T = T_{\mathrm{r}} \times N_{\mathrm{slph}} \times N_{\mathrm{ph}} \times N_{\mathrm{EX}}, \tag{13}$$

where T_{r} is sequence repetition time, N_{ph} is the number of phase encoding steps or samples, N_{slph} is the number of

phase encoding steps in the slab direction, and N_{EX} is the number of sequence averages.

Sampling in the frequency encoding direction is performed in the presence of a band-pass frequency filter to prevent noise from frequencies outside the sampled bandwidth from being aliased into the image. If the frequency encoding gradient amplitudes and sampling times are adjusted appropriately, no image aliasing will occur, but the noise is white and exists outside the imaging space. It is this noise which can be aliased in and decreases the signal-to-noise ratio of the image (63).

Image resolution is defined as the field of view divided by the number of samples in that direction. This is true as long as the point-spread function (PSF) is not broadened by signal decay in that direction. The point-spread function is the Fourier transform of the product of the sampling window and any weighting functions due to uneven sample weighting in k-space. To understand image resolution, consider the image of an object that is just one pixel in width and height. The resulting image can be predicted by the convolution of the object shape by the point-spread function. If an object were to be sampled to an infinite distance in k-space, the point-spread function would be a delta function because the sampling window would be at a constant level (whose Fourier transform is a delta function). Real-world imaging samples data over a finite range of samples. In this case, the sample window is a rectangular function whose transform is a sinc pulse. The wider the sample window, the closer the PSF approaches a delta function. If however, the weighting of the samples is not constant, the PSF will be broadened. This can be the case for some fast imaging methods (64).

Signal-to-Noise Ratio

The Signal-to-Noise Ratio (SNR) in MR imaging is the limiting factor in the choice of experimental parameters used in the imaging sequence. The SNR formula is

$$\mathrm{SNR} \propto \frac{NV\sqrt{T_s}(B_0)^k(e^{-T_e/T2})(1-e^{-T_r/T1})}{\sqrt{\mathrm{bw}}} \qquad (14)$$

where N is the number of spins in the voxel,
V is the voxel volume,
B_0 is the main magnetic field strength,
T_s is the total sampling time of the experiment,
bw is the bandwidth of the receiver filter,
and k is a factor that is one for conductive samples and 7/4 for nonconductive samples (discussed further in the instrumentation section) (65).

Voxel volume is a product of the voxel dimensions in the plane and the slice thickness. As such, it can be considered an indicator of resolution. As resolution improves, the volume element V decreases and decreases the SNR of the image (for constant imaging time). Imaging experiments which create a $T2$ or $T1$ based contrast will operate at less than the maximum SNR because T_e and T_r times used are less than optimum for signal preservation. Increases in imaging SNR can be obtained by increasing the B_0 field and or the total sampling time (usually by signal averaging) or by decreasing the receiver bandwidth.

IMAGE PROCESSING AND *K*-SPACE

The ubiquity of Fourier transform imaging in MR is due to the high SNR/unit time of the technique. This SNR efficiency results from collecting data from the entire object for each repetition of the sequence. This occurs because the gradients act to perform Fourier transforms of the object into k-space data. Then, the image reconstruction process Fourier transforms the k-space data into the image (Fig. 15). k-space is the complex matrix where acquired data are stored. The dimensions of k-space are wave number or inverse length. The structure of k-space has the lowest spatial frequencies in the center, whereas the outer edges contain the highest spatial frequencies. For common objects, such as human subjects, the center of k-space holds most of the image energy, and the edges contain the detail. It is important to remember that a data point in k-space has no direct remapping into the final image but represents only the signal from a component of spatial frequency at a particular phase. k-space has another important property that results because the data are a Fourier transform (one created by the encoding gradients) of a real object. The transform is Hermitian across the center of k-space that is, the complex datum at a point x, y, call it $(a + ib)$ is the complex conjugate of the datum at $-x, -y$ which is $(a - ib)$ where a is the real part and b the imaginary. Thus, only half of the k-space data must be collected to make an image because the other half of k-space can be recreated from the acquired half (66). The time between excitation and the echo (T_e time) can be minimized by acquiring only the right half of k-space. Total acquisition time can be cut approximately in half by acquiring only the top half of k-space. Note that flow and motion can cause k-space data to depart from the Hermitian property.

In two-dimensional imaging, k-space is a two-dimensional matrix. Reconstruction is accomplished by first Fourier transforming along one dimension of the data set. Then, a second Fourier transform is performed along the other axis to yield an image. The Fourier transform

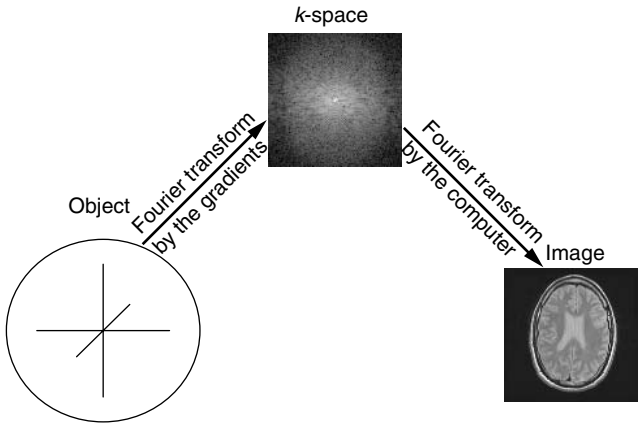


Figure 15. The concept underlying Fourier transform imaging, where the object is sampled through a single wire (the wire to the receiver) by encoding object data through the Fourier transform effect of the gradients into a single signal. The image is then created by a computer-based Fourier transform.

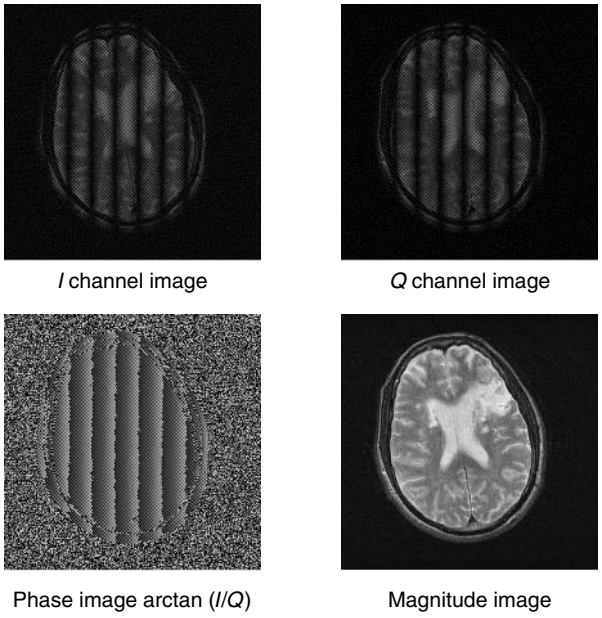


Figure 16. Images obtainable from the complex data matrix which results from Fourier transform reconstruction. By back-calculating the image in the image, it can be removed to make the real image readable. In practice, only the magnitude image is used.

of the complex raw data produces a complex image. Four images can be created from the complex data set. These are the real (I) and imaginary (Q) images, the phase image [arctan(I/Q)] and the magnitude image. Each of these images is shown in Fig. 16. Most images are displayed as magnitude, however, additional information is available about the sign of the signal if the real channel is used, provided that the phase variation is removed (67). In 3-D imaging, a transform along the third axis is performed. No third transformation is required along the third dimension in 2-D imaging because the slices are acquired in separate k-space matrices. The order in which Fourier transforms are performed is arbitrary because each encoding is orthogonal from the others. The images produced reflect any nonlinearities of the physical gradient fields. Because the imaging process samples to a linear grid, nonlinearities will show up as image stretching or compression (68). Knowledge of the actual field distribution of the gradient allows mathematical warping of the image to compensate partially for the physical imperfections of the gradients.

ENDOGENOUS IMAGE CONTRAST

The simplest mechanism of image contrast is spin density. The number of spins in a voxel is a primary determinant of the signal available from the voxel [see Eq. (14)]. If this were the only source of image contrast, however, then MRI would be little different from computed tomography where image contrast is a function of the tissue density.

The most significant discovery in NMR of intact human tissues was that different tissues had different relaxation rates (69). Further, pathological conditions caused the tissue relaxation rates $T1$ and $T2$ to change from those of normal tissue (70). These two facts have formed the basis of most MRI sequences. Relaxation-based contrast mechanisms include $T1$ and $T2$ weighted methods, as well as magnetization transfer (MT) and time-of-flight angiography. To use these methods fully, tissue relaxation rates must be known approximately to allow choosing sequence timing parameters appropriately. The choice of parameters depends on the materials involved and also on the strength of the main magnetic field. For intact animals, the range of $T1$ and $T2$ values can be of the order of milliseconds to seconds, most strongly dependent on tissue type (71). Typically, $T2$ values change only slightly as field strength changes whereas $T1$ and $T2^*$ values can change substantially. $T2^*$ values change due to the presence of paramagnetic (blood) and ferromagnetic substances such as ferritin (72) which causes transverse magnetization to dephase more rapidly as B_0 increases. $T1$ values change as field strength B_0 changes because the loss of longitudinal magnetization to the lattice is determined by the ratio of the correlation time of the molecule and the Larmor frequency of the spins. When the Larmor frequency of the spins is equal to the inverse of the correlation time (between the free spin and the surrounding tissue matrix), the relaxation or spin loss is most efficient. Sequence parameters are thus adjusted to achieve the maximum difference between normal (for anatomical visualization) and pathological tissue states (for detection of disease), not to achieve maximum SNR. In Fig. 17a, MR signal intensity is plotted against echo time for some of the tissues in the brain. Echo times that yield the maximum $T2$ weighted image contrast are not zero, where signal-to-noise is maximum. It is instead, as shown in the associated graph that shows the difference in the signals, in an intermediate range of echo times. For $T1$ contrast, a typical set of relaxation curves is shown in Fig. 17b where the maximum contrast for a simple spin-echo sequence is plotted in the same figure. Note that other methods of spin preparation such as inversion can greatly alter the contrast between tissues (73,74). Image contrast based on the combination of T_e and T_r for human tissues is shown in Fig. 18. Typical ranges for T_e and T_r are 5 to 200 ms for T_e and 300 to 4,000 ms for T_R. The mixed weighting condition (long T_e, short T_R) is not used for patients.

OTHER CONTRAST MECHANISMS

In addition to spin density, and $T1$, $T2$, and $T2^*$ contrast, spin systems, especially biological, have other contrast mechanisms that can be used to image aspects of system or organ function. These include the bulk flow of spins (usually blood), diffusional motion of spins, and chemical shift. The bulk flow of spins can be imaged by two methods. The first is a time-of-flight method where a short T_r sequence is used to suppress nonmoving spins by reducing the longitudinal magnetization (M_z) to near zero (75). Spins that move into the slice have full M_z (because they have not experienced an effective excitation) and thus produce strong signals. The second method is a phase contrast method where bipolar gradient lobes are played out along one gradient axis (76). Spins that do not

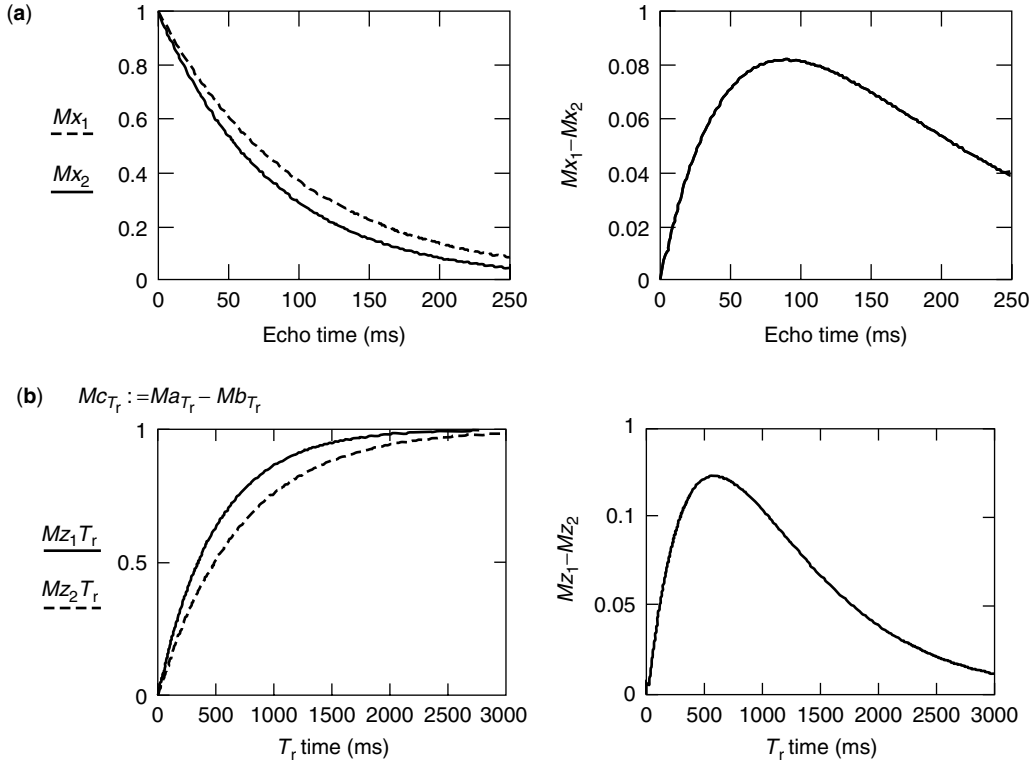


Figure 17. (**a**) Different spin populations that have different T2 relaxation rates show maximum contrast at a T_e time that is not zero, where signal would be maximum. (**b**) Two spin populations different in $T1$ relaxation, showing relaxation rates and the difference in signal versus T_r time. Again, the maximum contrast is not obtained where the signal is maximum (5 × T1).

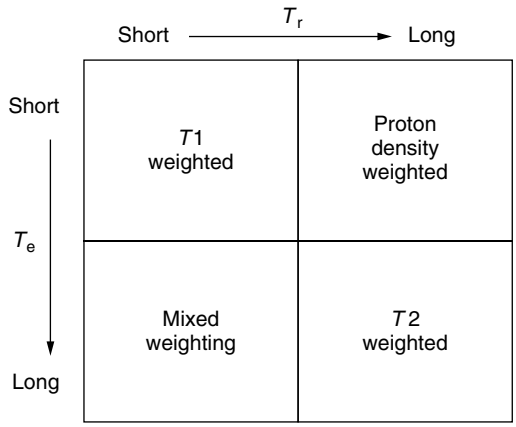


Figure 18. A matrix that shows the effect of T_e and T_r time on image contrast in spin-echo imaging. At long T_e and short T_r, mixed contrast weighting is obtained and therefore is not used in MRI of patients.

move have equal amounts of phase added and subtracted, resulting in a net zero phase change. The sequence is then repeated, and the polarity of the phase encoding lobes is reversed. Subtraction of the two images results in an image of only the moving spins. The advantage of this method is that the velocity is encoded and thus can be measured (unlike time of flight). However, an image that is sensitized to flow in all three axes requires six repetitions of the sequence. It is possible to reduce this number to four repetitions by suitable methods, but it is still a more time-consuming method. A comparison of the methods is shown in Fig. 19.

Spin diffusion can indicate the microstructure of a material. In biological systems, it is used to reveal defects in cellular membrane integrity and cellular swelling (77–81). The method is implemented by using very large bipolar gradients that are arranged on either side of an RF180 pulse (25). Then, signal loss is a function of the diffusion of the spins (D) and the amplitude (g), width (δ) and spacing (Δ) of the diffusion weighted gradients:

$$\frac{S}{S_0} = e^{-\gamma^2 g^2 \delta^2 (\Delta - (\delta/3))D}. \tag{15}$$

All of the terms in the exponent except D are combined and referred to as b. The b value of a diffusion weighted pulse sequence is used as a measure of the degree of diffusion weighting in the pulse sequence. At high b values, even spins that diffuse slowly move a sufficient distance, thus losing signal. The areas of the gradients used to create the diffusional motion sensitivity are quite large, and the amplitudes of the motions are quite small. In living biological systems, minor motions attendant to respiration and other physiological processes require that the imaging experiment be quite short. Typically, fast echo-planar imaging (EPI) methods are used which acquire all of k-space in a single excitation. This will be explained in the section on fast imaging (Fig. 20) (82).

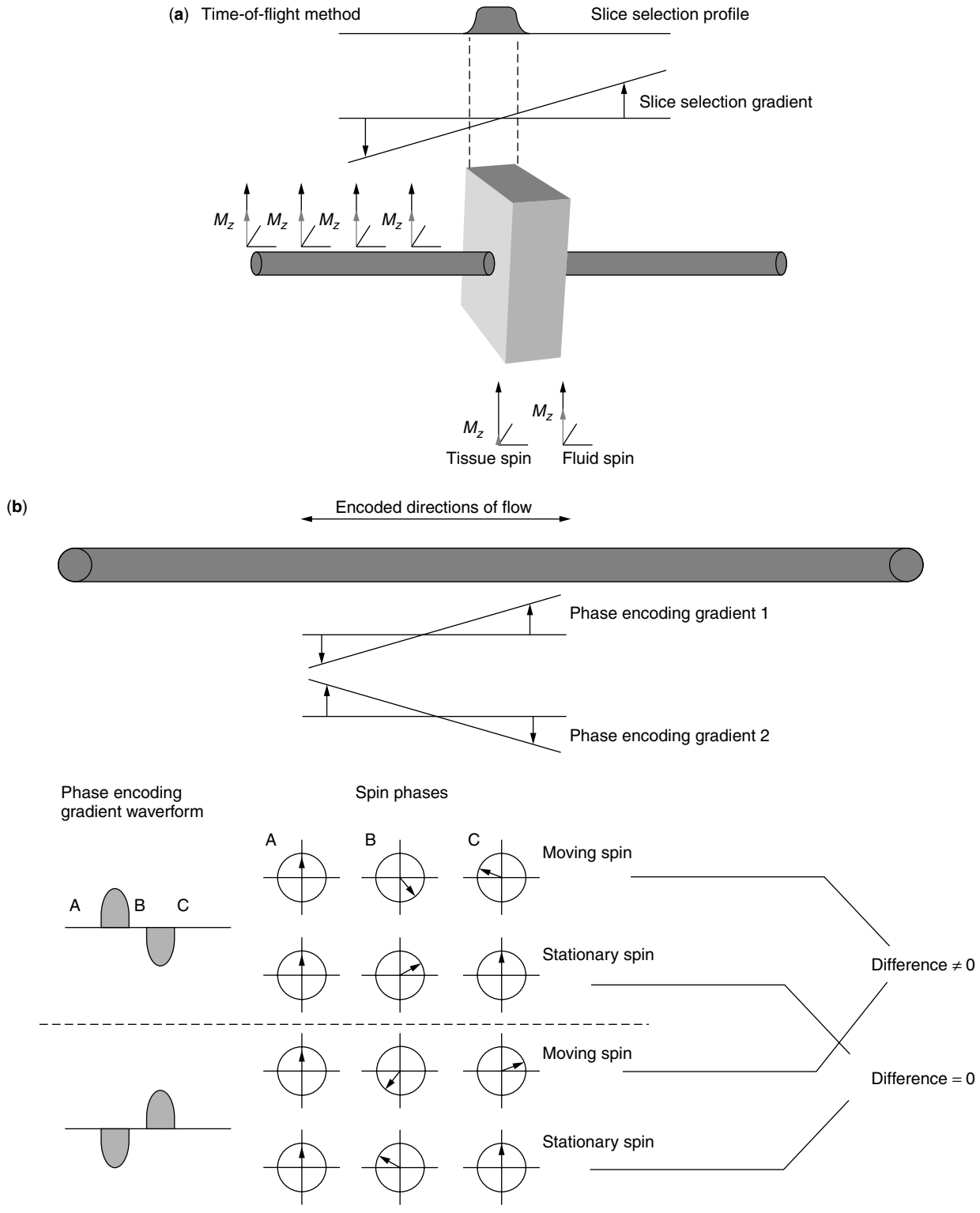


Figure 19. (**a**) Mechanics of the time-of-flight method of flow imaging. Spins in the slice that do not move have small M_z due to the short T_r of the pulse sequence. These stationary spins produce very little signal in the image. Flowing spins, however, have not been subject to B_1 excitation and have $M_z = M_0$. The flowing spin in the slice then gives a very large signal upon excitation. (**b**) Mechanics of the phase contrast method of flow imaging. Spins that move in the direction of the encoding gradient acquire a phase as a function of the distance traveled in the time between encoding lobes. Stationary spins acquire and lose the same amount of phase from each gradient lobe pair.

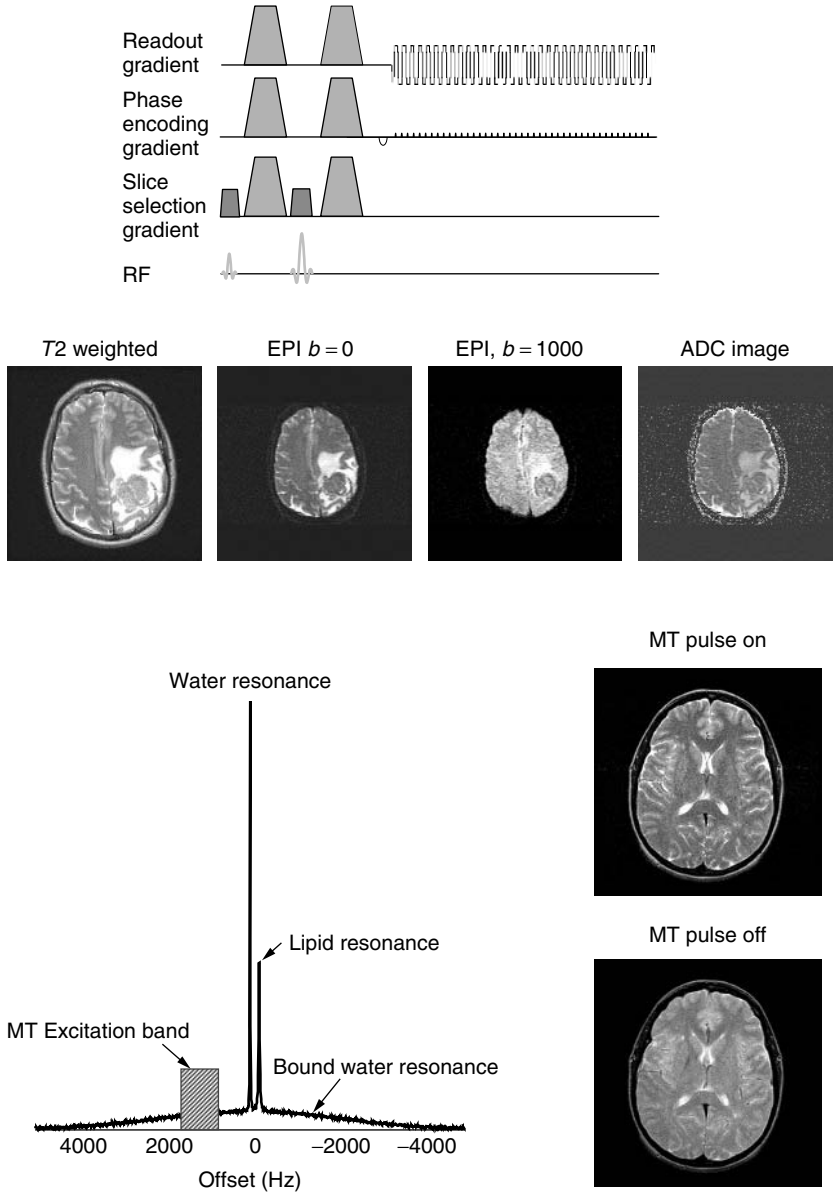


Figure 20. Diffusion imaging showing the diffusion encoding gradients on all three axes at once (for the simplicity of the diagram). In practice, each diffusion direction must be collected separately. The diffusion weighted image is made by collecting at images that have at least two different diffusion weightings and calculating the Apparent Diffusion Coefficient (ADC) image by fitting each pixel intensity to Eq. (15).

Figure 21. Magnetization transfer imaging, showing the saturation pulse relative to the imaged mobile protons and the bound background spins.

Magnetization transfer imaging is a method of modifying $T1$ relaxation of a sample by eliminating one of the pathways by which mobile spins lose energy to the lattice (83). In magnetization transfer imaging, an RF pulse is used that is significantly off-resonance. In imaging biological systems, mobile water protons communicate with the protons bound to macromolecules of the background. The macromolecule associated protons do not tumble freely which means that they have very short $T2$ times that render them invisible. The short $T2$ time also implies that the bandwidth of the resonance is very broad. A strong RF pulse delivered off-resonance will excite these spins without interfering with the mobile proton signal. The excited off-resonance protons become invisible to the mobile protons and force the $T1$ relaxation rate to increase. Not all biological systems demonstrate this behavior. Brain tissue and blood do, but adipose (fatty) tissue does not. The process of magnetization transfer imaging is shown in Fig. 21. An image of the magnetization transfer effect requires making two images, one for the pulse on and the other for the saturation pulse disabled. The subtraction of the two images gives an image of the magnetization transfer effect. Magnetization transfer is now used to investigate diseases of the myelin sheath around neurons (84,85).

FAST IMAGING

In many systems, image acquisition time must be minimized due to the changing nature of the object. In biological systems, this is typified by cardiac, respiratory, and gross patient motion. For process control systems such as flowing fluids or production line systems, the object must be sampled on timescales short compared to the time available for observation. Imaging that uses pure $T1$, $T2$, or proton density contrasts requires constant T_r periods

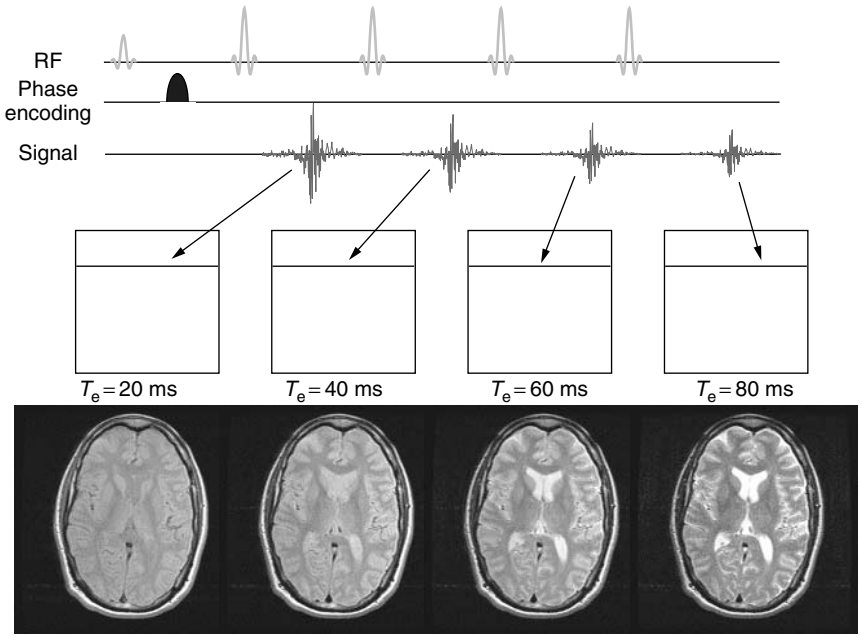


Figure 22. Conventional multiple echo spin-echo sequence where one phase encoding is used for all echoes in a sequence. Then, a four echo sequence creates four separate images that have four distinct T_e times.

for each line of k-space data. Therefore, the minimum acquisition time is $T_r \times N_p$ where N_p is the number of phase encoding steps. In biological systems where $T1$ values are of the order of 1 second, total acquisition time is of the order of two minutes. Several approaches exist to minimize total acquisition time. Each, however, requires trade-offs.

Fractional *k*-Space. The Hermitian nature of k-space permits concluding acquisition after half of the lines of k-space has been acquired. This method decreases total scan time by almost 50%. The trade-off is decreased signal-to-noise because the SNR is proportional to the square root of total acquisition time. Additionally, artifacts may arise from areas in the sample where the magnetic field changes abruptly, such as interfaces between tissue and air.

Gradient Echo

The gradient echo sequence can operate at short T_r times and decrease acquisition time from minutes to seconds (86). The use of a both a short T_r and the required short T_e time results in a mixed $T1$ and $T2^*$ weighting. The $T2^*$ rather than $T2$ weighting is due to the lack of an RF180 pulse. In some instances, $T2^*$ weighting is highly desirable, as when searching for pathology that may be associated with the accumulation of ferromagnetic material such as hemorrhage from trauma (87).

Fast Spin Echo

In a spin-echo pulse sequence, the RF180 pulse may be repeated to refocus the magnetization at different times. Each of these echoes is then stored in its own k-space array to form separate images, as shown in Fig. 22. If different phase encodings are applied to each of the echoes and the resulting data are stored in the same k-space, acquisition time will decrease as the inverse of the number of echoes (RARE) (64). The image contrast is determined principally by the echo time of the acquisitions near the center of the k-space (Fig. 23). The number of echoes can be extended to equal the total number of phase encodings. By this method, the entire image may be acquired in a single T_r time (88). The utility of fast spin echo depends on sufficiently long $T2$ times in the object. The k-space data are weighted by the $T2$ decay of the spins as shown in Fig. 24. k-space weighting causes blurring which worsens as $T2$ decreases. The $T2$s of the spins to be imaged dictate the maximum number of echoes (the echo train length).

Driven Equilibrium

In a conventional $T2$ weighted spin-echo sequence, the echo time T_e is long to remove signal contributions from spins that have short $T2$ times, and the repetition time T_r is long to prevent $T1$ weighting from mixing in with the contrast (89). To eliminate the need to wait for restoring longitudinal magnetization through $T1$ relaxation, the sequence can be repeated backward, as shown in Fig. 25. The use of driven equilibrium does not restore lost transverse magnetization ($T2$ losses) and may create artifacts.

Echo Planar

In a manner analogous to that of fast spin echo, multiple echoes may be acquired in a gradient echo sequence (Fig. 26) simply by inverting the amplitude of successive readout gradient pulses (90).

MAGNETIZATION PREPARATION

Gradient Echo Associated Methods

The lack of $T1$ contrast can be partially offset by using RF spoiling. RF spoiling varies the phase of

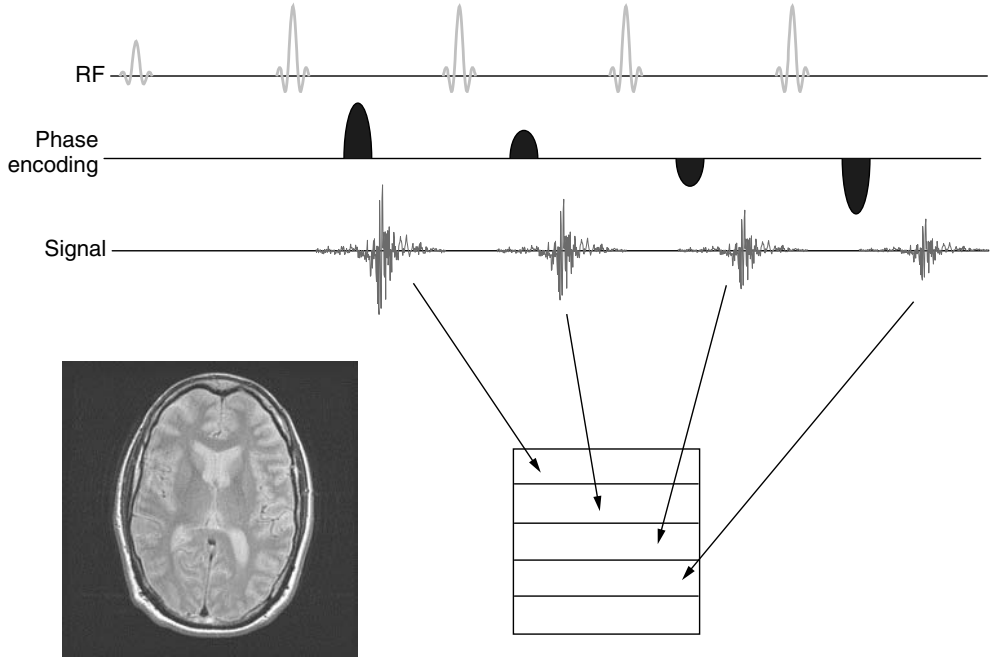


Figure 23. A fast spin-echo sequence whose echo timing is comparable to that of the conventional spin echo. Now, each echo has a different value of phase encoding applied and a single k-space matrix is filled in one-fourth the time of the conventional spin echo acquisition.

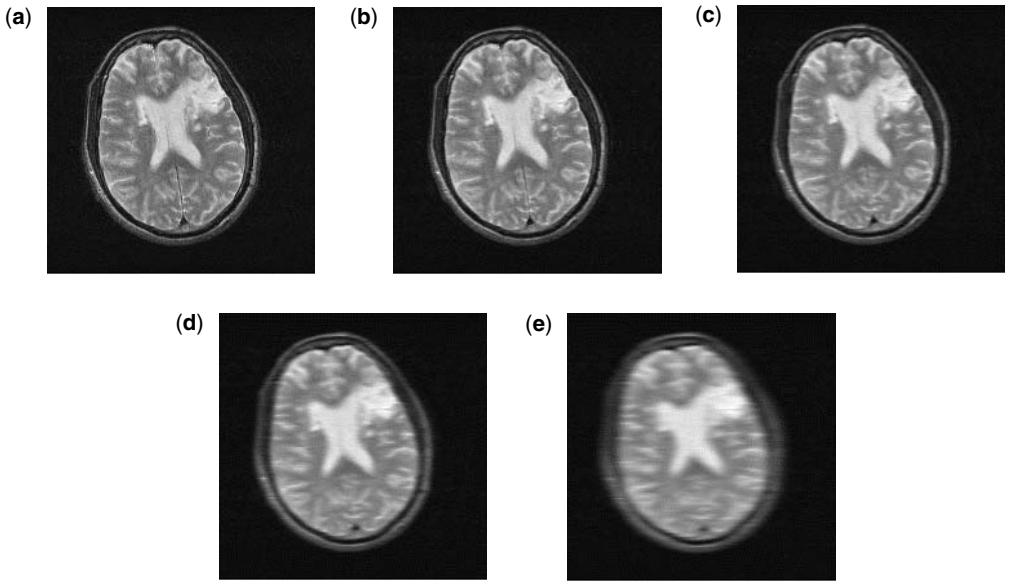


Figure 24. A fast spin-echo image data set where the PSF is increasingly broadened by filtering before reconstruction to emulate the effect of $T2$-related blurring on a fast spin-echo image. (**a**) No degradation, (**b**)–(**e**) Increasing degradation.

the RF pulses randomly between T_r periods to prevent the coherent interference of echoes from successive T_r periods (91).

Inversion Recovery

Most imaging methods start by nutating the longitudinal magnetization between 0 and 90°, but the spins can be nutated 180°. Immediately after such a pulse, the magnetization is directed along $-z$. Then, the path of relaxation is up along z, and no transverse component is being generated. A family of curves for a typical biological system is shown in Fig. 27. The magnetization so created can then be sampled using a spin-echo or gradient echo sequence. The image contrast is largely determined by the time between the inversion and the RF90 pulse. The

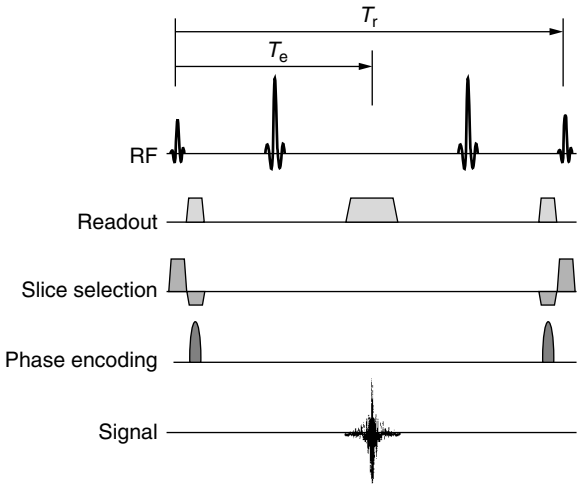


Figure 25. Driven equilibrium spin-echo sequence showing the sequence repeated backward after the echo time T_e. The phases of the RF pulses in the second half of the sequence are opposite those from the first part. If the initial phases of the RF90 and RF180 are $+x$ and $+y$, then the second RF180 must be $-y$ and the second RF90 must be $-x$.

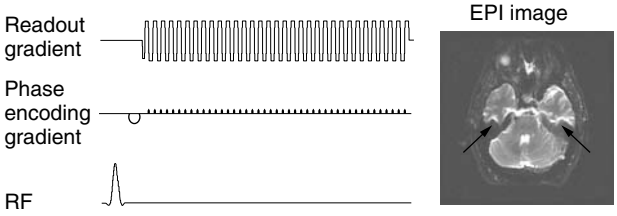


Figure 26. Echo planar imaging (EPI) pulse sequence and a typical image. Note the distortion at the air/tissue interface in the temporal lobes (arrows).

inversion method of spin preparation has three important features. The first is that the polarity of the sampled signals may be both positive and negative, depending on which side of the transverse plane the longitudinal magnetization was just before excitation. Most commercial MR imaging systems generate magnitude images from the complex Fourier transformed data, thus displaying only a positive image. However, it is possible to make real images that include signal polarity. The second important feature of inversion recovery is the ability to sample the inverted magnetization when one component has $M_z = 0$. An image will then be produced that has no contribution from spins sampled at a time equal to $0.693 \times T1$ after inversion. This has been used extensively to eliminate the signal from fat tissues. The third important feature is the ability to generate larger contrasts between spins of similar $T1$ than of $T1$ weighted spin-echo imaging (92,93).

Chemical Shift Saturation

The resonant frequency of a spin is a function of both the magnetic field in which it is immersed and the local magnetic fields it experiences. One of these field sources is the molecule in which it is bound. Hydrogen nuclei in water resonate 3.5 ppm downfield from hydrogen nuclei bound in lipid methylene moieties. Although small compared to the strength of the main magnetic field, the frequency difference is large enough to permit excitation of only one moiety. In practice, a single species may be excited before the imaging sequence, and its signal made incoherent by a strong gradient pulse that renders it invisible to the subsequent pulse sequence (94,95).

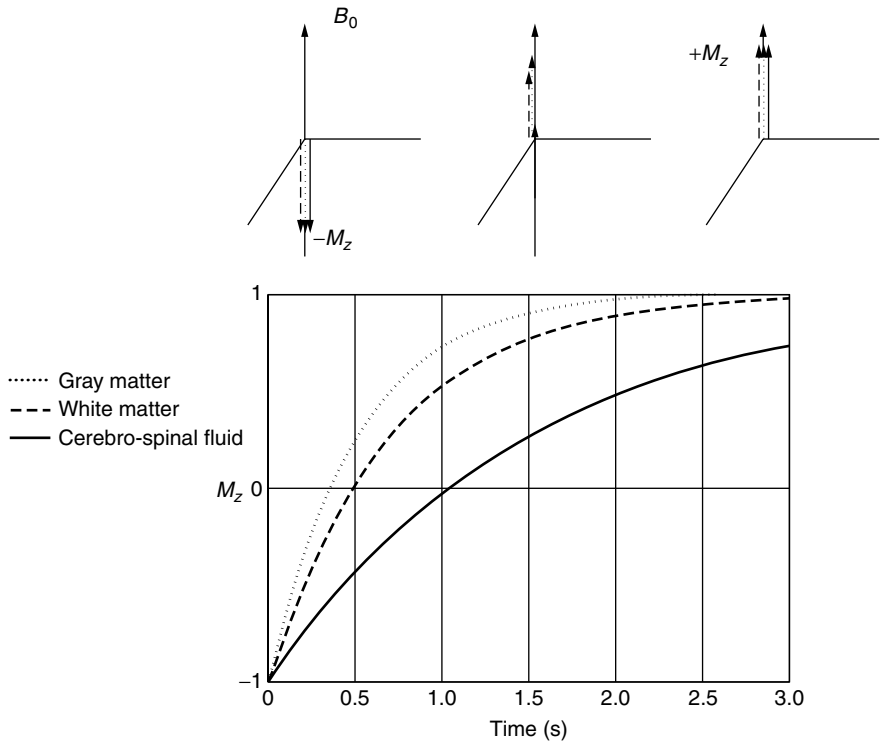


Figure 27. Longitudinal magnetization recovery after inversion versus inversion time for tissues in the brain. Note that M_z for each of the tissues crosses zero at a different inversion time.

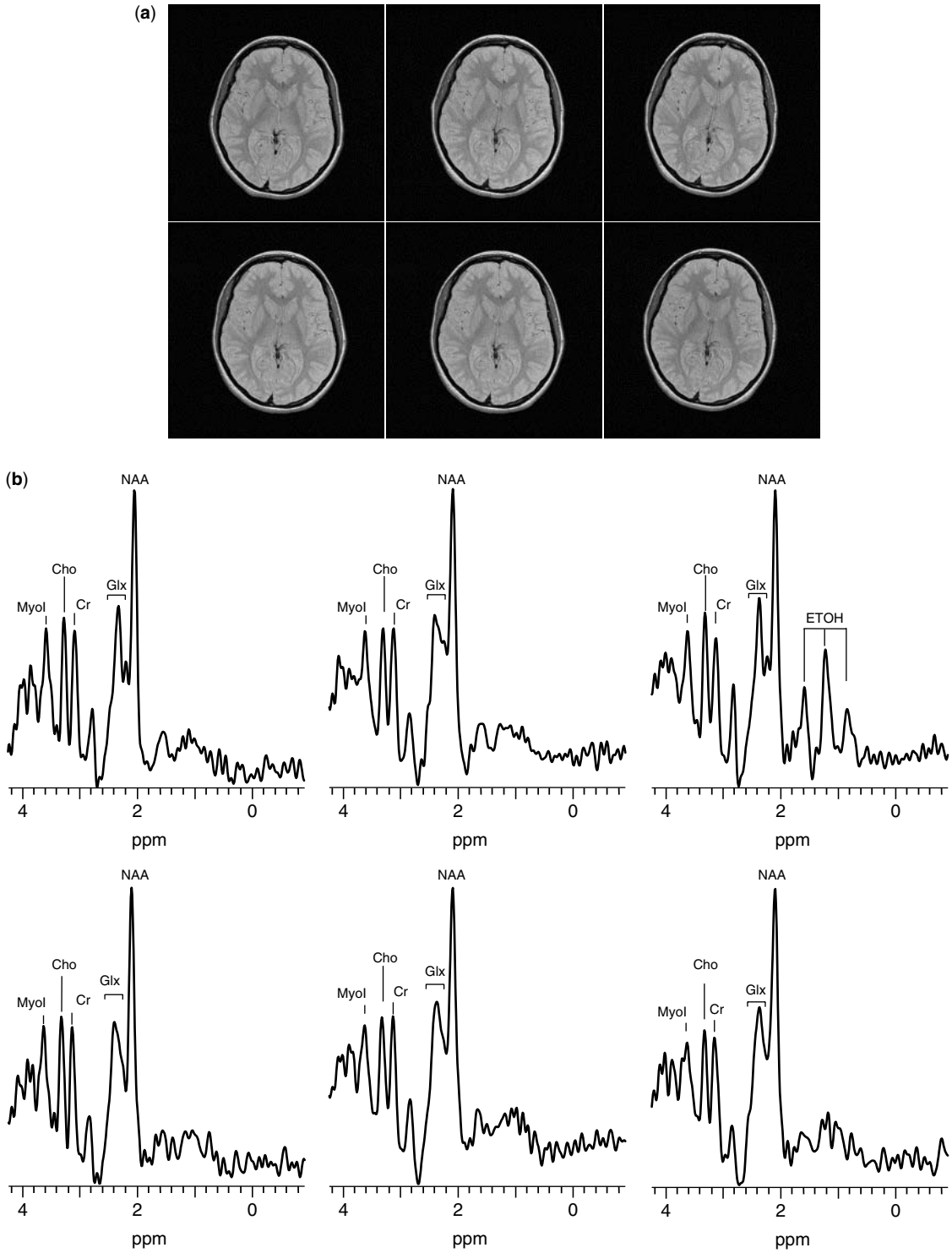


Figure 28. Proton density weighted, fast spin-echo images of a volunteer at the level of the basal ganglia collected in six separate sessions. Proton spectra were extracted from the left putamen for each of the studies. Water suppressed proton spectra were acquired at 0.5 Tesla using a spin-echo chemical shift imaging sequence (109).

CHEMICAL SHIFT IMAGING

The resonant frequency of a spin depends on the magnetic field from all contributions. This includes, most importantly, the main magnetic field. Other effects include gradients used for localization, local variations in the magnetic field produced by alterations in the bulk magnetic susceptibility of the material (susceptibility artifacts) and the chemical binding environment of the spin. For this last reason, a nucleus experiences a slight

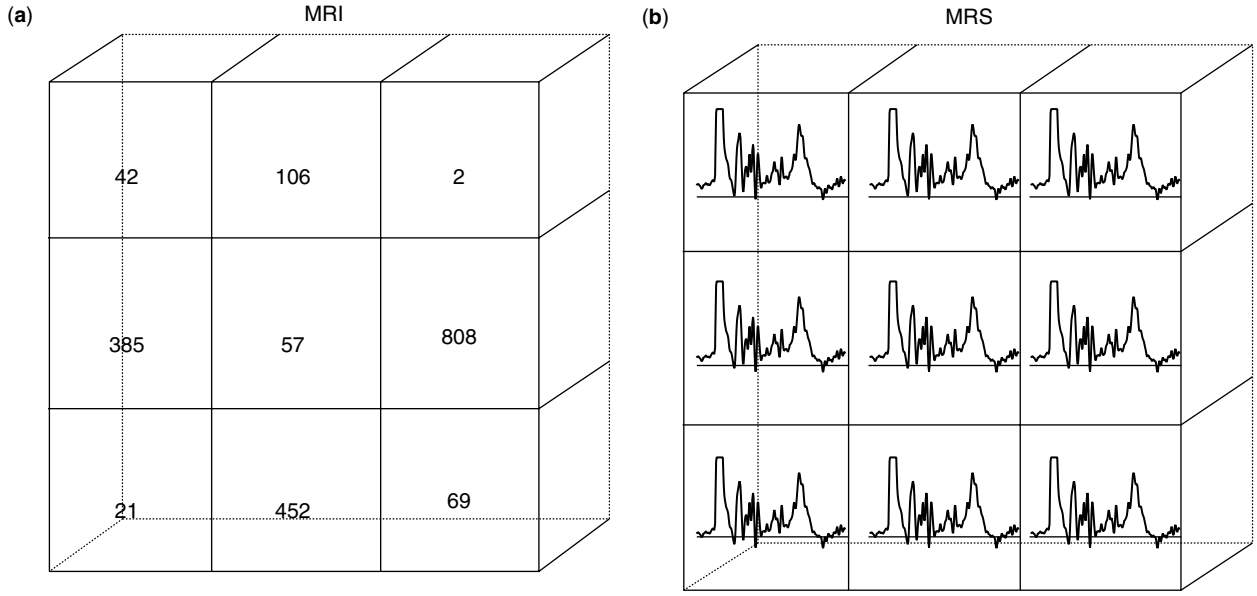


Figure 29. MRI image (**a**) is composed of scalar values. Data are available from each voxel, shown as individual spectra in (**b**) but are discarded in routine imaging. Note that the resonances in each voxel in (**b**) have much lower amplitude than water resonance (shown as truncated resonance).

shift in its resonant frequency which is characteristic of the molecule of which it is a part (the chemical shift). The chemical shift has been used to characterize chemical compounds in laboratory NMR spectrometers since the 1960s (15). Since the early 1980s, the technique has been applied in the human body within MRI scanners and is called magnetic resonance spectroscopy or MRS. One of the strongest advantages of MRS is the ability to demonstrate chemical change in the absence of morphological change. In Fig. 28, the images and corresponding MR spectra from six repeated studies of a volunteer are shown. In the third study (upper right hand corner), it was discovered incidentally that the volunteer had been drinking late into the night before the exam. The triplet resonance of ethanol is clearly shown in the spectrum (marked ETOH), whereas the corresponding image shows no change. The chemical shift can be exploited to create images of individual chemical moieties in an object by using chemical shift imaging (26). The information from all spins in each voxel of every MRI image is collapsed into a single scalar value to create a two-dimensional image (Fig. 29a). The readout gradient causes all of the spins in a voxel to be irretrievably mixed into a single scalar value. Using phase encoding gradients for both transverse axes, the chemical shift information is preserved and can be extracted by Fourier transform. An image created in this manner has three dimensions, two spatial and a third in chemical shift (Fig. 29b). Multiple images can be created from a single three-dimensional data set by binning the frequency dimensional data. Resulting images of individual resonances reveal the spatial distribution of metabolites. Figure 30 shows images from a patient who had a brain tumor. The first image is the $T2$ weighted image where the CSI sampled area is superimposed upon it in red. The lactate image demonstrates localized hypoxia, and the choline image shows membrane proliferation and by extension, growth of the tumor. The NAA (*n*-acetyl aspartate) image shows the distribution of healthy neurons, and the glutamate/glutamine images show the distribution of one of the principal neurotransmitters in the brain (96).

Trade-Offs

The limitation of MRI is the available signal. This signal is fundamentally limited by the polarization induced in the sample, which in turn is proportional to the strength of the main magnetic field and inversely proportional to the temperature of the sample. Increasing the magnetic field can cause other subtle degradations, including increased magnetic susceptibility artifacts and decreased contrast in $T1$ weighted scans due to the change of tissue $T1$ as field strength changes.

The maximum available signal is obtained at long T_r ($5 \times T1$ gives 99% of the available longitudinal magnetization) and the shortest T_e times allowed by the performance of the MRI system. Operation at these parameters yields only spin density contrast. $T1$ weighted contrast requires T_r times less than $5 \times T1$ and creates signal loss as $1 - e^{-T_r/T1}$. Likewise, $T2$ weighted sequences require T_e times that create signal loss as $e^{-T_e/T2}$.

Image signal-to-noise is a function of available signal (dependent on the image contrast parameters, as mentioned before), voxel volume, and total sampling time. In-plane spatial resolution is expensive in terms of signal-to-noise (SNR). Doubling the in-plane resolution without changing the field of view (FOV) decreases the voxel volume and thus the SNR by a factor of 4. The SNR limited nature of MRI means that the effects of trade-offs can be considered by referring to Eq. (14) (97).

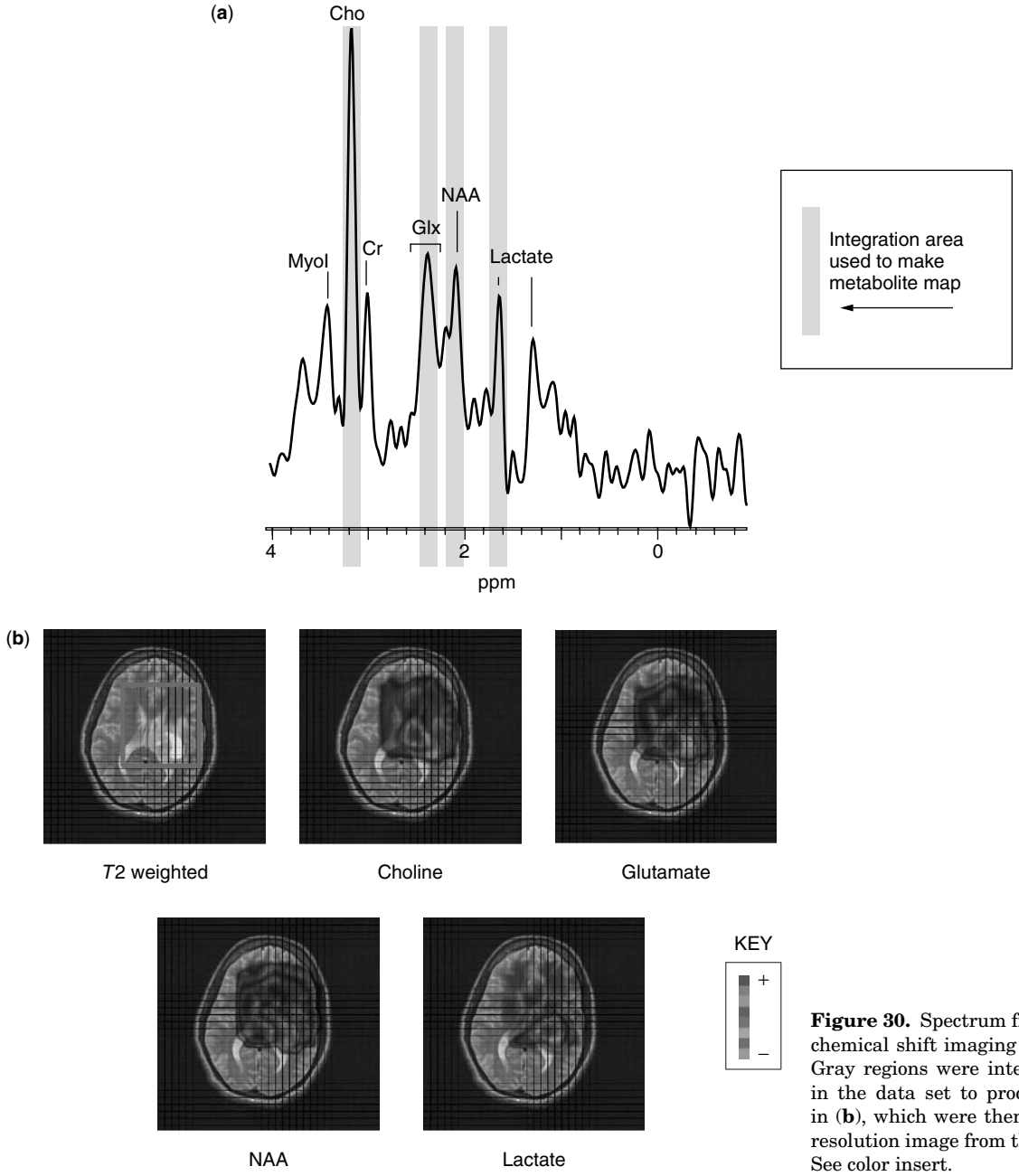


Figure 30. Spectrum from a single voxel in a chemical shift imaging data set of the brain. Gray regions were integrated for each voxel in the data set to produce false color maps in (**b**), which were then overlaid on the high resolution image from the same slice position. See color insert.

Instrumentation

An instrument for magnetic resonance imaging consists of a magnet to create the polarizing field, a gradient subsystem to create the localizing fields, a transceiver and radio frequency coils to transmit and receive the radio-frequency magnetic field, a control computer and a computer for reconstructing of the data, and an image visualization system (often combined with other computers).

Magnet. The fundamental and most critical piece of instrumentation in MRI is the magnet. Three technologies exist to create the main polarizing magnet field B_0 (98–101). The first is the permanent magnet. Typically made of ferrous or nickel alloys, the maximum field that can be generated is approximately 3,000 Gauss (0.3 Tesla). This technology is often employed in open architecture MRI systems (Fig. 31). Permanent field magnets cannot achieve the higher field strengths of superconducting magnets and are prone to field drift that is a strong function of ambient temperature.

The second method of producing the field is by resistive magnets. In these magnets, the field is produced by a large dc current flowing through windings, arranged as shown in Fig. 32. The core of the windings can be augmented by ferromagnetic alloys to improve the stability of the field. The deficiencies of this magnet technology include the requirement for constant current supply,

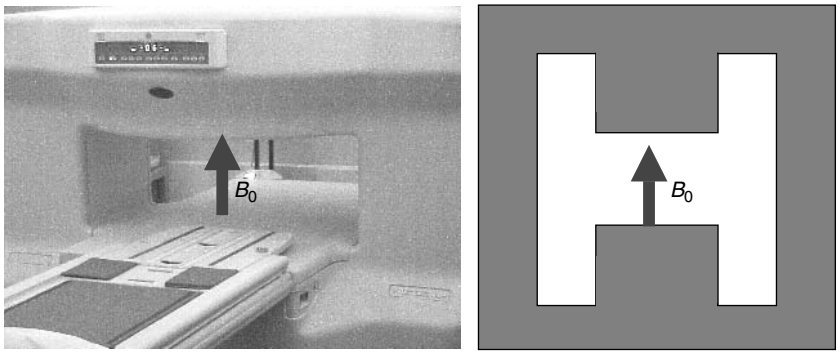


Figure 31. Photograph of permanent magnet MRI system showing the direction of the B_0 field and a cross section of the magnet within. Typical magnets are composed of boron–nickel hydride. See color insert.

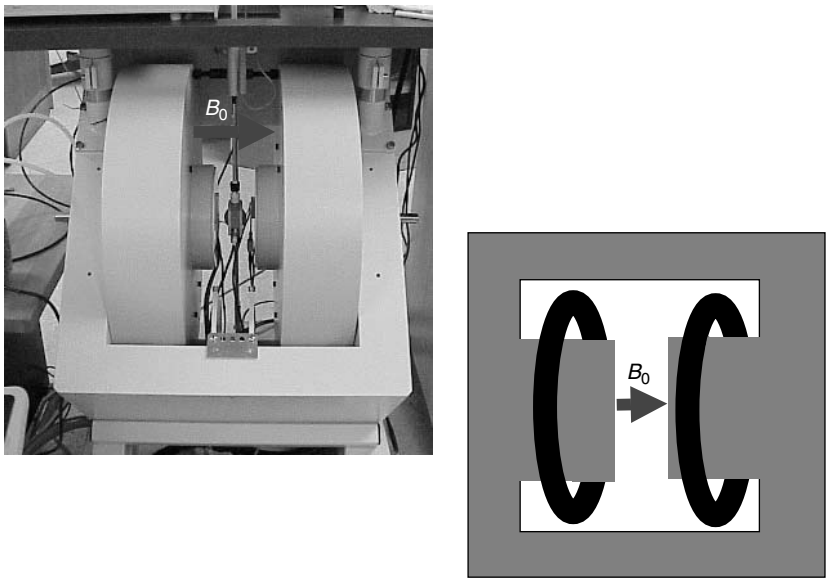


Figure 32. Resistive (electromagnet) design showing the windings and field direction. See color insert.

large cooling requirements, poor field stability, and the inability to achieve the field strength of superconducting magnets.

The third and most popular magnet technology is implemented using superconducting wire wound in an arrangement of four or more coils, as shown in Fig. 33. After the current is injected into the magnet, no further current is required. The magnet has excellent field stability, and no water cooling is required as for resistive magnets. The ambient temperature stability is much better than that of permanent magnets. However, most superconducting magnets require liquid helium to keep the magnet windings cold. Typical wire in such a magnet becomes superconducting at approximately 10°–14° K. Liquid helium at 4.2° K provides a bath that keeps the magnet superconducting. Liquid helium has a very small specific heat and boils off easily. If liquid helium boils off, the magnet can quench, lose field, and the magnet wire may be destroyed. The wire cannot support the required current when it is not superconducting and can melt when it suddenly becomes no longer superconducting. Many of the original magnet designs had a jacket of liquid nitrogen (temperature 77° K) surrounding the liquid helium bath to keep the liquid helium boil-off rate as low as possible. Now, many superconducting magnets use multiple, closed-cycle, helium refrigerators to keep the magnet windings cold, eliminating the need for liquid nitrogen. The downside of this magnet is that a loss of line power will cause the refrigerators to shut down, allowing the magnet windings to warm; without a liquid nitrogen bath to provide a thermal buffer, the liquid helium will boil off rapidly causing a possible wire-melting loss of superconductivity.

Gradients. The gradient fields are created by Maxwell pairs of coils that create a linear field across the volume to be imaged. Typically, three coil sets are used, one for each of the three principal coordinate axes, x, y, and z. Each coil set is connected to an amplifier that can drive tens to hundreds of amps of current into the coil. The self-inductance of the coil opposes the rapid rise in current which is desirable in an MRI system. Therefore, the amplifiers use voltages of the order of hundreds to thousands of volts for the segment of the gradient waveform where the current is rapidly increasing. When the current nearly reaches the final desired value, the voltage will decrease. In this way, smaller amplifiers can be used. The figures of merit for a gradient subsystem are the maximum field that can be created,

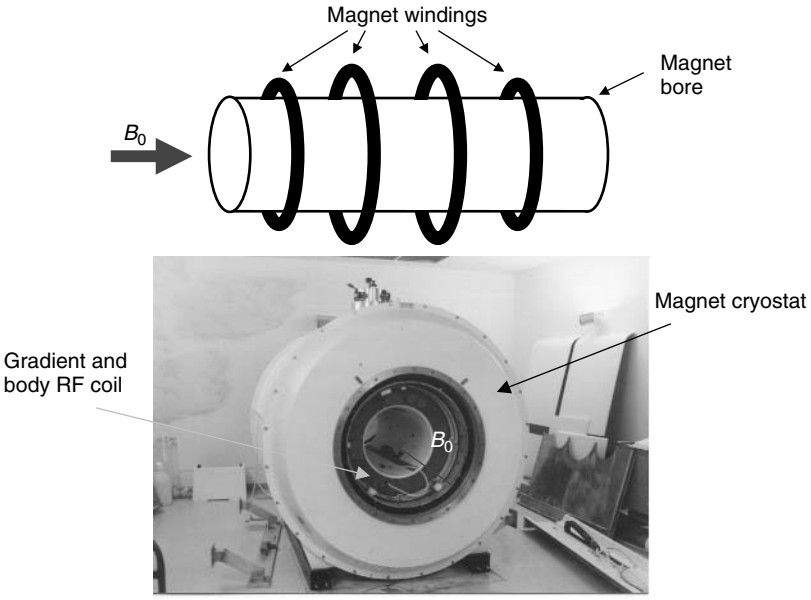


Figure 33. Superconducting magnet showing the direction of the main field and the arrangement of current-carrying coils. Most superconducting MRI magnets are built of four windings connected in series and arranged as shown. The variation in diameter is intended to improve the homogeneity of the field in the center of the magnet. See color insert.

measured in gauss/centimeter or millitesla/meter; the slew rate, measured in tesla/meter/second; and the linearity. Linearity is measured as the deviation from actual positional measurements of a phantom where the geometry of the image is known (102).

Transceiver. To observe the spin system in the rotating frame of reference, the spins must be communicated with at radio frequencies. At 1.5 tesla, proton spins precess at 63.86 MHz. By transmitting RF pulses whose carrier frequency is 63.86 MHz, the spins are efficiently nutated. Likewise, when receiving the signal induced in the radio-frequency coil, or antenna, the signal produced by the spins in the object is also at 63.86 MHz. A different field strength magnet, such as 0.5 tesla will have a different resonant frequency. For 0.5 tesla, the spins precess at 21.29 MHz. The transceiver creates the transmit field by mixing a carrier frequency (at the Larmor or resonant frequency of the magnet) whose wave shapes are created by a pulse control computer. The modulated carrier signal so produced is then fed into an RF power amplifier that provides sufficient power to nutate the spins in the sample. The power required from the RF power amplifier is a function of two variables. The first is the conductivity of the object. The more conductive the object, the more power must be applied to create the necessary nutation. Required RF power increases at the square of the increase in field strength for conductive bodies such as humans. The second variable is the size of the RF coil. A larger RF coil will dissipate more current in its own structure, and it will also usually contain a larger conductive object, and so will require more RF power (103).

RF Coils. Radio-frequency antennas for MRI, also known as RF coils, operate in an unusual manner for RF antennas. Most RF antennas can be designed by using far-field approximations of the behavior of the components. No so for MRI. Coils for MRI operate in the near-field in an attempt to maximize the SNR of the MR image (104). The SNR For MRI coils is determined by the relative contribution of the noise generated by the object and the noise generated as Johnson noise by the coil itself. As the sample becomes more conductive (typically higher in field strength), it is easier to make a coil whose contribution to noise is insignificant relative to the noise from the patient (105). RF coils for MRI are typically either volume coils that excite and receive NMR signals from a large volume or are surface coils that are sensitive only to local regions of a sample (106). Typical volume RF coils in an MRI system include body, head, and knee coils. The B_1 field created by the volume coil may be either linear or quadrature. A quadrature coil is analogous to a circularly polarized communications antenna. The benefits of quadrature coil design include a square root of 2 increase in image SNR and a reduction in required transmission power by a factor of 2 (107). Surface coils may be of any size or shape but most often are not used to transmit B_1 RF pulses. Instead, a volume coil such as the body coil is used to create a homogeneous RF transmission field (108). The surface coil has an inhomogeneous sensitivity profile that allows it to achieve high image SNR in the vicinity of the coil but would create a wide range of transmission field amplitudes. The large range of transmission fields would make the received image even less homogeneous.

The high power RF amplifier (100–20,000 watts) must be kept isolated from the RF preamplifier. The RF preamplifier receives a signal of the order of 10^{-12} watts that has an extremely low noise figure. The preamplifier would be destroyed quickly if it were not protected. The protection, called a T/R switch, is implemented as a series of diodes that can be switched on or off to direct the current flow.

Computer Systems. Computer system operations in an MRI scanner are divided into two distinct areas. The first is nominally referred to as the host computer. The

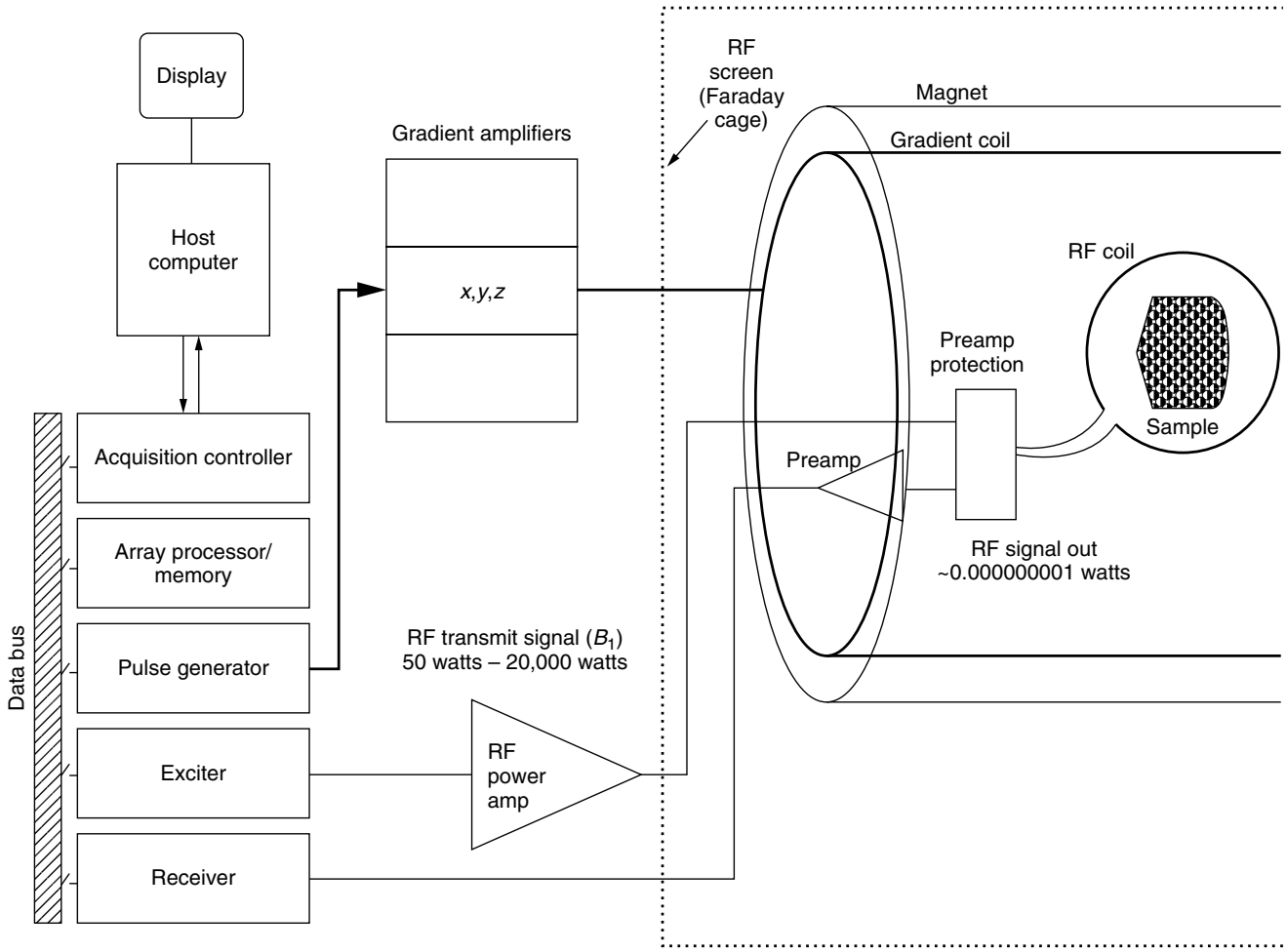


Figure 34. Block diagram of a typical MRI system.

host computer handles user interface, image display, and image storage functions. A second computer is required for real-time generation of pulse sequences (gradient, RF, and timing waveforms), as well as image reconstruction.

A block diagram of a typical MRI system is shown in Fig. 34.

BIBLIOGRAPHY

1. M. D. Hurlimann and D. D. Griffin, *J. Magn. Resonance* **143**, 120 (2000).
2. U. Yaramanci, *Annali Di Geofisica* **43**, 1,159 (2000).
3. Y. M. Zhang, P. C. Xia, and Y. J. Yu, *IEEE Trans. Appl. Superconductivity* **10**, 763 (2000).
4. D. E. Woessner, *Concepts Magn. Resonance* **13**, 77 (2001).
5. M. H. G. Amin, L. D. Hall, R. J. Chorley et al., *Progr. Phys. Geogr.* **22**, 135 (1998).
6. M. Appel, F. Stallmach, and H. Thomann, *J. Petr. Sci. Eng.* **19**, 45 (1998).
7. J. R. Bows, M. L. Patrick, K. P. Nott et al., *Int. J. Food Sci. Technol.* **36**, 243 (2001).
8. P. Cornillon and L. C. Salim, *Magn. Resonance Imaging* **18**, 335 (2000).
9. J. Gotz, W. Kreibich, M. Peciar et al., *J. Non-Newtonian Fluid Mech.* **98**, 117 (2001).
10. J. J. Gonzalez, R. C. Valle, S. Bobroff et al., *Postharvest Biol. Technol.* **22**, 179 (2001).
11. K. M. Keener, R. L. Stroshine, and J. A. Nyenhuis, *Trans. ASAE* **40**, 1,633 (1997).
12. I. L. Pykett, *IEEE Trans. Appl. Superconductivity* **10**, 721 (2000).
13. P. C. Lauterbur, *Nature* **242**, 190 (1973).
14. J. W. Prichard, J. R. Alger, and R. Turner, *British Medical Journal (BMJ)* **319**, 1,302 (1999).
15. E. Becker, *High Resolution NMR, Theory and Chemical Applications*, Academic Press, Orlando, FL, 1980.
16. R. R. Price, J. L. Creasy, C. H. Lorenz et al., *Invest. Radiol.* **27**, S27 (1992).
17. G. F. Mason, R. Gruetter, D. L. Rothman et al., *J. Cerebral Blood Flow Metab.* **15**, 12 (1995).
18. B. Mora, P. T. Narasimhan, B. D. Ross et al., *Proc. Natl. Acad. Sci. USA* **88**, 8,372 (1991).
19. A. Abragam, *Principles of Nuclear Magnetism*, Clarendon Press, Oxford, 1961.
20. G. A. Johnson, G. Cates, X. J. Chen et al., *Mag. Resonance Med.* **38**, 66 (1997).
21. K. Ruppert, J. R. Brookeman, K. D. Hagspiel et al., *Magn. Resonance Med.* **44**, 349 (2000).

22. Y. Wu, D. A. Chesler, M. J. Glimcher et al., *Proc. Natl. Acad. Sci. USA* **96**, 1,574 (1999).
23. G. F. Anderson and J. P. Poullier, *Health Affairs* **18**, 178 (1999).
24. T. Ernst, R. Kreis, and B. D. Ross, *J. Magn. Resonance Ser. B* **102**, 1 (1993).
25. E. O. Stejskal and J. E. Turner, *J. Chem. Phys.* **42**, 288 (1965).
26. T. T. Brown, B. Kincaid, and K. Ugurbil, *Proc. Natl. Acad. Sci. USA* **79**, 3,523 (1982).
27. J. A. Kaufman, S. C. Geller, M. J. Petersen et al., *Am. J. Roentgenol.* **163**, 203 (1994).
28. F. Guckel, G. Brix, K. Rempp et al., *J. Comput. Assisted Tomography* **18**, 344 (1994).
29. M. S. Cohen and S. Y. Bookheimer, *Trends in Neurosciences* **17**, 268 (1994).
30. C. T. Moonen, F. A. Barrios, J. R. Zigun et al., *Magn. Resonance Imaging* **12**, 379 (1994).
31. G. E. Santyr, *Magn. Resonance Imaging Clin. North America* **2**, 673 (1994).
32. A. D. Elster, J. C. King, V. P. Mathews et al., *Radiology* **190**, 541 (1994).
33. J. R. Hesselink and G. A. Press, *Radiol. Clin. North America* **26**, 873 (1988).
34. C. P. Slichter, *Principles of Magnetic Resonance*, Springer Verlag, Berlin, 1980.
35. T. C. Farrar, *Concepts Magn. Resonance* **2**, 1 (1990).
36. F. Bloch, *Phys. Rev.* **70**, 460 (1946).
37. T. C. Farrar, *An Introduction to Pulse NMR Spectroscopy*, Farragut Press, Chicago, 1987.
38. T. W. Redpath and J. M. Hutchison, *Phys. Med. Biol.* **27**, 443 (1982).
39. F. Bloch, W. W. Hansen, and M. Packard, *Phys. Rev.* **70**, 474 (1946).
40. P. Mansfield and P. G. Morris, *NMR Imaging in Biomedicine*, Academic Press, NY, 1982.
41. D. Nishimura, *Principles of Magnetic Resonance Imaging*, Stanford University Press, Stanford, CA, 1996.
42. R. C. Hawkes and S. Patz, *Magn. Resonance Med.* **4**, 9 (1987).
43. A. Oppelt, R. Graumann, and H. Barfuss, *Electromedica* **54**, 15 (1986).
44. H. Y. Carr and E. M. Purcell, *Phys. Rev.* **94**, 630 (1954).
45. S. Meiboom and D. Gill, *Rev. Sci. Instrum.* **29**, 6,881 (1958).
46. E. L. Hahn, *Phys. Rev.* **80**, 580 (1950).
47. C. T. W. Moonen and P. C. M. vanZijl, *J. Magn. Resonance* **88**, 28 (1990).
48. N. Bloembergen, E. M. Purcell, and R. V. Pound, *Phys. Rev.* **73**, 679 (1948).
49. R. L. Dixon and K. E. Ekstrand, *Med. Phys.* **9**, 807 (1982).
50. T. Jue, D. L. Rothman, J. A. Lohman et al., *Proc. Natl. Acad. Sci. USA* **85**, 971 (1988).
51. J. A. Jeneson, J. O. van Dobbenburgh, C. J. van Echteld et al., *Magn. Resonance Med.* **30**, 634 (1993).
52. R. Turner, *Magn. Resonance Imaging* **11**, 903 (1993).
53. R. Bowtell and P. Mansfield, *Magn. Resonance Med.* **17**, 15 (1991).
54. R. Damadian, *Philos. Trans. R. Soc. London — Ser. B: Biol. Sci.* **289**, 489 (1980).
55. P. Mansfield and A. A. Maudsley, *Phys. Med. Biol.* **21**, 847 (1976).
56. P. Mansfield, P. G. Morris, R. J. Ordidge et al., *Philos. Trans. R. Soc. London — Ser. B: Biol. Sci.* **289**, 503 (1980).
57. W. A. Edelstein, J. M. Hutchison, G. Johnson et al., *Phys. Med. Biol.* **25**, 751 (1980).
58. M. A. Reicher, R. B. Lufkin, S. Smith et al., *Am. J. Roentgenol.* **147**, 363 (1986).
59. V. M. Runge, M. L. Wood, D. M. Kaufman et al., *Radiographics* **8**, 507 (1988).
60. J. B. Ra, S. K. Hilal, and Z. H. Cho, *Magn. Resonance Med.* **3**, 296 (1986).
61. I. Shenberg and A. Macovski, *Med. Phys.* **13**, 164 (1986).
62. R. N. Bracewell, *The Fourier Tranform and Its Applications*, McGraw-Hill, NY, 1986.
63. D. I. Hoult and P. C. Lauterbur, *J. Magn. Resonance* **34**, 425 (1979).
64. J. Listerud, S. Einstein, E. Outwater et al., *Magn. Resonance Q.* **8**, 199 (1992).
65. W. A. Edelstein, G. H. Glover, C. J. Hardy et al., *Magn. Resonance Med.* **3**, 604 (1986).
66. D. A. Feinberg, J. D. Hale, J. C. Watts et al., *Radiology* **161**, 527 (1986).
67. R. E. Hendrick, T. R. Nelson, and W. R. Hendee, *Magn. Resonance Imaging* **2**, 279 (1984).
68. Y. P. Du and D. L. Parker, *Magn. Resonance Imaging* **14**, 201 (1996).
69. R. Damadian, *Science* **171**, 1,151 (1971).
70. A. F. Martino and R. Damadian, *Physiol. Chem. Phys. Med. NMR* **16**, 49 (1984).
71. P. A. Bottomley, T. H. Foster, R. E. Argersinger et al., *Med. Phys.* **11**, 425 (1984).
72. G. Bartzokis, M. Aravagiri, W. Oldendorf et al., *Magn. Resonance Med.* **29**, 459 (1993).
73. M. Graif, J. M. Pennock, J. Pringle et al., *Skeletal Radiol.* **18**, 439 (1989).
74. M. Dousset, R. Weissleder, R. E. Hendrick et al., *Radiology* **171**, 327 (1989).
75. F. W. Wehrli, A. Shimakawa, G. T. Gullberg et al., *Radiology* **160**, 781 (1986).
76. C. L. Dumoulin, S. P. Souza, M. F. Walker et al., *Magn. Resonance Med.* **9**, 139 (1989).
77. Y. Ozsunar and A. G. Sorensen, *Top. Magn. Resonance Imaging* **11**, 259 (2000).
78. N. J. Beauchamp Jr., A. M. Ulug, T. J. Passe et al., *Radiographics* **18**, 1,269 (1998).
79. R. H. Rosenwasser and R. A. Armonda, *Clin. Neurosurg.* **46**, 237 (2000).
80. S. L. Keir and J. M. Wardlaw, *Stroke* (Online) **31**, 2,723 (2000).
81. H. A. Rowley, P. E. Grant, and T. P. Roberts, *Neuroimaging Clin. North America* **9**, 343 (1999).
82. P. Bornert and D. Jensen, *Magn. Resonance Imaging* **12**, 1,033 (1994).
83. S. D. Wolff and R. S. Balaban, *Magn. Resonance Med.* **10**, 135 (1989).
84. M. A. van Buchem, *J. Comput. Assisted Tomography* **23**, S9 (1999).
85. L. J. Bagley, R. I. Grossman, and J. C. McGowan, *Neurology* **53**, S49 (1999).
86. I. R. Young, J. A. Payne, A. G. Collins et al., *Magn. Resonance Med.* **4**, 333 (1987).
87. E. C. Unger, M. S. Cohen, and T. R. Brown, *Magn. Resonance Imaging* **7**, 163 (1989).

88. R. C. Semelka, N. L. Kelekis, D. Thomasson et al., *J. Magn. Resonance Imaging* **6**, 698 (1996).
89. C. M. van Uijen and J. H. den Boef, *Magn. Resonance Med.* **1**, 502 (1984).
90. M. K. Stehling, R. Turner, and P. Mansfield, *Science* **254**, 43 (1991).
91. D. Chien and R. R. Edelman, *Magn. Resonance Q.* **7**, 31 (1991).
92. I. R. Young, A. S. Hall, and G. M. Bydder, *Magn. Resonance Med.* **5**, 99 (1987).
93. L. Kjaer and O. Henriksen, *Acta Radiologica* **29**, 231 (1988).
94. T. W. Chan, J. Listerud, and H. Y. Kressel, *Radiology* **181**, 41 (1991).
95. G. Sze, *Can. Assoc. Radiol. J.* **42**, 190 (1991).
96. R. W. Prost, L. Mark, and S. -J. Li, in *ISMRM*, Sydney, Australia, 1998.
97. W. A. Edelstein, P. A. Bottomley, H. R. Hart et al., *J. Comput. Assisted Tomography* **7**, 391 (1983).
98. F. Schmitt, A. Dewdney, and W. Renz, *Magn. Resonance Imaging Clin. North America* **7**, 733 (1999).
99. L. Marti-Bonmati and M. Kormano, *Eur. Radiol.* **7**, 263 (1997).
100. D. Hawksworth, *Magn. Resonance Med.* **17**, 27 (1991).
101. J. H. Battocletti, *Crit. Rev. Biomed. Eng.* **11**, 313 (1984).
102. J. W. Carlson, K. A. Derby, K. C. Hawryszko et al., *Magn. Resonance Med.* **26**, 191 (1992).
103. G. N. Holland and J. R. MacFall, *J. of Magnetic Resonance Imaging* **2**, 241 (1992).
104. J. H. Letcher, *Magn. Resonance Imaging* **7**, 581 (1989).
105. D. I. Hoult and R. E. Richards, *J. Magn. Resonance* **24**, 71 (1976).
106. J. F. Schenck, H. R. Hart, T. H. Foster et al., *Magn. Resonance Annu.* 123 (1986).
107. D. I. Hoult, C. N. Chen, and V. J. Sank, *Magn. Resonance Med.* **1**, 339 (1984).
108. C. E. Hayes, W. A. Edelstein, J. F. Schenck et al., *J. Magn. Resonance* **63**, 622 (1985).
109. R. W. Prost, in *Elliptical Excitation Chemical Shift Imaging at 0.5 Tesla in Neuro Degenerative Disease*, Medical College of Wisconsin, Milwaukee, WI, 1999, p. 194.

MAGNETOSPHERIC IMAGING

J. L. BURCH
Southwest Research Institute
San Antonio, TX

INTRODUCTION

The sun interacts with the earth to create a very dynamic and interesting region called the magnetosphere, that envelops the atmosphere and shields the immediate human environment from interplanetary space. As the technological age has advanced, the importance of the magnetosphere as a medium that must be monitored and whose behavior must be predicted has rapidly become apparent. The magnetosphere protects the earth's upper atmosphere from direct exposure to the flow of charged particles from the sun, known as the solar wind, and it is also the site of disturbances — geomagnetic storms and magnetospheric substorms — that can adversely affect human activities. For example, such disturbances generate strong electrical currents overhead in the ionosphere and below the earth's surface, that can disrupt communications and power systems. They also accelerate magnetospheric ions and electrons to extremely high energies, creating populations of high-energy particles that can damage electronic systems in space and that pose potentially serious health risks to astronauts.

The sun contributes to the formation of the earth's magnetosphere in two important ways. First, the solar wind — the supersonic stream of plasma that flows away from the sun and carries with it the embedded magnetic field, known as the interplanetary magnetic field (IMF) — applies mechanical pressure to the geomagnetic field, compressing it on the day side and stretching it into a tail many hundreds of earth radii long on the night side (1 earth radius = 6,378 kilometers). Only a small fraction of the solar-wind plasma enters the magnetosphere, but a larger fraction of the solar-wind electric field is admitted that drives the large-scale flow of plasma within the magnetosphere. Secondly, solar X rays and ultraviolet radiation strip electrons from atoms in the upper atmosphere, creating the partially ionized region of the atmosphere, known as the ionosphere. Like all plasmas, the ionosphere is electrically conducting and thus is strongly influenced by electromagnetic forces. Accelerated by electric fields and guided by the earth's magnetic field, ions and electrons flow continuously upward from the high-latitude ionosphere into the magnetosphere. This outflow is an important source of magnetospheric plasma.

To date, the magnetosphere has been investigated largely by measurements made by single spacecraft along their orbital tracks. The understanding derived from such localized measurements is, by necessity, partial and fragmented. Magnetospheric imaging is needed to provide the missing global context that will allow space researchers, for the first time, to study the earth's magnetosphere as a coherent global system of interacting components, driven by the highly variable input of mass, momentum, and energy from the solar wind.

The many spacecraft that have explored the magnetosphere have identified its average configuration and the different populations of plasmas and energetic particles within it. Strong shear flows and pressure gradients at the boundaries that separate these populations often lead to current systems that link the outer magnetosphere with the upper atmosphere. The resulting coupling waxes and wanes owing to the occurrence of geomagnetic storms and the more impulsive and shorter lived magnetospheric substorms. The most dramatic and familiar manifestations of these disturbances are the aurora borealis and aurora australis. These polar lights are like images on a TV screen upon which is projected the global distribution of charged particles and currents in the magnetosphere. Figure 1 shows an artist's conception of the average configuration of the magnetosphere.

In situ measurements during the past 35 years have yielded a wealth of statistical information about the

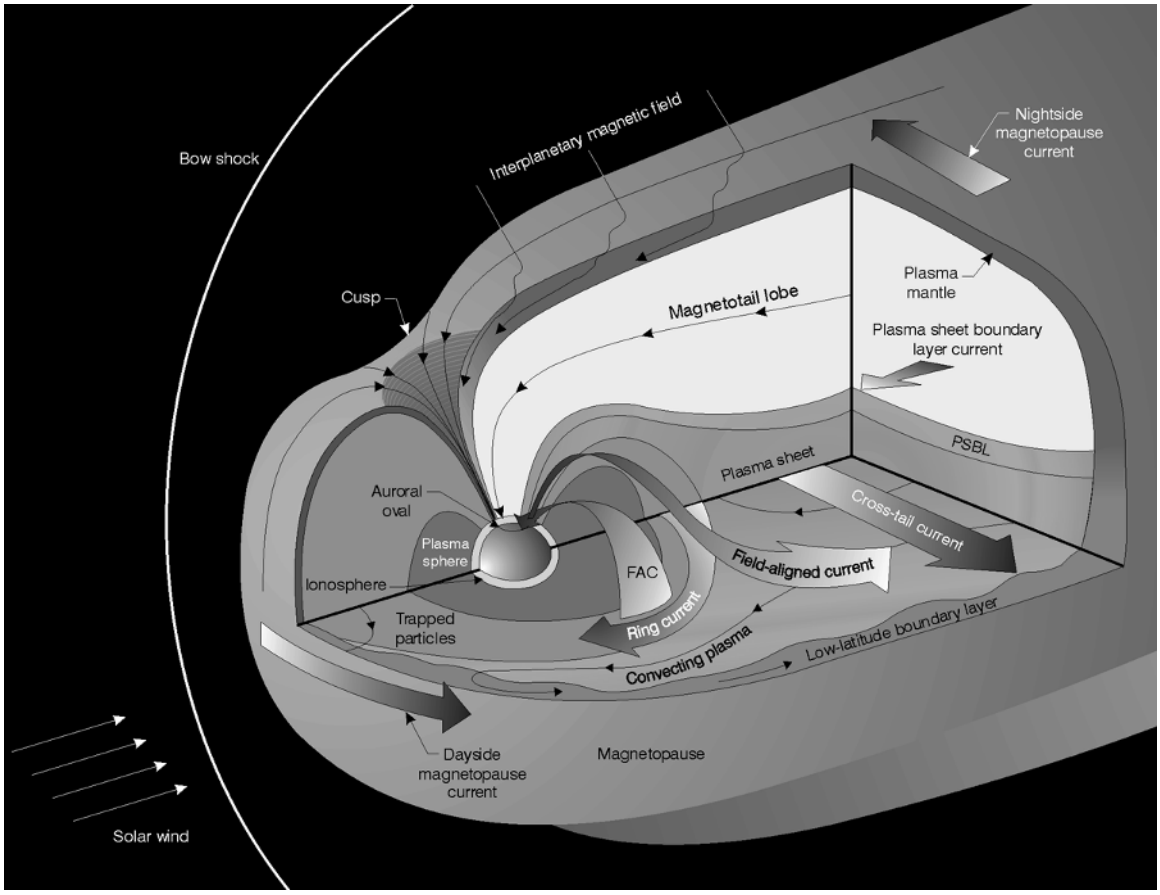


Figure 1. Artist's drawing showing the principal plasma regions and boundaries of the earth's magnetosphere. Different imaging techniques are used according to the properties of the plasma population to be imaged. Traditional photon imaging at ultraviolet wavelengths is used for the low-energy plasma of the plasmasphere, and the detection of energetic neutral atoms is used to image the higher energy plasmas of the ring current/radiation belts, inner plasma sheet, and cusp.

magnetosphere and its constituent plasma regimes. They have also provided many examples of dynamic changes of magnetospheric plasma parameters at specific times and places in response to changes in the solar-wind input and to internal disturbances related to substorms. Statistical global averages and individual events are, however, not sufficient to understand the dynamics and interconnection of this highly structured system. Fundamental questions concerning plasma entry into the magnetosphere, global plasma circulation and energization, and the global response of the magnetospheric system to internal and external forcing remain unanswered. These processes occur on time scales of minutes to hours, yet, currently available statistical averages are on time scales of months to years. To further our understanding of the physical processes that affect the magnetospheric plasma requires nearly instantaneous measurement of its structure, which can be obtained only by magnetospheric imaging.

To image a large, complex region such as the magnetosphere, which is invisible to traditional astronomical techniques, it is necessary to image its component parts. The particular imaging technique to be employed depends to a large degree on the properties of the plasmas that are to be imaged. One technique for magnetospheric imaging is *extreme ultraviolet* (EUV) *imaging*. This technique uses the fact that helium ions in the sun emit ultraviolet light at 30.4 nm. This light is absorbed and then rapidly reemitted by helium ions in the magnetosphere. The greatest concentration of helium ions is in the plasmasphere, the high-altitude extension of the ionosphere. Most of the ions in the plasmasphere are hydrogen, but hydrogen ions produce no light emissions. On the other hand, there is a nearly constant fraction ($\sim$10–15%) of helium in the plasmasphere, which can be used as a tracer for the region. Recently, the first images of the plasmasphere taken from outside the region were obtained by an EUV imager on the Japanese Nozomi (Planet-B) spacecraft while it was in earth parking orbit before it was injected into an interplanetary trajectory to Mars (Fig. 2). The image in Fig. 2 took nearly 21 hours to construct because the spacecraft was not spinning and had to rely on its orbital motion to complete a raster scan of a large portion, but not all, of the plasmasphere (1).

For more energetic ion populations, a different technique, *energetic neutral atom* (ENA) *imaging*, is used. Because the target ions in the earth's inner magnetosphere are trapped by the geomagnetic field, they can be

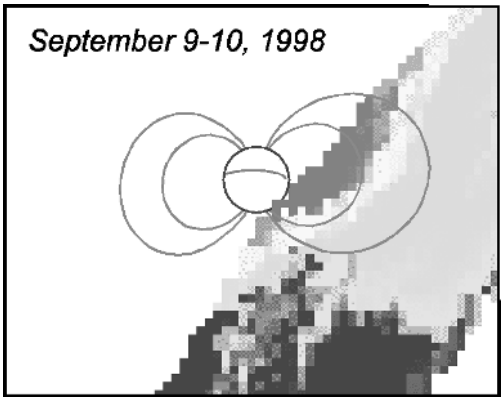


Figure 2. A portion of the earth's plasmasphere imaged at extreme ultraviolet wavelengths by the Japanese Nozomi (Planet B) spacecraft. The image was created from photons emitted from singly ionized helium at a wavelength of 30.4 nm and accumulated during a nearly 21-hour period in September 1998 while the spacecraft was in a parking orbit around the earth (1). (Image from M. Nakamura et al., *Geophysical Research Letters*, Vol. 27, No. 2, 141–144, January 15, 2000. Copyright 2000 American Geophysical Union. Reproduced by permission of the American Geophysical Union.)

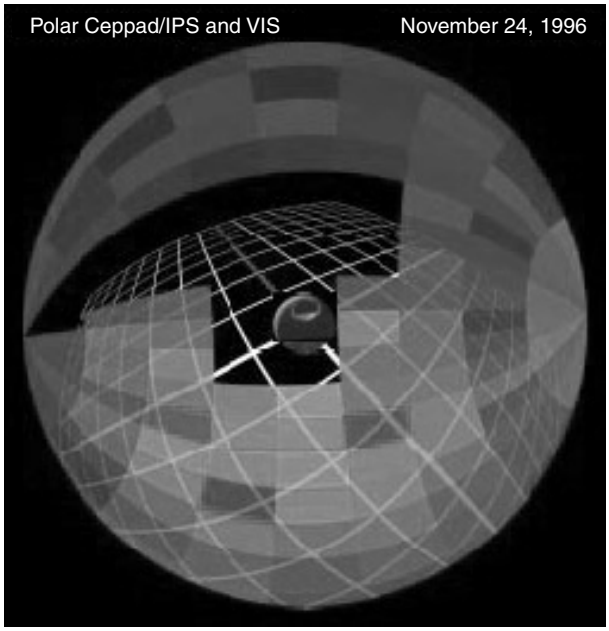


Figure 3. Energetic neutral atom image of the earth's space environment during a magnetospheric disturbance. The image was acquired with the CEPPAD instrument on NASA's Polar spacecraft. The view is from high northern latitudes at an altitude of about 25,000 km. The earth is shown by an ultraviolet image obtained on Polar. The neutral atoms have energies between 75 and 113 keV. The image is a 10-minute average shown in a hemispheric fisheye projection. CEPPAD is a charged particle detector and was not designed for ENA imaging. However, owing to the low charged particle background at the particular segment of the spacecraft orbit where this image was acquired, CEPPAD was able to detect the energetic neutrals produced by charge exchange between the energetic radiation belt particles and exospheric neutral hydrogen (4).

imaged on a global scale from remote locations only if they are converted to neutral atoms and released from their magnetic confinement to travel along line-of-sight paths to a remote imager. This conversion occurs by charge exchange. In this process, a few percent of the energetic ions in a particular population pick up electrons from the neutral hydrogen atoms that occupy the outer reaches of the earth's atmosphere (the exosphere). Charge exchange preserves the parent ion's direction, speed, and mass so that suitable ENA cameras can obtain detailed information about the magnetospheric ion populations.

ENA imaging of ions at ring-current energies has been accomplished with instruments originally designed to measure energetic ions on at least two spacecraft (2,3). Figure 3 shows an image obtained from the NASA Polar spacecraft (4). These instruments were able to form the ENA images when the spacecraft were in the region above the earth's polar cap, where the fluxes of energetic ions were below the instrument threshold. The energetic ion instruments were essentially pinhole cameras that had solid-state detectors at the image plane. Because only a few per cent of the energetic ions in the ring current and radiation belt undergo charge exchange and the distance of the spacecraft from the ion populations was several earth radii, the ENA fluxes in the region were very low even though they still considerably exceeded the energetic ion fluxes. For example, the image in Fig. 3 took about one and a half minutes to form, and considerable smoothing was required because the pixel size of the raw images was $20^\circ \times 12^\circ$. Thus, to achieve an acceptable signal-to-noise ratio, ENA cameras need to be considerably larger than the ion detectors that typically have flown on magnetospheric satellites.

A necessary component of neutral atom imaging is the translation of the neutral atom images into images of the parent ion populations (5,6). The generation of ion images from neutral atom images requires knowledge of the hydrogen atom density of the upper atmosphere, which can be provided by imaging the hydrogen atoms of the earth's exosphere using resonantly scattered solar Lyman-α emissions (121.6 nm). The Lyman-α "glow" of the exosphere is called the geocorona. The neutral atom populations being imaged are "optically thin," or transparent. For such a transparent medium, the signal obtained is the integrated signal along the line of sight through the imaged region.

The charge-exchange process also makes possible ultraviolet imaging of energetic protons that precipitate into the atmosphere and produce the proton aurora. Until now, the aurora is the only magnetospheric phenomenon that has been imaged routinely. These images are made at wavelengths at which nitrogen and oxygen are excited by precipitating energetic electrons. Using grating-based techniques, it has become possible to image energetic neutral hydrogen atoms produced by charge exchange of energetic protons just before they impact the upper atmosphere to produce the proton aurora (7). In this case, rather than imaging the aurora itself, the camera would actually be imaging the protons that produce the aurora. Such imaging is very challenging in that the resonantly scattered solar Lyman-α radiation (121.6 nm) from the

energetic hydrogen atoms has to be separated from the radiation from the "cold" hydrogen atoms at the same wavelength in the upper atmosphere. The separation is made possible because the energetic atom emissions are red-shifted when viewed from high altitudes in the magnetosphere. For particles whose energies are several keV, a red-shifted emission at 121.8 nm can be imaged.

For the purposes of this article, magnetospheric imaging will be defined as the collection of techniques that can provide large-scale images of various magnetospheric charged particle populations on time scales that are relevant to those of magnetospheric disturbances (i.e., a few minutes). This definition will include imaging precipitating energetic protons. Radio sounding of the magnetosphere — a technique employed in NASA's IMAGE (Imager for Magnetopause to Aurora Global Exploration) mission (8) — is beyond the scope of the article. However, the radio sounding system on IMAGE (9) has already demonstrated the capability of remotely mapping the plasmasphere, cusp, and magnetopause. The current state-of-the-art magnetospheric imaging technology is embodied in the instrumentation developed for the IMAGE mission (launched on 25 March, 2000), which is the first mission dedicated to global imaging of the terrestrial magnetosphere. Thus the description of imaging techniques in the following sections will focus specifically on the IMAGE instruments. The discussion is ordered according to the particle populations to be imaged.

EXTREME ULTRAVIOLET (EUV) IMAGING OF THE PLASMASPHERE

The plasmasphere is populated by the outflow of ionospheric plasma along mid- and low-latitude magnetic field lines (i.e., those that map to magnetic latitudes of $\sim 60^\circ$ and less). The plasma of the plasmasphere "corotates" with the earth, unlike the plasma of the magnetosphere's central plasma sheet, which in general "convects" (i.e., flows) sunward toward the dayside magnetopause. When the magnetosphere is disturbed by a magnetic storm, enhanced convection "erodes" the outer plasmasphere, captures plasma in the afternoon-dusk sector, and transports it outward and sunward toward the magnetopause. As geomagnetic activity abates, the ionosphere "refills" the plasmasphere. By imaging the helium ions from above the plasmasphere, it is possible to map the refilling of the plasmasphere during quiet times and the stripping away of its outer regions during magnetic storms. Direct measurements of the ions from spacecraft have led to intriguing, but unproven, notions that long plasma tails or detached islands are produced and are driven out through the dayside boundary of the magnetosphere by high-altitude winds. Global helium ion images can to test these theories definitively.

To achieve the goal of imaging the earth's plasmasphere, an extreme ultraviolet imager must

1. accommodate a maximum 30.4-nm brightness of $\sim$10 R (Rayleighs) in the plasmasphere as well as the much brighter, localized ionospheric source;
2. image emissions of 0.1–0.3 R in an integration time of several minutes;
3. have a wide field of view (extending out to at least a seven earth radii geocentric distance) to encompass the entire plasmasphere in a single "snapshot," and
4. reject bright contaminating emissions such as 121.6 nm from the geocorona and interplanetary medium and the He 58.4-nm line emission from the upper atmosphere.

The EUV imager on the IMAGE spacecraft is an instrument that meets these requirements (10).

Figure 4 illustrates the optical design of the three IMAGE EUV sensor heads. Their characteristics are summarized in Table 1. The entrance aperture consists of four segments that collectively form a nearly complete annulus. Each of the four segments of the annulus contains a thin (150-nm) aluminum filter that transmits the He^+ 30.4-nm line and excludes the bright geocoronal H Lyman-α line at 121.6 nm, as well as other emissions. Light that passes through the filter reaches the mirror. For good reflectivity at the target wavelength, the mirror has a multilayer coating formed of depleted uranium and silicon. Light reflected from the mirror is focused on the spherical focal surface of the microchannel plate (MCP) detector.

To achieve the 90° field of view required for plasmaspheric imaging, three EUV sensor heads are used, and their respective 30° fields of view overlap by 3° to compensate for vignetting effects. The front of the sensor head is insulated from the main body and held at +15 V to exclude low-energy ions from the sensor head. The sensor heads are copiously vented to avoid damage to the filters

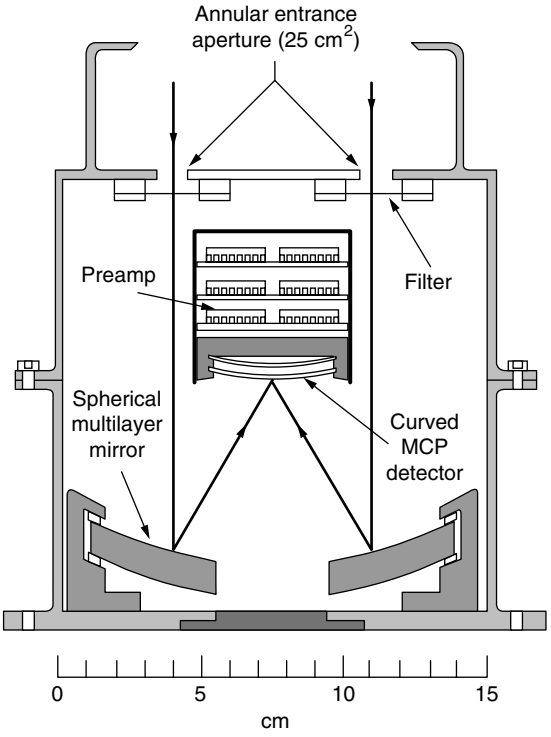


Figure 4. Cross-sectional view of one of the three IMAGE EUV sensor heads (10).

Table 1. EUV Sensor Characteristics

Property	Value
Aperture area	21.8 cm^2
Aperture diameter (outer)	8.2 cm
Aperture diameter (inner)	6.0 cm
Mirror diameter	12.3 cm
Mirror radius of curvature	13.5 cm
Focal ratio	f/0.8
Detector active diameter	4.0 cm
Radius of curvature of focal surface	7.0 cm

from the acoustic environment at launch. All of the vent openings are covered by a fine mesh that is also held at +15 V.

Filters

The filter material was chosen to provide high transmission at the target wavelength of 30.4 nm while attenuating the unwanted hydrogen Ly-α emission from the earth's exosphere and the interplanetary medium. The filters consist of a 150-nm thick layer of aluminum supported by a nickel mesh that has a wire spacing of 0.36 mm and an open area of 93%. The aluminum material was permitted to oxidize naturally on both sides. This oxide coating stabilizes the material and provides the additional benefit of reducing the transmission of the He 58.4-nm line. This emission is expected to be weak in the plasmasphere but can be quite bright in the earth's ionosphere. A strong response to the ionospheric He 58.4-nm emission could lead to high count rates when the earth is in the field of view and thus compromise measurements of 30.4-nm radiation. The measured transmission of the IMAGE EUV filters at 30.4 nm and 58.4 nm is 33% and 12%, respectively. At 121.6 nm (Ly-α), the transmission is essentially unmeasurable, but models show that it is approximately 10^{-7}.

Multilayer Mirrors

Almost all materials absorb highly in the extreme ultraviolet. This means that only a few interfaces in a multilayer contribute to its reflectance. Surface contamination and corrosion are particular problems because as little as 0.05 nm of an oil film or surface oxide can significantly change a mirror's reflectance. The Ly-α 121.6-nm and He I 58.4-nm lines are much brighter than the 30.4-nm line. The 121.6-nm line is blocked by the Al filter, but Al is relatively transparent to 58.4-nm radiation. There is no acceptable filter for blocking 58.4-nm radiation that meets the other mission requirements. The optical response of the mirrors must be retained over a 6° range of incidence angles (12°–18° from normal), which is relatively broad for multilayer mirrors. The EUV instrument uses a new kind of extreme ultraviolet mirror that retains high 30.4-nm reflectivity and low 58.4-nm reflectivity, which is challenging because most materials have much higher single-layer reflectivities at 58.4 nm than at 30.4 nm. To reflect well at 30.4 nm and poorly at 58.4 nm, the top part of the stack, which

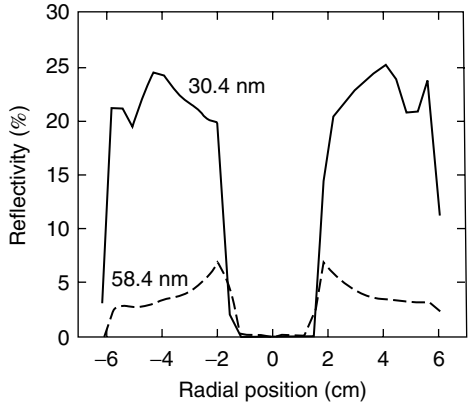


Figure 5. The measured reflectivity of one of the IMAGE EUV multilayer mirrors (10).

reflects well at 30.4 nm, must act as an antireflection layer at 58.4 nm. Such mirrors that have dual optical function in the extreme ultraviolet have not been designed or built previously. Depleted uranium proved to be the only metallic material for the high-index (the so-called absorber) layer which also produced low 58.4 nm reflectance. The genetic algorithm (11) calls attention to the possibility of using oxidized uranium as the top layer material and shows the proper range of thicknesses to consider. The top layer of the mirror consists of a thin layer of UO_X produced by oxidizing of 1.5 nm of uranium within a few minutes of exposure to air. (If the film oxidizes to UO_3, the layer will swell to more than three times its original thickness). This oxide functions as the top high-index layer in the multilayer when considering high reflectance at 30.4 nm. This is the first time this has been done for EUV multilayers. The oxide is also largely responsible for the low reflectance at 58.4 nm. Beneath the uranium oxide layer are 6.5 periods of bilayers that have 12.8 ± 0.01 nm of silicon and 5.3 ± 0.1 nm of uranium. The bottom layer consists of one 10.6-nm layer of uranium, twice the uranium thickness used in the other layers. This extrathick bottom layer allows the coating to be released from the glass, if recoating is necessary, without compromising optical performance. All films were deposited by dc magnetron sputtering. The detailed process for fabricating the EUV mirrors is described by Sandel et al. (10). Figure 5 shows the measured reflectivity of one of the IMAGE EUV multilayer mirrors.

The aluminum filters, spherical multilayer mirrors, and spherical MCPs of the IMAGE EUV instrument provide excellent imaging of the plasmasphere. Figure 6 shows a global plasmaspheric image obtained by the EUV instrument every ten minutes during the two-year IMAGE mission.

ENERGETIC NEUTRAL ATOM IMAGING AT HIGH ENERGIES: RADIATION BELTS AND RING CURRENT

Charged particles that have energies from a few tens of keV up to several MeV are trapped in the earth's magnetic field and form the well-known Van Allen belts. One of the primary replenishing mechanisms for the radiation

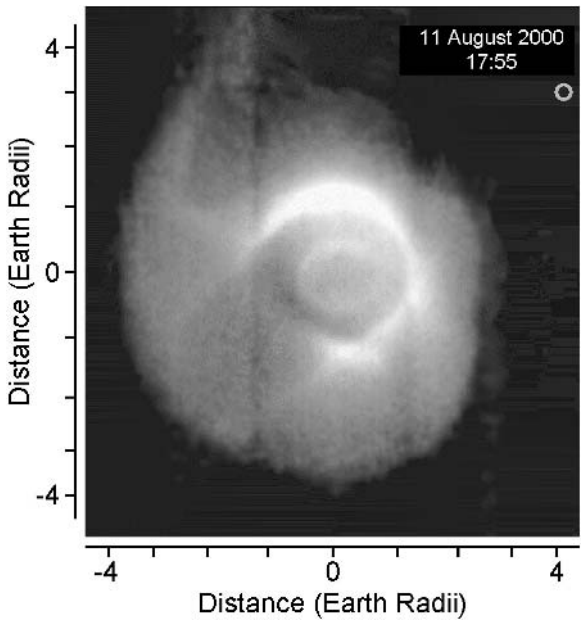


Figure 6. Global view of the Earth's plasmasphere on August 11, 2000, imaged with the EUV imager on NASA's IMAGE spacecraft. The extreme ultraviolet emissions are solar 30.4-nm photons that have been resonantly scattered by singly ionized helium. The view is toward the Earth's northern hemisphere from an altitude of roughly 26,500 km. The small circle in the upper right indicates the direction toward the Sun. The darker region in the lower left results from shadowing by the Earth. A "tail" of plasma can be seen extending from the dusk plasmasphere in the upper left. (Image from B. R. Sandel et al., *Geophysical Research Letters*, Vol. 28, No. 8, 1439–1442, April 15, 2001. Copyright 2001 American Geophysical Union. Reproduced by permission of the American Geophysical Union.) See color insert.

belts is magnetic storms, during which plasma is injected toward the inner magnetosphere from the plasma sheet (see Fig. 1) and forms a ring current that eventually surrounds the earth. The ring current is responsible for the magnetic perturbations that are observed on the surface of the earth during storms. There is also an ionospheric component to the ring current because ionospheric ions are accelerated out into the magnetosphere in association with the enhanced auroral activity that occurs during the storms. The freshly injected ring-current ions have energies between several tens of keV and a few hundred keV. The ions have been measured by numerous spacecraft flying through the region, but understanding the source and loss processes of the ring current requires global imaging, which can be done with energetic neutral atom (ENA) cameras.

To achieve the goal of imaging the earth's radiation belts and ring current, an ENA imager must

1. detect atoms whose energies range from a few tens of kev to several hundred keV and have an effective geometric factor of $\sim$1 cm^{2}sr,
2. obtain at least crude energy resolution (ΔE/E$\sim$0.8) and mass resolution (separate H and O),
3. form images whose a spatial resolution in the ring current and radiation belts is $1R_E$ or less,
4. have a wide field of view (extending out to at least a seven earth radii geocentric distance) to encompass the entire ring current and radiation belts in a single "snapshot,"
5. reject bright contaminating emissions such as 121.6 nm from the exosphere and interplanetary medium, and
6. reject energetic ion and electron fluxes.

The High-Energy Neutral Atom imager (HENA) on the IMAGE spacecraft is an instrument that meets these requirements (12). The HENA sensor characteristics are summarized in Table 2.

HENA Optics

A schematic view of HENA is shown in Fig. 7. The instrument is a slit camera that has a $90^\circ \times 120^\circ$ field of view that operates over an energy range from 20–500 keV. A collimator that sweeps out energetic ions and electrons is in front of the entrance slit. The slit is covered by a thin foil that absorbs a large fraction of the ultraviolet flux and also provides a start signal for time-of-flight analysis by virtue of secondary electron emission from its back side. An MCP that extends the length of the foil detects the secondary electrons and, using its one-dimensional position-sensitive anode, locates the entrance position of the atom along the slit. The ENAs that have passed through the foil suffer a certain amount of scattering but continue in approximately a straight-line path to the back plane of the instrument where they encounter one of two types of detectors–a position-sensitive solid-state detector (SSD) or an MCP mounted behind a second foil. Secondary electrons produced by the arriving ENAs at either the SSD or the front of the second foil are steered by wire electrodes to trigger the coincidence/SSD-stop MCP, so-called because it provides fast stop pulses for the SSD detector and also, together with the start and stop pulses from the start and

Table 2. HENA Characteristics

Energy range	$\sim$20–500 keV/nuc.
Energy resolution [$\Delta E/E$]	$\leq$0.25
Velocity resolution	$\sim$50 km/s (1 ns TOF)
Composition	H, He, O, heavies
Field of view	$120^\circ \times 90^\circ$
Angular coverage	$\sim 3\pi$ sr, spinning
Angular resolution ($\Delta\phi\Delta\theta$)	$\sim 4^\circ \times 6^\circ$, high energy (>80 keV/nuc)
Angular resolution is degraded by ENA scattering in foils at low E	$<8^\circ \times <12^\circ$ above $\sim$40 keV/nuc (see Fig. 9)
Time resolution (statistics limited)	2.0 s., PHA events 1 spin, 2 min, images
Sensitivity, $G \times \varepsilon$(cm^2-sr)	$\sim$1.6 for O, $\sim$0.3 for H
$G \times \varepsilon$ for $3^\circ \times 3^\circ$ pixel (image accumulation bin)	0.0027, tapered at FOV edges
Pixel array dimensions[a]	$4M \times 8E \times 40$ A $\times$ 120 A
UV rejection	$<3 \times 10^{-7}$(Ly-α, I below 10 kR)
Electron rejection	$\sim 10^{-5}$ ($E <$ collimator rejection)
Ion rejection	$\sim 10^{-5}$ ($E <$ collimator rejection)
Dynamic range	$\sim 10^7$

[a] $E \Rightarrow$ energies; $M \Rightarrow$ masses; $A \Rightarrow$ angles.

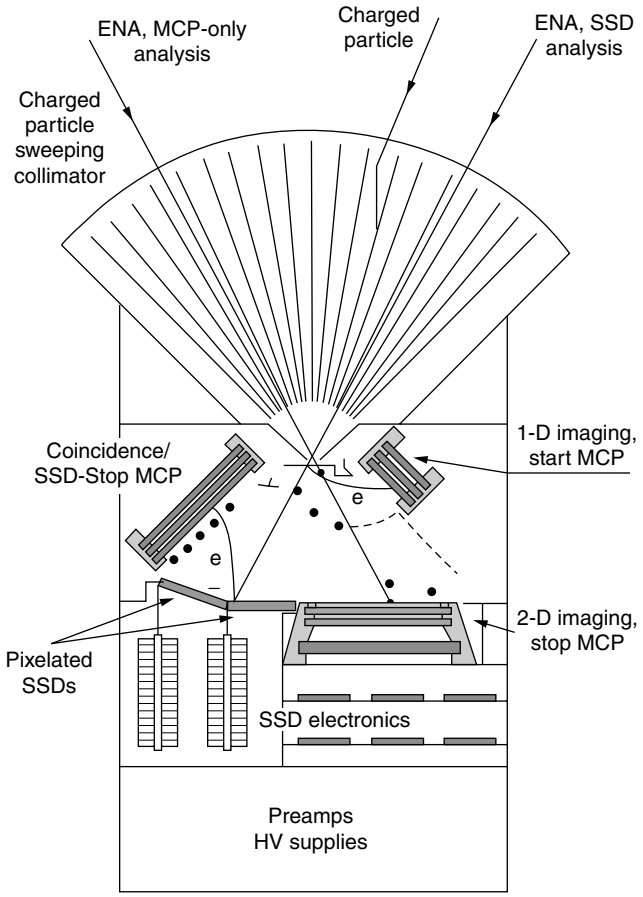


Figure 7. Cross-sectional schematic view of the IMAGE high-energy neutral atom (HENA) imager (12).

stop MCPs, provides triple-coincidence rejection of signals from photons or high energy electrons. The dots in Fig. 7 indicate the locations of the wire electrodes for steering the secondary electrons to their respective MCPs.

Based on its $90^\circ \times 120^\circ$ field of view, the HENA instrument could obtain suitable images of the inner magnetosphere from a nadir-pointing spacecraft. However, other instruments on board IMAGE require a spinning spacecraft. The spin rate is 0.5 rpm, and the spin axis is perpendicular to the plane of Fig. 7. Thus a single target pixel moves across the MCP and SSD detector planes as the spacecraft spins and results in both an MCP image and a SSD image (that has its higher resolution mass analysis) of the ring current and radiation belts every two minutes.

If an ENA is directed toward the MCP portion of the focal plane, secondary electrons produced on the exit surface of the second foil are accelerated into the MCP that has a 2-D imaging anode that maps the position of impact and registers the stop timing pulse for the TOF measurement. In addition to this TOF and trajectory measurement, secondary electrons produced as the ENA enters the back foil are electrostatically accelerated and guided to the coincidence/SSD stop MCP (C/S MCP) (see Fig. 7). The electron travel time for the backscattered electrons from the back foil to the C/S MCP is constrained to <40 ns by the steering potentials.

When an ENA strikes the SSD, both position (pixel ID) and total energy (through pulse-height analysis) are derived. Secondary electrons generated at the surface are accelerated into the C/S MCP and have a travel time between the SSD and the MCP of between 6 and 10 ns, depending on which row of SSD pixels is struck. The signal produced in the C/S MCP is corrected for electron travel time and used for the SSD TOF. The TOF solution measures the velocity of the ENA, which, when combined with the total energy measurement, allows calculating the atom's mass.

A different technique may be used on the MCP side of the back plane for mass determination. The number of secondary electrons produced in each foil depends on the atomic number of the neutral atom; of the two most common neutral atoms expected (H and O), it has been found that oxygen produces several times the number of secondaries that hydrogen produces. Exploiting this phenomenon, the pulse-height of the MCP signal is used to determine the species based on that measurement. This method is optimized by pulse-height analysis (PHA) of both the signal from the front of the MCP and the signal from the back of the MCP, as shown in Fig. 8. This technique is the primary approach to mass determination on the Cassini INCA sensor (13), which has no SSD system. For HENA, it is used as a complement to the SSD system.

Foils and UV Sensitivity

There are two thin foils in HENA, the Si–polyimide–C foil at the entrance slit and the C–polyimide–C foil at the back plane in front of the 2-D MCP array. These foils are the source of secondary electrons for start and stop pulses used in the TOF analysis. The second foil also provides backscattered electrons for the coincidence pulse. However, electrons are also produced in the foils from photoionization by ultraviolet light, predominantly in the Ly-α line from the earth's hydrogen exosphere. Therefore, the foil must be thick enough to attenuate the Ly-α to an acceptable level, so that the position and TOF circuitry

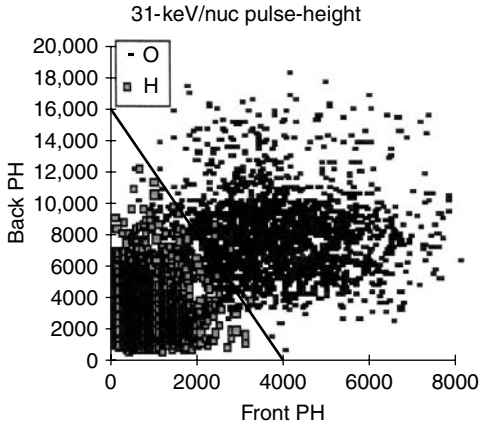


Figure 8. Separation of oxygen and hydrogen using MCP pulse-height signatures. The 31-keV O and H measurements are overplotted. The hydrogen separates as lower MCP pulse-heights. At higher energies, this separation becomes more complete. Data from calibration of Cassini/MIMI/INCA sensor (13).

are not swamped by false counts. Any single photon can produce a start or a stop signal, but not both. As long as the time between a false Ly-α start and a separate false Ly-α stop is long compared with the maximum valid TOF period ($\sim$100 nsec), this background will not produce false events. The electronic design assumes nominal upper limits for Ly-α start and stop events of $\leq 10^5\ s^{-1}$, and $\leq 1,000\ s^{-1}$, respectively. The Ly-α flux from the sunlit atmosphere at earth (excluding the Sun itself) is $\sim$10–30 kR (14), which corresponds to $\sim 0.32-1.0 \times 10^{10}$ photons/s incident on the front foil (which has a $\sim 4\ cm^2 sr$ geometry factor). The forward photoelectron yield of a foil that has a C exit surface by Ly-α photons is $\sim$1% (15). The HENA foil comprises three layers: $6.5\ \mu g\,cm^{-2}$ Si, $7.0\ \mu g\,cm^{-2}$ polyimide, and $1.0\ \mu g\,cm^{-2}$ C that has measured 1.5×10^{-3} Ly-α transmittance to reach a conservative design goal of $\leq 0.5-1.5 \times 10^5$ Ly-α induced secondary electrons per second, when the brightest portion of the earth's atmosphere/exosphere fills the FOV. Laboratory calibrations of the heritage INCA instrument have verified this system's capability to continue to image and reject Ly-α background at UV-induced counting rates in excess of $3 \times 10^5\ s^{-1}$. The front foil is mounted on a 70-line/inch nickel mesh (82% transmission), supplied by Buckby-Mears. The stop foil, composed of $7.0\ \mu g\,cm^{-2}$ of polyimide and $5.0\ \mu g\,cm^{-2}$ carbon, will reduce the Ly-α flux to the back, 2-D imaging MCP by about another factor of 100. When the sun is in view, no practical foil composition–thickness combination could be identified to sufficiently suppress it because the quantum efficiencies for secondary electron production are so high at EUV wavelengths. Therefore, to avoid high rates a mechanical shutter is closed once per spin while the sun is in view.

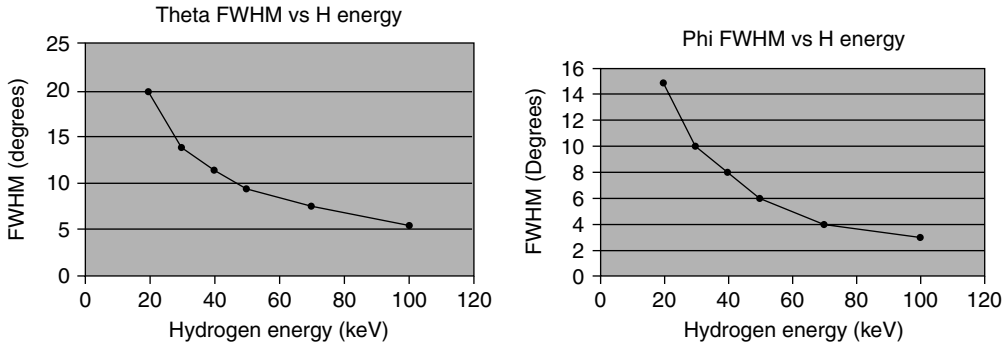


Figure 9. Calibration data that show angular resolution of the IMAGE HENA instrument vs. hydrogen energy (12).

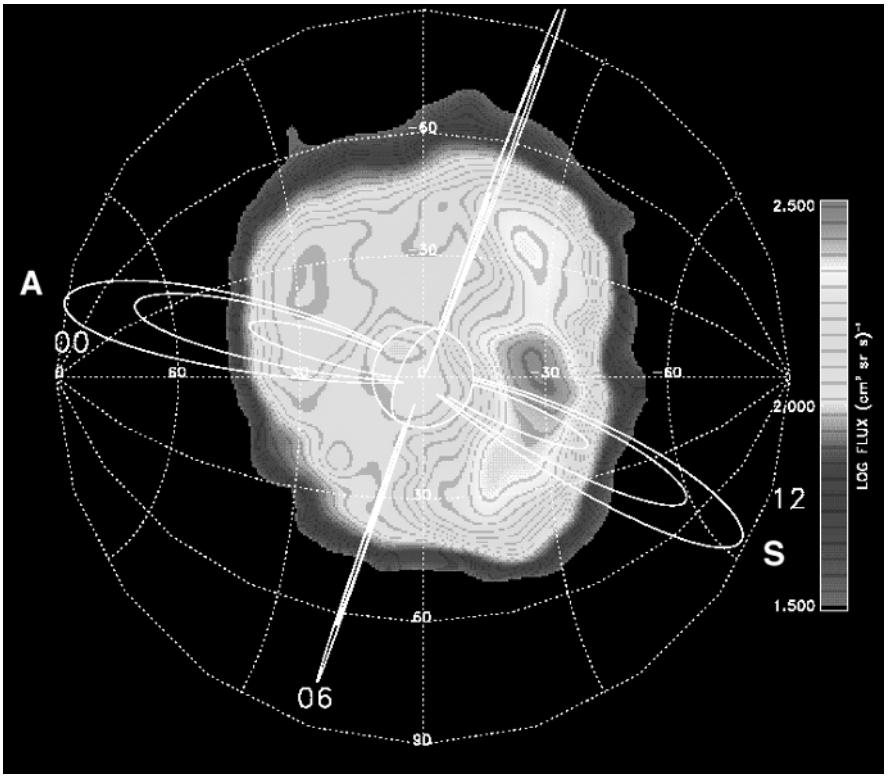


Figure 10. Energetic neutral atom image that shows flux at 39–60 keV during a magnetospheric substorm on 10 June 2000. The view is from the IMAGE satellite located above the North Pole of the Earth (the white circle). Magnetic field lines that have equatorial crossings at four, eight, and twelve earth radii are also shown in white. Noon and midnight field lines are denoted by S and A, respectively. The plot shows a hemispherical fish-eye view.

Angular Resolution and Expected Results

The HENA angular resolution as revealed by calibration generally respond to the measurement requirements of the radiation belt and ring current (see Fig. 9 and Table 2). Typically, the azimuth angle ϕ meets or exceeds those requirements (shown in Fig. 9), whereas the elevation angle θ meets or falls slightly short.

The response of the HENA instrument to a ring-current event on 10 June 2000 is shown in Fig. 10. HENA returns such images every two minutes throughout the two-year IMAGE mission.

ENERGETIC NEUTRAL ATOM IMAGING AT MEDIUM ENERGIES: THE PLASMA SHEET AND CUSP

The plasma sheet and cusp (see Fig. 1) are the most important channels of plasma injection into the inner magnetosphere. The northern and southern cusps are the only certain locations where solar-wind plasma has direct access to the magnetosphere, whereas the plasma sheet is the great reservoir of plasma that is available for injection into the inner magnetosphere during magnetic storms and magnetospheric substorms. Occupying regions of lower magnetic field, these regions have much lower energies than the ring current and radiation belts. The cusp ions have typical energies in the range of 1–2 keV, and the plasma sheet has typical ion energies of about 5 keV.

The requirements for imaging the plasma sheet and cusp are as follows:

1. detect atoms with energies from 1–30 keV whose effective geometric factor is $\sim$0.1 cm^{2}sr,
2. obtain at least crude energy resolution ($\Delta E/E \sim 0.8$) and mass resolution (separate H and O),
3. form images whose a spatial resolution in the ring current and radiation belts is 1 R_E or better,
4. have a wide field of view (extending out to at least a seven earth radii geocentric distance) to encompass the entire ring current and radiation belts in a single "snapshot,"
5. reject bright contaminating emissions such as 121.6 nm from the exosphere and interplanetary medium, and
6. reject energetic ion and electron fluxes.

ENA imaging can be used for both of these regions, and a slit-camera approach such as used for the higher energy ions is feasible. However, the foils used to generate start pulses for 1 keV ions must be much thinner than those in HENA. These ultrathin foils ($\sim$0.5 μ g cm^{-2}) admit too much Lyman α, and the MCPs would be swamped with counts. Thus, an additional element is needed to suppress the ultraviolet radiation by at least a factor of 10^{-5}. This element is a gold UV-blocking grating. The medium-energy neutral atom imager (MENA) on the IMAGE mission is a grating-based slit camera that meets the requirements for imaging the cusp and inner plasma sheet, as well as the low-energy end of the ring current (16).

A schematic view of the MENA instrument is shown in Fig. 11 and the instrument performance data are summarized in Table 3. The neutral atoms first pass through a multielement fan-shaped collimator whose plates carry alternately 0 V and +10 kV to sweep out energetic ions and electrons whose energies are up to a few hundred keV. The atoms then encounter the UV grating (Fig. 12). The MENA gratings were developed and manufactured at MIT using holographic lithography techniques (17). Each grating consists of an ensemble of gold bars, each 400 nm thick and 160 nm wide. The bars lie parallel to one another at a periodic spacing of 200 nm. The $\sim$60-nm spacing between the gold bars is the critical parameter for absorbing of hydrogen Ly α. The gold bars are supported on nickel grids. The grating acts as a strong polarizer and a tuned lossy wave guide. The photon transmission of the gold gratings has been modeled as a function of wavelength and measured in the laboratory (18–20). The results, shown in Fig. 13, indicate a transmission of $\sim 3 \times 10^{-5}$ at hydrogen Ly α (121.6 nm). When convolved with the Ly-α emission profile from the earth's exosphere (21), the total background count rate can be estimated at less than 250 s^{-1}, as shown in Fig. 14. As discussed later, the anticoincidence of the TOF circuitry will reject nearly all of these counts, leaving a residual rate of only $\sim 10^{-3}$ s^{-1} that results from accidental Start/Stop coincidences.

ENAs that pass through the gold grating impact a thin (nominally 0.6 or 0.7 μg/cm^2) carbon foil that has been affixed to the back face of the grating. The purpose of the foil is to generate secondary electrons for a start pulse. The number of secondary electrons emitted provides information regarding the ENA species, as discussed for HENA.

The entire grating/foil assembly for each sensor consists of five carbon foils, gold gratings, and stainless steel grating frames. It is clamped into a lighttight mount in the sensor aperture and electrically biased to -1 kV with respect to chassis ground. Because the secondary electrons that are emitted have a broad angular distribution and energies of a few eV, the -1 kV bias is required to accelerate electrons toward the detector, thereby retaining the z coordinate of the ENA passage through the aperture. A grounded (85% transmission) grid is placed 1 mm downstream from the carbon foil, inside the sensor. Secondary electrons are accelerated to 1 keV through this grid and pass through a nearly field-free, 3-cm region before striking the detector over the Start anode segment. Thus, they provide the TOF Start signal and the z coordinate of detector impact, which is interpreted as the z coordinate of ENA passage through the aperture, or Start position. The pulse height of the MCP charge burst, or Start Height, is measured and recorded to assist in the statistical measurement of PHA for species identification.

ENAs undergo energy- and mass-dependent scattering during passage through the foil, causing some degradation in the determination of polar angle for incident ENAs. Figure 15 contains results of a laboratory study of ENA scattering in a nominal 0.5-μg/cm^2 foil (22). These data may be interpreted directly as a lower bound on the scattering half angle for ENAs that pass through the

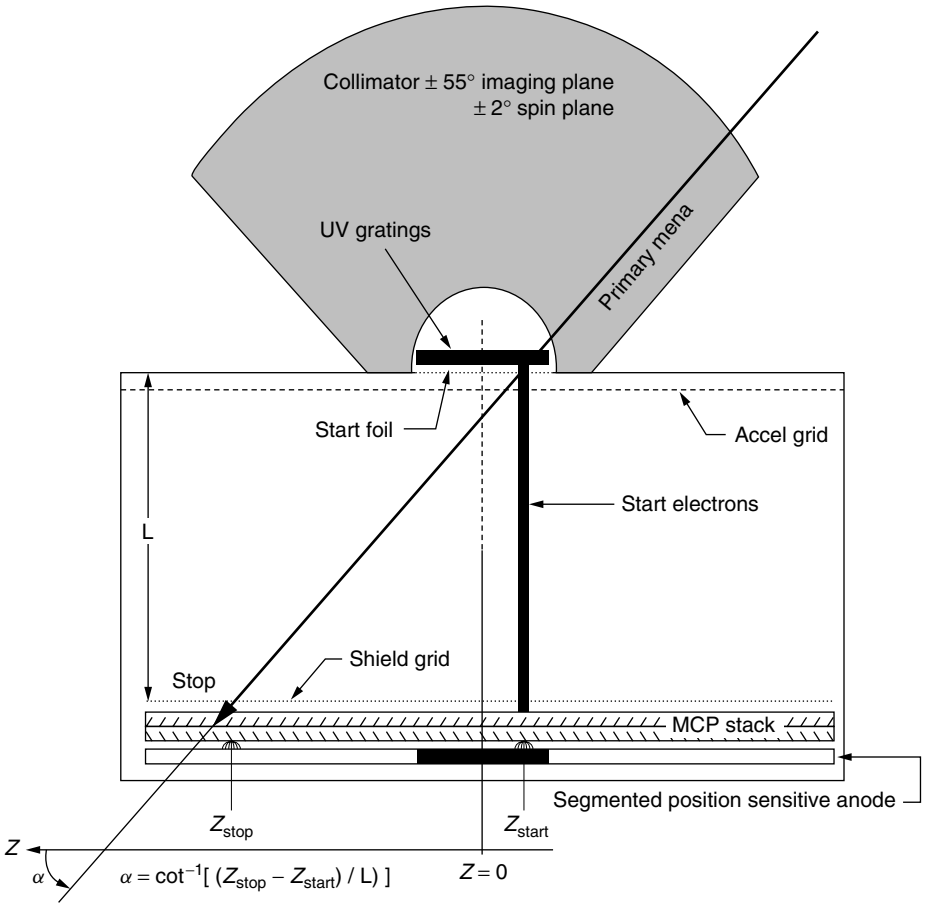


Figure 11. Schematic view of a MENA sensor. A neutral atom passes through the Start foil, producing secondary electrons. Secondary electrons are accelerated toward the Start segment of the detector, whereas the ENA will impact the detector Stop segment. Particles incident on the Start and Stop segments of the detector will provide TOF timing signals, which together with their respective pulse height and ID position on the detector provide the required information for polar incidence angle, energy, and species determination of the ENA (16).

Table 3. MENA Characteristics

Quantity	Requirement	Performance (O)	Performance (H)	Units
Angle FOV				
Azimuthal	360	360	360	Degrees
Polar	90	140	140	Degrees
Angle resolution				
Azimuthal	8	5	5	Degrees
Polar	8	12	8	Degrees
Energy range	1–30	3–70	1–70	keV
Energy resolution	80	80	80	%
Species ID	H, O	Probabilistic	Probabilistic	Unitless
Effective area[a]	1.0	0.83	0.83	Cm^2

[a]Three sensors combined.

MENA aperture and therefore as a lower bound on the (half-width) polar angle resolution achievable using the MENA imager. The angular resolution in the collimated azimuth direction is not affected because the collimator fixes the resolution at approximately 5° FWHM.

An incoming ENA continues (nearly) along its original trajectory and strikes the detector over the Stop anode segment, providing the TOF Stop signal, the z coordinate of ENA impact on the detector (or Stop Position), and the pulse height of the Stop MCP charge burst, or Stop Height. Like the Start Height, the Stop Height is related to the ENA species. The path length of the ENA from the aperture to the detector is $3/\cos\alpha$ cm, where α is the measured polar incidence angle, as shown in Fig. 11. The timing of the Start and Stop signals is compared to produce a time of flight.

Because of the −1-kV accelerating potential applied to the carbon foil, a secondary electron's transit time from foil to detector is ~1.6 ns, assuming an initial secondary electron energy of 2 eV and a 45° emission angle. During this time, an electron will have traveled less than 1 mm transverse to the aperture normal (z direction). This

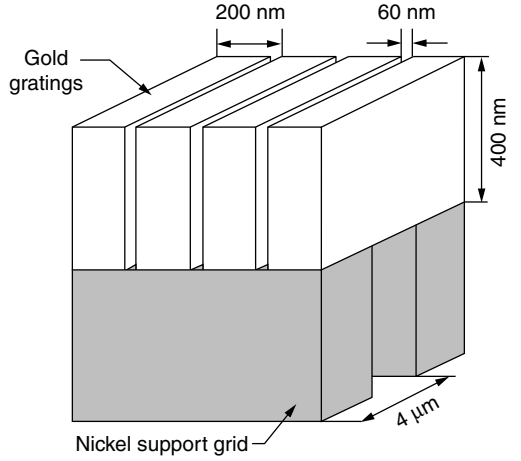


Figure 12. Schematic diagram of UV transmission gratings. The drawing is not to scale. Therefore, disparate dimensions like the 60-nm gaps between gold bars and the 4-μm nickel grid period are displayed on similar scales (16).

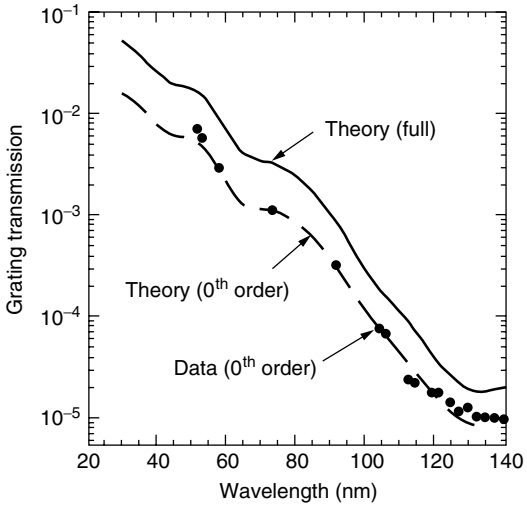


Figure 13. Model and measured Lyα photon transmission through MENA prototype gratings. The solid curve represents model results, points indicate measurements.

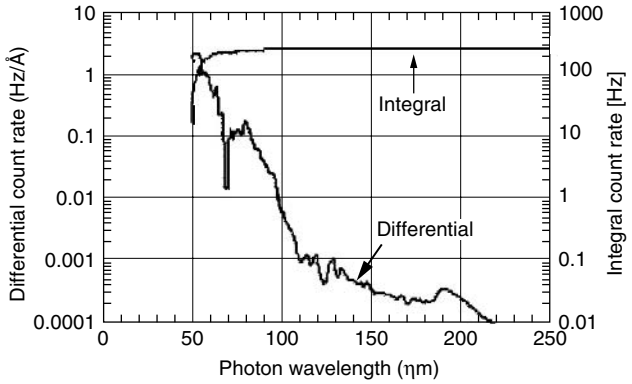


Figure 14. Background noise rate due to geocoronal photons. Shown are the total (Starts + Stops) expected MCP count rates due to UV photons near the IMAGE apogee. Both differential count rates and their integral are shown and labeled (16).

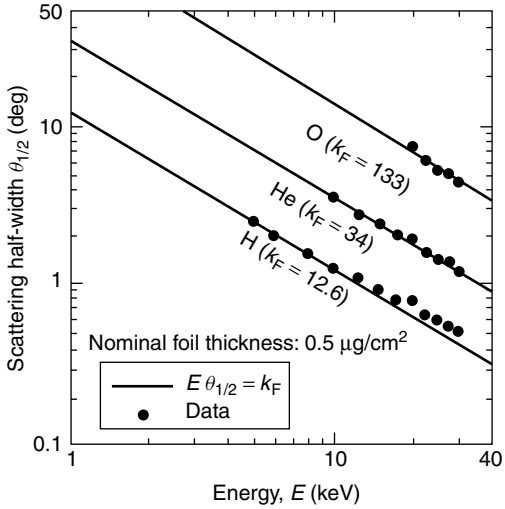


Figure 15. Angular scattering of ENAs upon passage through a nominal 0.5-μg cm^{-2} foil (22).

displacement adds to the uncertainty in measuring of the z coordinate of the ENA as it passes through the aperture, and therefore, in the ENA polar angle which is derived from it.

As in EUV and HENA, the MENA instrument also uses an MCP and position-sensitive anode system. However, because of the mechanical collimation along the slit provided by the collimator, only one-dimensional position sensing is required. Therefore, a spinning platform is required for generating magnetospheric images from MENA.

As shown in Fig. 11, the MENA instrument uses the same MCP for start and stop counts, and the center part is reserved for starts. This simplification leads to a blind spot in the middle of each sensor. All counts in that portion are assumed to be start pulses. At the expected MENA count rates (hundreds of counts/s), such accidental coincidences caused by misidentifying a stop pulse as a start pulse are insignificant. To cover the blind spot will require at least two MENA sensors. For symmetry, three sensors are actually used and are offset by 30° to prevent the blind spots from overlapping.

A flight image obtained on by MENA 10 June 2000 is shown in Fig. 16. Such images are obtained every two minutes throughout the IMAGE mission.

ENERGETIC NEUTRAL ATOM IMAGING AT LOW ENERGIES: IONOSPHERIC PLASMA OUTFLOW

The upper ionosphere is heated by resistive currents (Joule heating) and by wave-particle interactions that are excited by magnetic-field-aligned currents associated with charged particle precipitation and the associated auroral activity. The net result of these magnetosphere–ionosphere interactions is an outflow of ions (primarily oxygen and hydrogen) into the magnetosphere. This source of ions competes on an equal footing with the solar wind as a provider of plasma to the magnetosphere. It is also the major loss process for hydrogen and oxygen,

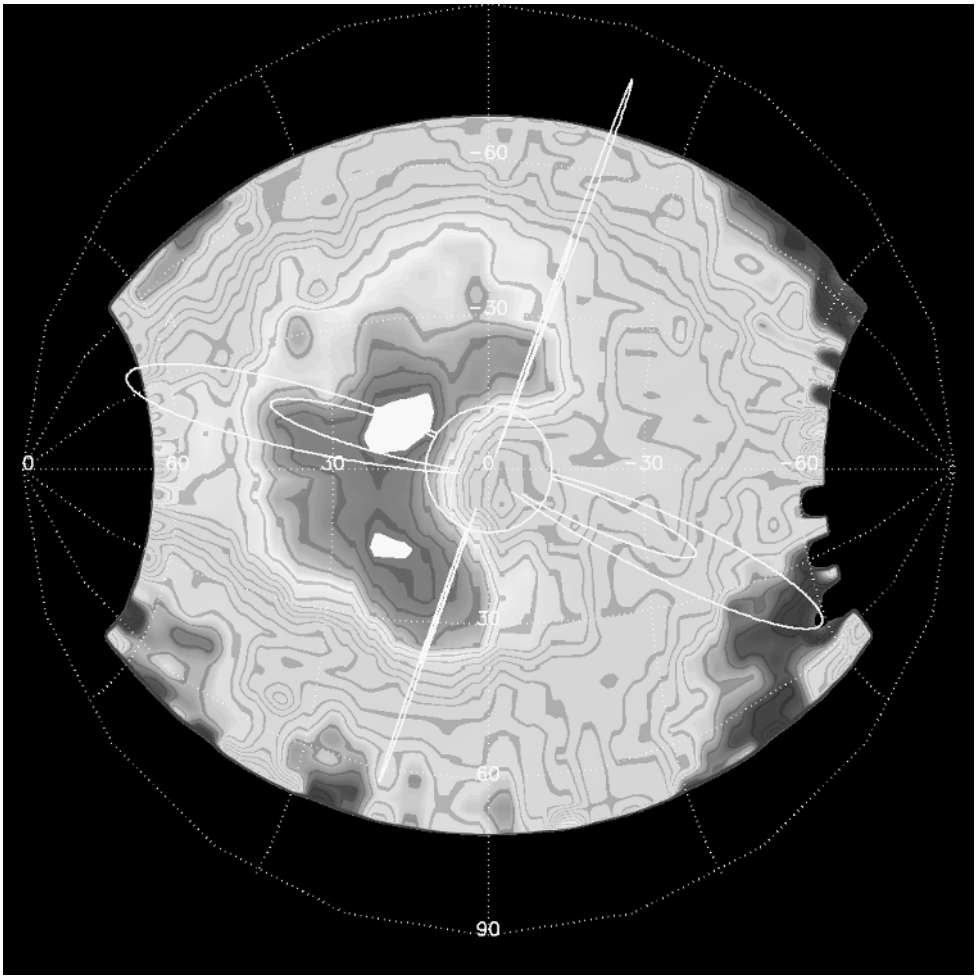


Figure 16. Energetic neutral atom image at 8.6 keV during a magnetospheric substorm on 10 June 2000. Format and time are the same as in Fig. 10.

and hence, water from the earth's atmosphere. It is estimated that the loss rate from ionospheric outflow for O^+ is a few times 10^{25} ions s^{-1} (23,24).

It is known that ionospheric outflows are highly variable in time (25), but it has been difficult to study the variability on appropriate time scales. Most studies have been statistical where time scales range from a few hours to 11 years. Improvement of these time scales requires global imaging of the ion outflow. Because the ions undergo charge exchange with upper atmospheric atoms as they emerge from the ionosphere, global imaging of ion outflows can, in principle, be realized from neutral atom imaging, as for the more energetic magnetospheric ions discussed earlier. However, the thin-foil approach described for the IMAGE HENA and MENA instruments cannot be used within the <1 keV energy range of interest for the outflowing ions because ions at these energies cannot penetrate even the thinnest available foils. Alternatively, incoming neutral atoms can be converted to ions and then analyzed by conventional ion mass spectrometric techniques. Preferably, conversion is accomplished without disturbing the energy, angle, and mass properties of incoming neutral atoms.

The minimum requirements for imaging the ionospheric outflow are as follows:

1. image neutral atoms whose energies range from 10 eV to several hundred eV, and separate high and low energies;
2. field of view of at least two earth radii from a geocentric distance of two earth radii (90° full field);
3. have at least 100 pixels within the field of view; and
4. separate hydrogen and oxygen.

The low-energy neutral atom imager (LENA) on the IMAGE spacecraft meets these requirements (26). The LENA sensor depends for its operation on converting incident neutral atoms into ions that retain the general characteristics of their parent atoms. The ions are velocity-analyzed by conventional techniques of electrostatic and time-of-flight analyses, then detected and imaged by conventional charged particle techniques. The method for the neutral-to-ion conversion is charge exchange upon reflection from a solid surface. A fraction of the incoming neutral atoms is converted to negative ions that are nearly specularly reflected from the surface, then accelerated and collected by an ion extraction lens that focuses and

disperses them in energy. They are passed through a broom magnet that removes electrons and an electrostatic analyzer that maintains energy dispersion while removing photons; then they are delivered at high energy to a time-of-flight sensor and 2-D imaging system that sorts the particles by polar angle of arrival and energy. LENA performance and resource requirements are summarized in Table 4.

Exterior to the conversion surface, LENA is a pinhole camera that has a ~ 1 cm^2 pinhole. Figure 17 illustrates the main components of LENA and the optics path through the instrument. An optically polished near-conical conversion surface (consisting of four facets 22.5° wide each) intercepts incoming neutrals passed by the collimator/charged particle rejecter at an incidence angle of 75° from normal. The conversion surface is composed of four flat facets, each a trapezoid of optically polished tungsten. The entire structure that supports the conversion surface is floated at a high negative potential in the range from 15–20 kV. The conversion product negative ions (and any secondary electrons produced on the surface) are accelerated away from the conversion surface support

Table 4. LENA Characteristics

Parameter	Value
Energy range (incident neutral atom)	15–1250 eV
Energy resolution	$E/\delta E = 1$ at FWHM
Mass range	1–20 amu (H^+ and O^+)
Mass resolution	$M/\delta M = 4$ at FWHM
Angular coverage	Sampling: $8° \times 8° \times 12$ sectors per spin: $360° \times 90°$ in 45×12 samples
Angular resolution/response	$8° \times 8° = 0.02$ steradians
Total field of view	$2.8\ \pi$ sr
Pixels per image	3.2 $k = 3$ Energy $\times 45$ Az $\times 12$ Polar $\times 2$ Mass
Pixel physical aperture	1.0 cm^2 ($A_{\text{eff}} \leq 1 \times 10^{-3}$ cm^{-2})
Pixel solid angle	0.02 sr per pixel $\times 12$ pixels
Time resolution 1-D polar × energy	2.7 s
Time Resolution 2-D (az × polar) × energy	120 s
Time Resolution 3-D (az × polar × energy)	1 spin period (120 s)
Dynamic range	10^4

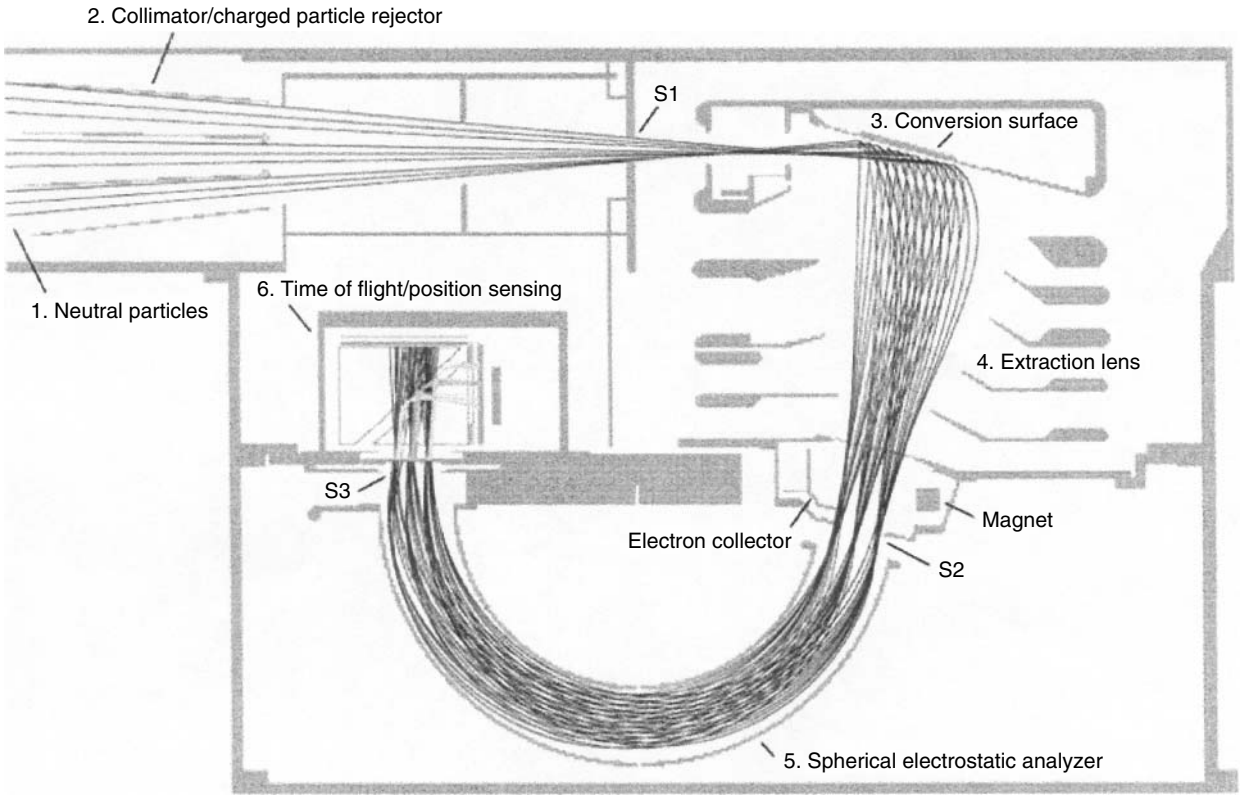


Figure 17. Two-dimensional cross section of the LENA imager optics (26).

structure and collected by the ion optics system. The surface converts incident atoms to negatively charged ions of corresponding species (or in some cases sputters surface adsorbates, again as negatively charged ions).

The emitted negative ions are accelerated, focused, and dispersed in energy by an ion extraction lens. The lens can be trimmed to selected energy ranges from 10–300 eV and 25–750 eV. A broom magnet removes photoelectrons and secondary electrons if they fall within the LENA energy range. Negative ions then encounter a spherical electrostatic analyzer that maintains the radial energy dispersion of the ions. The main purpose of the electrostatic analyzer is to remove residual UV photons and present the negative ions to the imaging time-of-flight (ITOF) system for mass, energy, and polar angle analysis. The ions arrive at the ITOF system with energy that corresponds to the optics high voltage, nominally 20 keV, but their original energies can still be derived from the radial dispersion, which is preserved through the spherical analyzer.

The ITOF system consists of carbon foil apertures, an electrostatic mirror system that reflects secondary electrons emitted from the back sides of the foils, and four sets of start and stop detectors (one for each conversion surface facet). The detectors are chevron stacks of microchannel plates whose length/diameter = 60. The start stacks are rectangular, and the stop stacks are trapezoidal corresponding to projections of the conversion surface facets. Start event charge clouds fall directly upon a 2-D wedge and strip anode (27) that has four preamplifiers; the ratio of their pulses define radial and polar angle coordinates for the events. The preamplifiers are located within the TOF detection unit, and pulses are passed to a position sensing system. Start and stop events generate fast pulses regardless of their point of impact on the detector system. These pulses are passed directly from the fast pulse anodes to the TOF electronics unit.

The response of LENA to an incident neutral atom is governed by the dimensions of the collimator, the entrance aperture slit, the conversion surface, the ion optics, the ITOF optics, and finally the response of the detectors. The response is fully described by the product of the effective aperture area ΔA, the pixel solid angle $\Delta\Omega$, and the energy range ΔE for each accumulation. In practice, these quantities may be functions of the incident particle species or energy and of internal potentials applied to the optics elements or to the MCP detectors.

Angular Response

LENA is designed to have a 90° polar × 8° azimuthal acceptance angle about the entrance aperture. This angle is defined geometrically by the collimator and by the dimensions of the conversion surface. The polar angle is subdivided into twelve 7.5° polar bins, and binning of azimuthal angle and imaging of the neutral atoms is achieved by rotating of the spacecraft. Conversion of neutral atoms takes place on four distinct tungsten surfaces equally spaced across the polar angle dimension. Each surface is associated with distinct microchannel plate assemblies for both the start and stop sections of the TOF system, as well as distinct carbon foils. The relative efficiency of detecting neutral atoms within any one of these four angular sections is constant, but the efficiency among the sections varies by as much as a factor of 3.

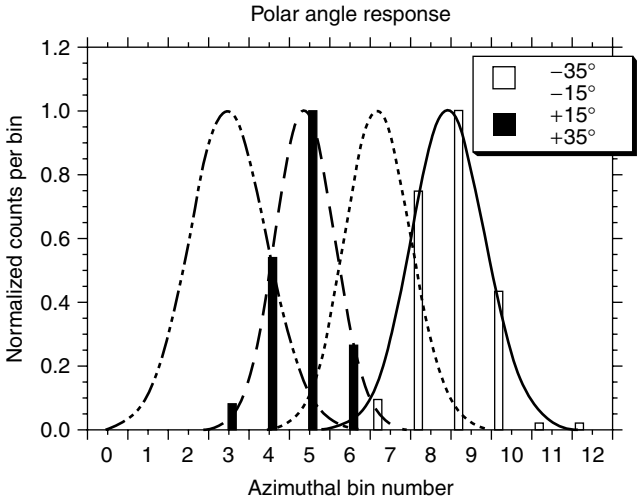


Figure 18. Response function of IMAGE-LENA in polar angle for several fixed incoming beam directions (26).

The angular resolution is ultimately limited by angular scattering of the atoms on the conversion surface, optical aberrations in the ion extraction lens, scattering of the detected secondary electrons from the carbon foil, and measurement error on the position-sensitive detector. These aberrations lead to a dispersion in the angular response, as illustrated in Fig. 18.

Energy Response

LENA was designed to operate across neutral atom energy ranges of 10–300 eV and 25 eV–750 eV, depending on the potential placed on the steering electrode. Within these ranges, these measured energies correspond to those of the scattered negative ion from the conversion surface. Particles incident on the surface lose energy across a broad distribution which peaks at about 80% of the incident energy for oxygen and 95% for hydrogen. In addition, low-energy negative ions, sputtered from the surface by the incident neutral particle, contribute to the signal. The ion extraction lens on LENA is designed to focus and disperse the scattered ions, according to energy, across the carbon foil detector plane on the TOF analyzer (Slit S3 in Fig. 17). The location of the incident ions on the carbon foil is determined by imaging the secondary electrons from the carbon foil onto a position-sensitive detector. In this way, measurements of both the energies and angles of the ions are made simultaneously. The nominal energy bins (at 20-kV optics potential) are centered at ~50, 150, and 250 eV (ion energy) at the lowest steering potential. At the highest steering potential, the bins are shifted to ~300, 500, and 700 eV, respectively. These energy ranges scale approximately with the optics potential in use at any given time.

Time-of-Flight Response

LENA includes a carbon foil TOF analyzer designed to separate converted hydrogen from converted heavy ions.

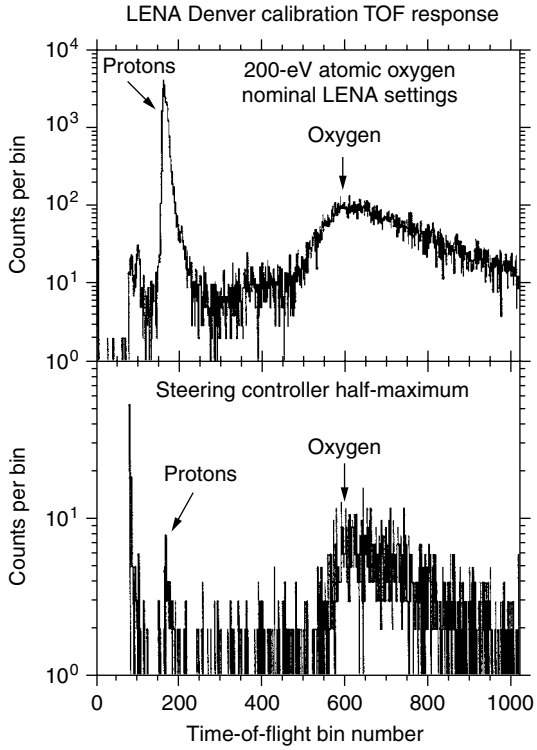


Figure 19. LENA time-of-flight responses to oxygen, including sputtered H atoms (top panel), and to oxygen where sputtered atoms are removed (bottom panel) (26).

An example of the TOF spectrum is shown in Fig. 19. The spectrum shows clear separation of the well-peaked H atom signal and the O atom signal which is more broadly distributed due to straggling in the foils. This spectrum was taken from an incident oxygen atom beam, but the spectrum includes an H atom signal due to sputtering from the tungsten surface. Because the sputtered atoms come off the surface with low energy relative to the energy of the scattered atoms, it is possible to remove most of the sputtered atoms from the detector by using the steering electrode. This is shown in Fig. 19, where the ratio of hydrogen to oxygen changes from 1.78 in the upper panel to 0.07 in the lower panel. When the sputtered ions are not removed, compositional information is lost, but the sensitivity of the instrument is improved.

An example of the LENA flight data is shown in Fig. 20, in which the horizontal axis is the 90° instrument polar angle and the vertical axis is the 360° spin angle of the IMAGE spacecraft.

FAR-ULTRAVIOLET IMAGING OF THE PROTON AND ELECTRON AURORAS

The aurora is an invaluable diagnostic of the state of the magnetosphere. Auroral emissions are excited predominantly by precipitating energetic electrons, and it is the electron aurora that has been extensively imaged in visible, ultraviolet, and X-ray wavelengths. Precipitating protons also produce auroral emissions, which have proven far more challenging to image. Auroral Lyman-α emissions of neutral hydrogen are produced by charge exchange of energetic protons that cascade into the atmosphere (28–31). As the protons penetrate into the atmosphere and collide with atmospheric gas, they capture electrons from the major neutral constituents (N_2, O_2, O) through charge exchange reactions. They become fast neutral hydrogen atoms, some of which will radiate the Doppler-shifted Lyman-α emission. In subsequent collisions, the energetic H atoms may lose their electron in an electron stripping collision and reappear as protons whose motion is constrained by the magnetic field. The cycle will repeat hundreds to tens of thousands of times, depending on the primary proton energy. The difficulty in imaging the precipitating energetic protons stems from the fact that they emit photons only when they gain an electron in an excited state and then transit to the ground state, thereby radiating at the same Ly-α wavelength as the cold hydrogen atoms of the exosphere. Imaging is possible

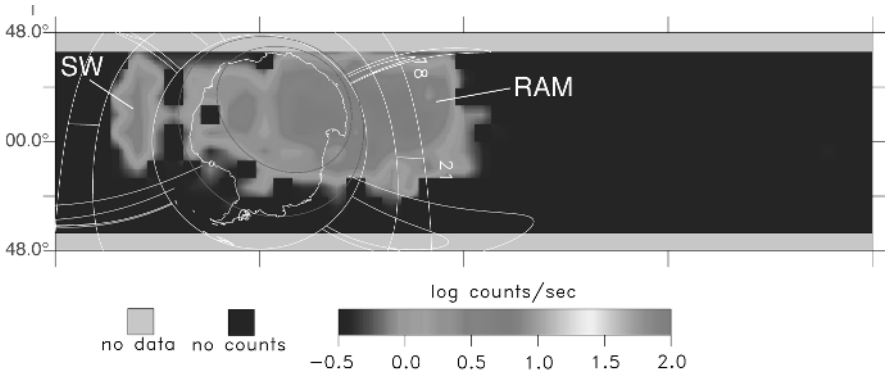


Figure 20. Low-energy neutral atom image obtained by the IMAGE LENA instrument on May 25, 2000. Vertical axis is instrument polar angle. The image was acquired when the spacecraft was near perigee over the southern hemisphere. The white circle indicates the Earth; the outline of Antarctica is shown. Also shown are representative magnetic field lines (in white), local time sectors, and the nominal auroral oval (in red). The energy range of the neutral atoms is 10 eV to 1000 eV. The fluxes to the left (labeled SW) are believed to originate in the solar wind, perhaps through charge exchange in the distant part of the Earth's exosphere. The fluxes to the right are due to outflowing ionospheric ions. (Image from T. E. Moore et al., *Geophysical Research Letters*, Vol. 28, No. 6, 1143–1146, March 15, 2001. Copyright 2001 American Geophysical Union. Reproduced by permission of the American Geophysical Union.)

because of the red shift of the energetic H emissions when viewed from above, but this red shift amounts to only a few tenths of a nanometer.

A similar spectral resolution requirement exists for imaging the electron aurora. In this case, the actual atmospheric emissions produced by bombarding oxygen and nitrogen atoms and molecules can be imaged in the ultraviolet. However, airglow emissions at the same wavelengths are produced by atmospheric chemistry, and these emissions compete favorably with the auroral emissions, particularly on the dayside and for weak auroras (32). Although the brightest auroral and airglow line is the 130.4-nm OI emission, it is multiply scattered in the atmosphere, making it difficult to derive accurate two-dimensional distributions of the auroral emissions. On the other hand, the 135.6-nm line is scattered only to a limited degree (33) and is an excellent emission feature for imaging the aurora that can be related to the total electron precipitation. Thus, a primary measurement requirement for UV imaging is the detection and spectral separation of the 135.6-nm emission from that of 130.4 nm. This requirement necessitates using a spectrometer because narrowband filter technology cannot satisfy the requirement.

The requirements for imaging the global proton aurora and the electron aurora at 135.6 nm are

1. image the entire auroral oval over a 5-minute period at a spatial resolution of <100 km;
2. spectrally separate the energetic proton precipitation (<1 kR of Doppler-shifted Lyman-α emissions near 121.8 nm which corresponds to a proton energy of 8 keV) from the intense, cold geocorona (>10 kR of Lyman-α emissions at 121.567 nm); and
3. separate spectrally the 135.6-nm OI emission from the multiply scattered 130.4-nm emission.

These requirements are satisfied by the Far Ultraviolet Spectrographic Imager (FUV-SI) on the IMAGE spacecraft (34).

The first requirement forces a relatively large field of view: $15^\circ \times 15^\circ$. The second requirement necessitates maximizing the spectral resolution within the available resources; an intrinsic spectral resolution of 0.2 nm was chosen to separate the Doppler-shifted auroral Lyman-α from the exospheric background. Because the Doppler-shifted Lyman-α signal is about 100 times less intense than the Lyman-α exospheric emissions at 121.6 nm, internal scattering in the instrument must be minimized. To be able to interpret the Doppler-shifted auroral Lyman-α measurements in terms of pure proton aurora requires rejecting of nitrogen emissions near 120 nm (triplet lines at 119.955, 120.022, and 120.071 nm) that are produced by all auroras, including the electron aurora. The third requirement provides that the transmission of the 130.4-nm triplet line should be less than 1% of the 135.6-nm line, necessitating an overall spectral resolution less than 3 nm in the 130–135 spectral region. The spatial resolution requirement specifies that, from the IMAGE satellite apogee of seven earth radii, the angular resolution should

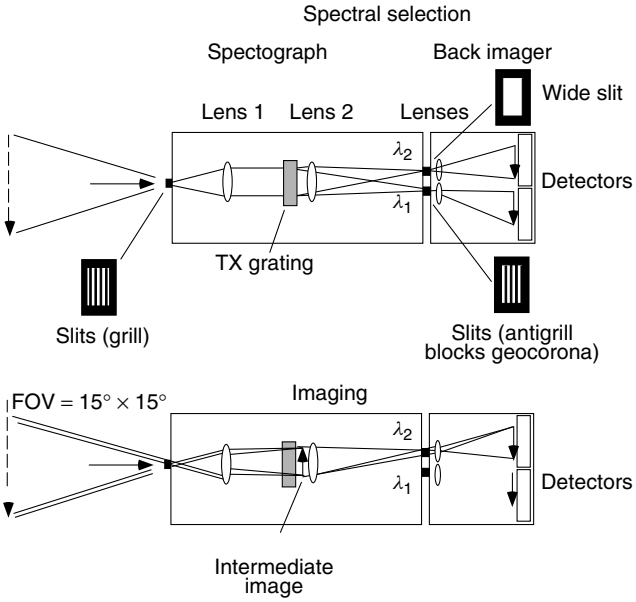


Figure 21. Simplified schematic illustrating the operating principle of the FUV-SI as an imaging monochromator. The optical elements are represented here by lenses. Mirrors are used in the actual instrument. The top panel shows the wavelength selection; the bottom panel illustrates the imaging function (34).

be 7×7 arcmin for each of the 128 pixels $\times$ 128 pixels within the total field of view.

The FUV-SI is essentially an imaging monochromator, whose principle of operation is illustrated in Fig. 21. The top diagram in Fig. 21 shows the wavelength selection, and the bottom part shows the imaging properties of such an instrument. For simplicity, refractive lens elements are used to represent the optical components. In the actual FUV-SI instrument, all optical elements are mirrors that pass far-ultraviolet radiation and are achromatic across the wavelength range covered. Consequently, the optics is folded, resulting in a more complicated arrangement. The first element in the imaging monochromator is an input aperture (a grill of slits in the case of FUV-SI) followed by a collimating lens that produces parallel light from any point in the plane of the slits. The grating in this illustration is a transmission grating placed at a distance from the collimating lens, so that any parallel light coming through the slit is focused to a point at the grating. Thus, an image of the object at infinity in front of the instrument is formed close to the grating. The grating disperses the light according to wavelength. The wavelength-related angular dispersion of the light is the same for all points of the image. Thus, to a first order, imaging and wavelength selection are independent of each other. The grating is followed by a spectrographic camera lens that focuses parallel light from the grating at the plane of the exit slit. A spectral region defined by the position and size of the slit(s) receives an equal amount of light from any point in the image. The back imager, the last part of the system, focuses the intermediate image, located at the grating, on the detector. The location and width of the exit slit(s) define the spectral region and the spectral width of the light that passes through the

system. Therefore, the spectrometer acts as an afocal filter in the optical system. Two wavelength regions correspond to two separate exit slits. One region is for imaging the Doppler-shifted Lyman-α radiation (121.8 nm) induced by the proton aurora, and the other is the 135.6-nm oxygen emission. The two regions in Fig. 21 are denoted as λ_1 and λ_2.

The bottom half of Fig. 21 shows the imaging properties of the imaging monochromator. Parallel light that enters the entrance slit (grill) from an object is focused at the plane of the grating. This intermediate image is reimaged on the detectors by the back imager camera lens located behind the slit(s).

The most remarkable property of this type of imaging monochromator is that its imaging and wavelength properties are independent of each other because light rays that leave from a point of the intermediate image at the grating are brought in focus on the detector, regardless of their direction or their associated wavelength. Different wavelengths are diffracted in different directions. Filtering is accomplished by the size and position of the slit(s). In this system, it is possible, for example, to have a wide exit slit that passes a large spectral region combined with a sharp focus on the detector that represents high spatial resolution. Thus, wide spectral passbands can be accommodated by high spatial imaging resolutions. Alternatively, it is possible to use very fine spectral slits and integrate over large detector pixels.

It is also possible to have several separate wavelength channels using this approach. The arrangement needs separate back imager optics for each wavelength channel, and each separate back imager optics has to provide some deflection to ensure that the wavelength regions end up on separate detectors. For example, the two small reimaging lenses in Fig. 21 must be separate to ensure that two passbands λ_1 and λ_2 form images at different locations and, if necessary, on separate detectors.

To have a fairly large throughput but high enough spectral resolution, the FUV-SI instrument uses a wide input slit that is subdivided into nine subslits to form a slit grill (35). The output slit of the Lyman-α channel also contains a similar slit grill that has been designed to block the light at the precise wavelength of the more intense, cold, geocoronal emission but transmit efficiently in the wavelength region that contains the weaker Doppler-shifted emissions. The nine parallel slits improve light collection by a factor of 9, so, that the FUV-SI is equivalent in efficiency to nine single-slit instruments of the same size.

The IMAGE FUV-SI instrument is schematically illustrated in Fig. 22. The light enters the spectrometer through the entrance slit; is reflected by the collimator mirror, which produces parallel light from any point on the slit; and focuses it as an intermediate image approximately at the grating. The function of the two back imagers is to transfer the intermediate image onto two spatially separated detectors for 121.8 nm and 135.6 nm. The pattern of slits for the 121.8-nm back imager is designed to be in anticoincidence with the input slit grill. The slit for the back imager labeled 135.6 is simply a wide planar slit because a 3-nm spectral bandwidth is adequate to discriminate against 130.4-nm emissions. To satisfy the spectral resolution requirements of 0.2 nm without lengthening the spectrometer, at least two surfaces are needed to provide sufficient aberrational correction for the spectrograph. The instrument uses Wadsworth mounting in which a collimating mirror forms a parallel beam for a separate grating. Minimizing the aberrations also demands using the central, or paraxial, field of the collimator. This dictates using a system configuration in which the incoming light is on axis and passes through a central hole of the grating. The SI has a 3,000-line per mm holographically ruled grating 181 $\times$ 156 mm (Table 5).

The IMAGE satellite is a slow spinner (2-minute spin period), so at apogee the earth is within the field of view of the imager for less than 10% of the spin. To maximize the sensitivity of the instrument, a time-delay integration (TDI) scheme is used in which multiple images taken as the earth passes through the field of view are spatially registered with respect to the earth and are overlapped

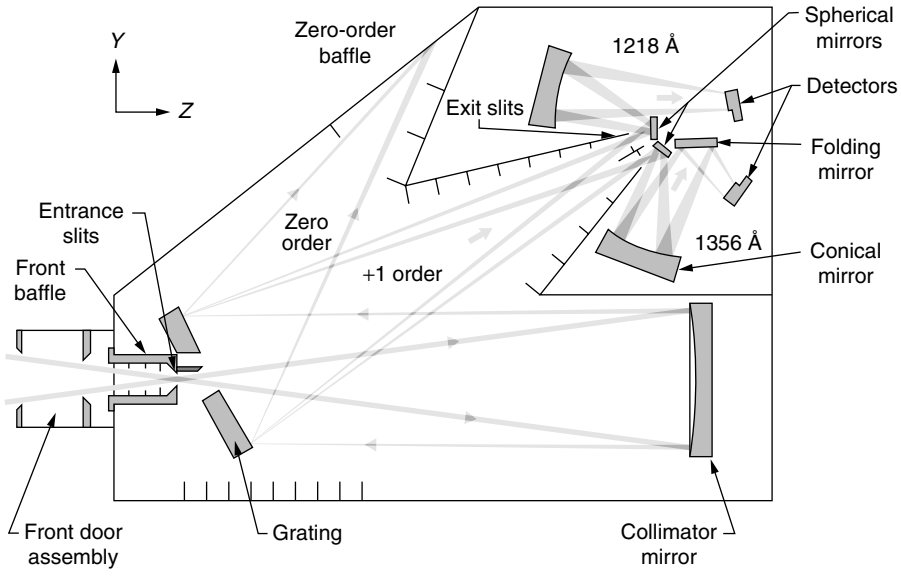


Figure 22. Schematic of the FUV-SI instrument (34).

Table 5. FUV SI Characteristics

Mechanical Configuration	Entrance Slits	Exit Slits
Overall slit width (μm)	391.6	356.5
Spacing (μm)	1076.6	980.1
Curvature	No	$R = 1300$ mm

Optical Equivalent	Wavelength
Slit width	0.19 nm
Bar width	0.3323 nm
Slit spacing	0.5223 nm
Linear dispersion	1820 μm/nm

Optical Design	Radius of Curvature (mm)	Conic Constant	Offset (mm)	Size (mm) x	y
Collimator mirror	1000 (concave)	−1.727	0	181	156
Grating	1000 (concave)	0	0	142	164
Spherical mirrors	136.461 (convex)	0	0	52	20.25
Conical mirrors	170.725 (concave)	0.114	53.5	144	76
Flat 135.6 nm	—	—	—	72	45.5

in the instrument memory. An advantage of the spinning platform is that the blind spot produced by the hole in the grating (a requirement of the paraxial Wadsworth configuration) is partially filled in.

For ease of manufacturing, two identical imagers were used: one works at 121.8 nm, the other one works at 135.6 nm. Each imager is a two-mirror optical system. The first mirror is spherical. The second is conical (elliptical) and works in an off-axis configuration. The two imagers are perfectly symmetrical (only the back focal length is slightly different, owing to a geometric difference in the intermediate image to be relayed). Thus two identical spherical and two identical conical mirrors were manufactured. A flat folding mirror was added on the 135.6-nm imager, so that the two detectors would fit in the available space. The imagers are not optimized independently of the Wadsworth monochromator. Astigmatism, which was not critical for spectral selection at the exit slit plane, was the main aberration to be corrected by the imagers to generate the best possible image on the detectors.

The FUV-SI instrument is designed to minimize the 120- and 121.6-nm lines while transmitting the largest bandwidth in the region from approximately 121.7–122.3 nm. The grating angle and the collimator conic constant are optimized to reduce aberrations at 121 nm, that is, between the 120–121.6-nm lines. It was relatively easy to meet the spectral resolving power requirements of the instrument in the 130 to 135-nm

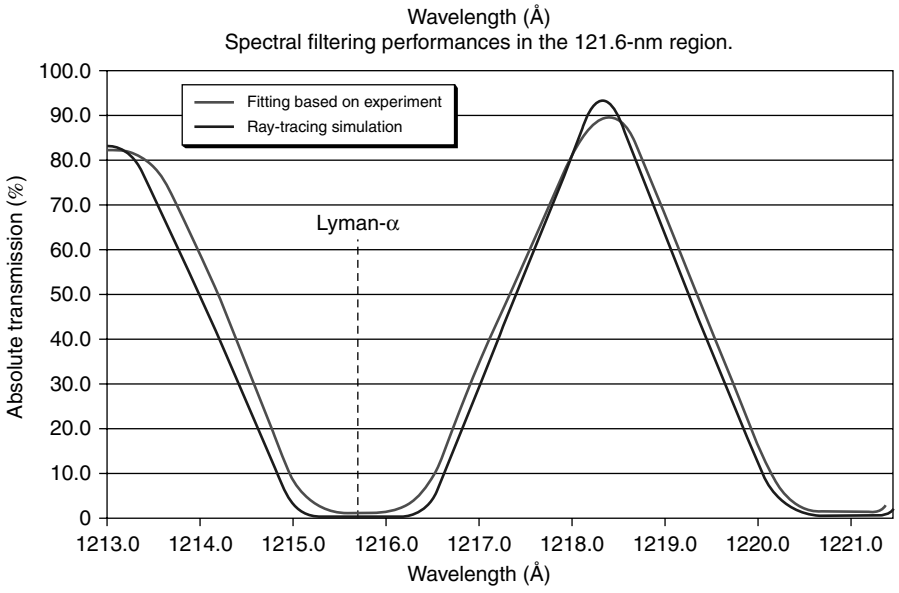


Figure 23. Laboratory measurement of the spectral performance of the FUV-SI instrument for the proton aurora (34).

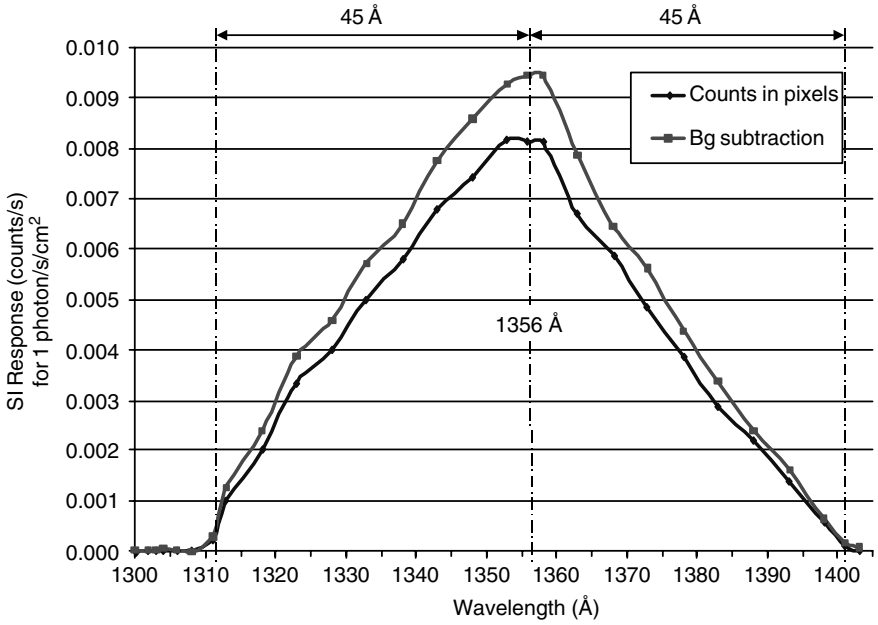


Figure 24. Laboratory measurement of the spectral performance of the FUV-SI instrument for the electron aurora (34). Bottom curve is the response after the background (Bg) has been subtracted.

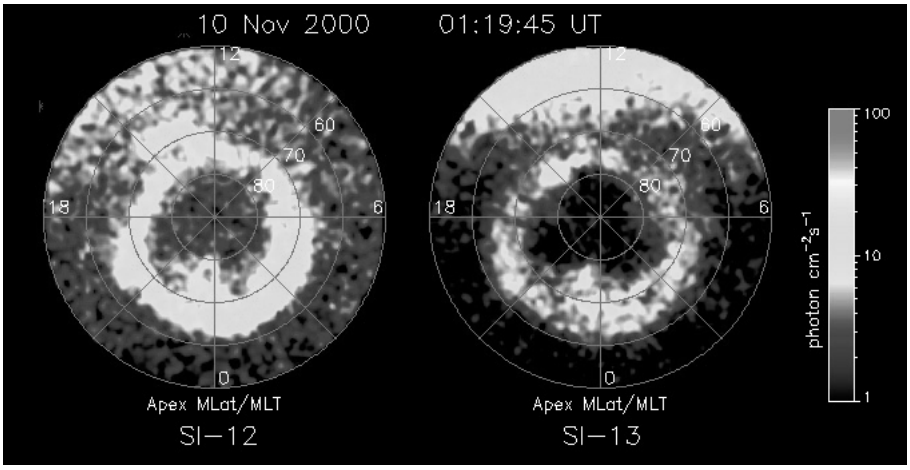


Figure 25. Auroral images obtained with the IMAGE FUV-SI instrument on November 10, 2000. The auroral emissions are mapped in magnetic latitude and magnetic local time, with local noon at the top and midnight at the bottom. The image on the left is the proton aurora; the image on the right is the electron aurora. The proton aurora image shows an emission patch equatorward of the main oval in the post-noon afternoon sector that is not seen in the image of the electron aurora. See color insert.

region to accomplish 130.4-nm blocking for the second imager.

FUV-SI Instrument Performance

The spectral performance of the FUV-SI, as measured in the laboratory, is shown in Fig. 23 and Fig. 24 for the proton and electron aurora, respectively. A worst-case Lyman-α leak of 1.3% is derived from the data, which are limited by the fact that the calibration monochromator had a lower spectral resolution than that of the instrument. Figure 24 shows excellent rejection of 130.4-nm radiation and a peak response at 135.6 nm, as desired. The absolute response for a 5-second exposure at the two wavelengths of interest was measured at 1.6×10^{-2} counts per Rayleigh per pixel and 1.8×10^{-2} counts per Rayleigh per pixel, for 135.6 and 121.8 nm, respectively.

Flight images of the electron and proton aurora obtained by the FUV-SI instrument are shown in Fig. 25.

SUMMARY AND CONCLUSIONS

Magnetospheric imaging is an emerging field that has now just been fully implemented. The Imager for Magnetopause-to-Aurora Global Exploration (IMAGE) spacecraft was launched in March 2000 and since then has been sending back global images of the plasma regions in the earth's inner magnetosphere. This spacecraft carries instruments that perform all of the known techniques for magnetospheric imaging. That is why a review of the state of the art in magnetospheric imaging is essentially a review of IMAGE instrumentation. One of the techniques employed on the IMAGE mission — energetic neutral atom imaging — will also be used to image the plasma environments of other planets. The INCA instrument on the Cassini Saturn Orbiter (36) — the predecessor of the IMAGE HENA imager — will provide ENA images of Saturn's inner magnetosphere after Cassini arrives in the Saturn system in the summer of 2004. ENA imaging of the interaction of the solar wind with the Martian

exosphere (37) is also a key objective of the European Mars Express mission, which is scheduled for launch in 2003. The application of ENA and other magnetospheric imaging technologies in the twenty-first century promises exciting advances in our understanding of terrestrial and planetary space environments.

BIBLIOGRAPHY

1. M. Nakamura et al., *Geophys. Res. Lett.* **27**, 141–144 (2000).
2. E. C. Roelof, *Geophys. Res. Lett.* **14**, 652–655 (1987).
3. M. G. Henderson et al., *Geophys. Res. Lett.* **24**, 1,167–1,170 (1997).
4. M. G. Henderson et al., *Adv. Space Res.* **25**, 2,407 (2000).
5. E. C. Roelof and A. J. Skinner, *Space Sci. Rev.* **91**, 437–460 (2000).
6. J. Perez, M. -C. Fok, and T. E. Moore, *Space Sci. Rev.* **91**, 421–436 (2000).
7. S. B. Mende et al., *Space Sci. Rev.* **91**, 287–318 (2000).
8. J. L. Burch, *Space Sci. Rev.* **91**, 1–14 (2000).
9. B. W. Reinisch et al., *Space Sci. Rev.* **91**, 319–359 (2000).
10. B. R. Sandel et al., *Space Sci. Rev.* **91**, 197–242 (2000).
11. S. Lunt and R. S. Turley, *Proc. Utah Acad.* submitted (1999).
12. D. G. Mitchell et al., *Space Sci. Rev.* **91**, 67–112 (2000).
13. D. G. Mitchell et al., in R. F. Pfaff Jr., J. E. Borovsky, and D. T. Young, eds., *Measurement Techniques in Space Plasmas: Fields*, American Geophysical Union, Washington DC, 1998, pp. 281–287.
14. R. L. Rairden, L. A. Frank, and J. D. Craven, *J. Geophys. Res.* **91**, 13 613–13 630 (1986).
15. K. C. Hsieh, E. Keppler, and G. Schmidtke, *J. Appl. Phys.* **51**, 2,242 (1980).
16. C. J. Pollock et al., *Space Sci. Rev.* **91**, 113–154 (2000).
17. J. T. M. Van Beek et al., *J. Vac. Sci. Technol.* **B16**, 3,911–3,916 (1998).
18. M. A. Gruntman, *Appl. Opt.* **34**, 5,732–5,737 (1995).
19. E. E. Scime, E. H. Anderson, D. J. McComas, and M. L. Schattenburg, *Appl. Opt.* **34**, 648–654 (1995).
20. M. M. Balkey, E. E. Scime, M. L. Schattenburg, and J. van Beek, *Appl. Opt.* **37**, 5,087–5,092 (1998).
21. R. R. Meier, *Space Sci. Rev.* **58**, 1–186 (1991).
22. H. O. Funsten, D. J. McComas, and B. L. Barraclough, *Opt. Eng.* **32**, 3,090–3,095 (1993).
23. C. J. Pollock et al., *J. Geophys. Res.* **95**, 18 969–18 980 (1990).
24. T. E. Moore et al., *Science* **277**, 349–351 (1997).
25. T. E. Moore et al., *Geophys. Res. Lett.* **26**, 1–4 (1999).
26. T. E. Moore et al., *Space Sci. Rev.* **91**, 155–195 (2000).
27. D. M. Walton, A. M. James, and J. A. Bowles, in R. F. Pfaff Jr., J. E. Borovsky, and D. T. Young, eds., *Measurement Techniques in Space Plasmas: Particles*, American Geophysical Union, Washington DC, 1998, pp. 295–300.
28. R. H. Eather, *Rev. Geophys. Space Phys.* **5**, 207–285 (1967).
29. B. C. Edgar, W. T. Miles, and A. E. S. Green, *J. Geophys. Res.* **78**, 6,595–6,606 (1973).
30. B. Basu, J. R. Jasperse, D. J. Strickland, and R. E. Daniell, *J. Geophys. Res.* **98**, 21 517–21 532 (1993).
31. D. J. Strickland, R. E. Daniell, J. R. Jasperse, and B. Basu, *J. Geophys. Res.* **98**, 21 533–21 548 (1993).
32. D. J. Strickland, J. R. Jasperse, and J. A. Whalen, *J. Geophys. Res.* **88**, 8,051–8,062 (1983).
33. D. J. Strickland and D. E. Andersen Jr., *J. Geophys. Res.* **88**, 9,260–9,264 (1983).
34. S. B. Mende et al., *Space Sci. Rev.* **91**, 287–318 (2000).
35. M. -P. Lemaitre et al., *Science* **225**, 171–172 (1984).
36. S. M. Krimigis et al., *Science*, in press (2000).
37. E. Kallio, J. G. Luhmann, and S. Barabash, *J. Geophys. Res.* **102**, 22 183–22 197 (1997).

MOTION PICTURE PHOTOGRAPHY

THOMAS W. HOPE
Hope Reports
Rochester, NY

It is not the intention of this article on motion picture photography to be all encompassing. Motion picture photography as a professional industry has many facets. In effect, this is a compendium. Subjects pertaining more closely to imaging are covered in greater detail than others not directly involving imaging. For definitions of many terms used here, consult the glossary. Motion picture photography, dating from the 1890s, is one of the oldest of modern imaging, technologies that remains current today.

APPLICATIONS (1)

As a predecessor to television, motion pictures have been a medium that has provided entertainment in movie theaters for more than a century. Many prime time television programs and most national TV commercials are shot on motion picture film. For years, the family recorded its activities in home movies. Motion pictures have played a major role in education from preschool through high school and college. Business and industry have used it for training, sales, and marketing and to provide information to the public as well as to specific audiences such as stockholders. Motion pictures have been a vital element of health sciences for education as well as information. Federal, state, and local governments have used the motion picture for a wide variety of purposes similar to those of industry. Churches and synagogues, youth groups, civic organizations, clubs, retirement and nursing homes all have used movies for teaching, informing, and entertaining. Today many of the functions listed here are carried out on videotape and, more recently, by digital technology using the computer to combine multimedia programs for instruction and other applications as well as title creation and postproduction functions. Much basic program production, however, is still initiated on motion picture film.

THEORY OF MOTION PICTURE PHOTOGRAPHY

Motion picture theory is simple and clear-cut. Motion film is composed of a series of still pictures. When the still pictures are projected progressively and rapidly onto a screen, the eye perceives motion, hence they become a motion picture. This is termed *persistence of vision*. In a conventional sound motion picture, 24 individual pictures

are projected onto the screen in one second. The screen is blank (black) for 1/96th second. While picture number one is blanked out, the film moves to picture number two. The eye and mind retain the image of picture number one for that fraction of a second when the screen is black. If, for example, something in motion is being photographed, the difference between image one and image two is minute. The eye and memory record a smooth transition from picture one to picture two. This is perceived as motion.

Two Basic Components

1. A motion picture film has to be created, that is, produced. This can be relatively simple, or it can be highly involved and extend over a long period of time.
2. To be viewed, the finished motion picture is projected onto a screen.

Four Key Elements

- *Motion picture film,* the keystone of motion picture photography
- *Motion picture camera* that photographs a series of pictures in rapid sequence
- *Motion picture printer* in a laboratory to reproduce prints (positives from a negative or original positive).
- *Motion picture projector* that allows a viewer to see a series of still pictures as motion when projected in rapid sequence onto a screen.

ORIGIN OF MOTION PICTURE PHOTOGRAPHY

The concept of motion picture science is based on work by at least 16 men from 1824 up to 1889 (2). The earliest evidence (1824) is Peter Mark Roget's paper, *The Persistence of Vision with Regard to Moving Objects*, delivered at England's Royal Society. This stimulated experimentation in England by Sir John Herschel and Michael Faraday and on the continent by Dr. Simon Ritter von Stampfer in Austria and Dr. Joseph Antoine Plateau of Belgium's University of Ghent. In 1853, another Austrian, Baron Franz von Uchatius, projected images from a disc onto a screen using a magic lantern. In the United States, Philadelphia engineer, Coleman Sellers (1860), patented the *kinematoscope*. Using still pictures mounted on the blades of a paddle wheel, it produced a *zoetropic* (animation) effect. Two other men followed up on Baron Uchatius's invention, one in France and the other, Henry Heyl, in Philadelphia, who patented the *plasmatrope*. In 1872, the well-known Eadweard Muybridge experiment, using a rotating machine designed by Leland Stanford's engineer, John D. Isaacs (California), depicted the gait of a horse.

35 mm

In 1887, Thomas A. Edison (New Jersey) experimented with pictures using his phonograph principle. Edison's assistant, W.K.L. Dickson, actually did the experimentation and testing in developing the Kinetograph. He tried different sizes of film. It is believed that he cut an Eastman 70 mm still film in half, which he found easier to work with. In 1889, he finally settled on 35 mm film that had four perforations and a 4 : 3 aspect ratio. Essentially, both have remained the production and theater projection standards for more than a century, well into the twenty-first century (3,4).

Aspect Ratio

Aspect ratio is the relation of the width to height of the projected picture on the screen. Motion picture pioneers widely discussed and debated the subject in the early days. Paintings were carefully studied and analyzed. What would be most pleasing to a viewer (11)? The standard aspect ratio of 4 : 3 for regular movie theater showings has prevailed through the years. As the motion picture industry sought to increase viewing pleasure and excitement, various wide-screen applications were developed. The ratio most used is 1.85 : 1 in the United States, whereas 1.66 : 1 is more common in Europe.

Motion Picture Film Created

During 1887/1889, Edison went to George Eastman (Rochester, NY), who developed a 50-foot strip of film on *nitrocellulose* for Edison. Thus, motion picture film was conceived. Motion picture shows were made in Edison's famous *Black Maria* tar-paper studio to show on his *kinetoscope* peep show device. The kinetograph camera, weighing almost a ton, was used to shoot the films (3). Edison and Eastman also realized the need for a term to describe this new medium. After many exchanges of letters exploring Greek- and Latin-based terminology, they decided on *motion picture* as the best name for this new technology. George Eastman was fascinated with words; *motion picture* was both easy to understand and to say. In France, brothers Louis and Auguste Lumière designed the *cinematographe*. Their demonstration in Paris, March 22, 1895, was a huge success. Thomas Armat (Washington, DC) gets credit for the principle of the modern projector by the unveiling of his *vitascope* in September 1895. Others, including the Lumiere brothers, made additional contributions that led to modern 35 mm motion picture photography, as we know it today.

FILM

Photographic film used for motion pictures is normally a thin strip or ribbon. The exception is wide format films, 65 mm and 70 mm used for *wide-screen* shows. Professional motion picture people such as cameramen, film directors, and film editors dealing with film refer to motion picture film as *film stock*.

COMPOSITION OF FILM

Base

All film has a *base* that is a clear, transparent substance like clear plastic. The first films used for movies from 1890

up into the 1950s had a *nitrate film* base which is highly flammable.

Safety Acetate Film. In 1906, a slow burning, safety base film using *cellulose acetate* was developed, called *acetate film*. It is still in use, but it took almost 50 years before nitrate-based film was eliminated.

Polyester Film. More recently, a very tough base was developed from *polyethylene terephthalate* by E.I. duPont de Nemours & Co., trade named *Cronar*®. This base is used for print film and color duplicate negatives. It is not used as a camera negative because it is too tough to cut and splice and, in cold weather, becomes too stiff for camera use. Eastman Kodak uses the trade name, *Estar*™. The Fuji Photo Film name for its base is Polyester. It also has excellent chemical stability, which makes it ideal for long-term archival use.

Emulsion

In the manufacture of motion picture film, a light-sensitive compound called an *emulsion* is coated on the base. The emulsion (sometimes multilayered) is the key substance on which images are exposed.

TYPES OF FILMS

Black-and-White Film

Black-and-white film (B&W) comes in several types — negative, positive or print film, and reversal film. B&W intermediate film is used for duplicating negatives, for title preparation, for special effects, for sound recording, and for archival film elements.

Negative and Reversal

Negative Film. Used in a camera to photograph shots. That action is referred to as *shooting* a film. There are different kinds of B&W negative movie films. The most common is panchromatic, which is sensitive to all colors. Orthochromatic film is sensitive primarily to blue and green.

The image recorded on negative film is the reverse of the visual image seen by the eye. It takes a skilled person to read negative film properly. Someone unfamiliar with film is usually unable to identify the picture on a negative.

Reversal Film. Processes to a positive image after camera exposure.

Exposure Film Speed. A critical aspect of any film is its exposure speed. Film speed refers to its sensitivity to light. Slow speeds require stronger light such as sunlight or powerful indoor lights on a sound stage. The faster the film speed, the better that film can photograph in subdued light. Motion picture films use an *EI number* (exposure index) instead of ASA (American Standards Association) or ISO (International Standards Organization) used with still film. EI film speed ratings are half of ASA/ISO ratings. That makes EI 500T the same as ASA 1000.

Projection Film — Print (Positive)

Print Film. For projection, the image on negative or duplicate-negative film is printed on B&W positive stock that reverses the image from the negative. The resulting picture is identical to the image seen by the human eye, except without color.

Color Film

Color film is more complex than B&W. It comes as negative, reversal, intermediate, and print film. Color intermediate film is used for duplication and special effects. In shooting color film, color "temperature" is a critical factor. For example, outdoor color film is balanced for 6,500 K — the ideal temperature for color film. "Tungsten balanced" film is designed for studio lamps at 3,200 K.

Negative, Intermediate, and Reversal

Color negative film has three color emulsions — yellow, magenta and cyan — and two filter layers, three protective layers, the support base, and a rem-jet backing layer that is removed in processing (5).

Color intermediate film is essentially a negative film. It is used in two stages, first to make a color master positive and then a color duplicate negative. A fine grain copy of the original negative is used for quantity printing of the *release prints*. The release prints go to theaters or are sold to schools and other customers.

Color reversal film is exposed in the camera and then developed to a positive image that is used in a projector or telecine. A common example is Kodachrome™ for amateurs. Kodak Ektachrome 16 mm film was developed

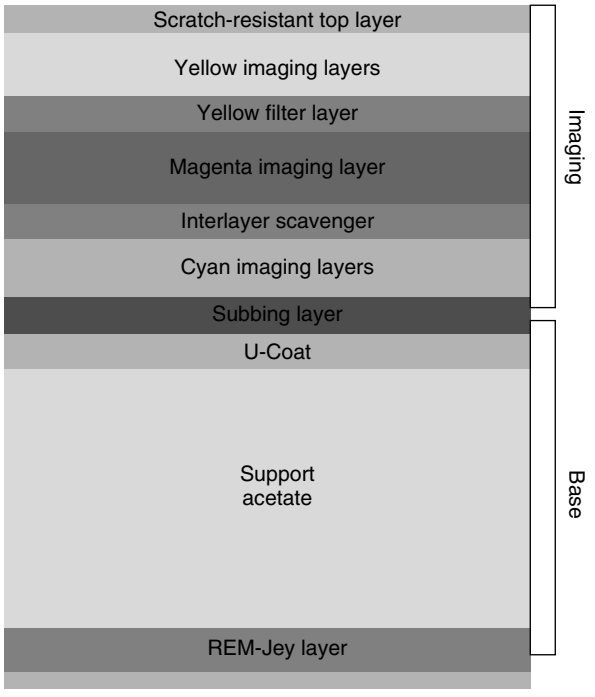


Figure 1. Cross section of Kodak Vision™ color negative film (courtesy Eastman Kodak Co.) (5).

for news gathering before videotape was introduced. It is widely used for automobile crash testing.

Projection Film — Print (Positive)

Color print film has three color emulsions — magenta, cyan, and yellow — two interlayer scavengers, a scratch-resistant top layer, an antihalation layer, a subbing layer, and, below the support base, a U-coat, a process-surviving antistatic layer, and at the bottom, a protective scratch-resistant layer and lubricant. The scavenger layer absorbs or inhibits the migration of unwanted chemicals.

Because of the many layers and their complexity, high-speed color film is a more difficult product to manufacture. It took years to bring color film speed up to that of black and white.

Color Film Evolution

The first successful color motion picture film came from Technicolor, a major Hollywood motion picture laboratory, when it introduced a two-color film for theaters in 1922. In 1932, Technicolor introduced a three-color, 35 mm film system called *imbibition* or *IB* for short. For the next 20 years, Technicolor's IB was the principal color film for theatrical motion picture productions. It was expensive to shoot because it required a large camera that literally exposed three separate B&W films simultaneously through red, green, and blue filters.

Kodachrome color film, introduced in 1935 by Eastman Kodak Co., was the first practical 16 mm color motion picture film. It is a color reversal film also available as 35 mm still picture film. In 1936, Kodak developed it for the 8 mm motion picture market. Although initially intended for the amateur, Kodachrome film found a market in the industrial nontheatrical film field and eventually in education and other markets.

Subsequently, other color reversal motion picture films were introduced. Agfa (Germany; after World War II it was Orno in East Germany) and Gevaert (Belgium) were widely used. Another was Anscochrome by General Aniline and Film. Finally, in 1958, a 16 mm film designed for duplication was developed by Kodak. Ektachrome Commercial replaced Kodachrome and similar films not designed for duplicating. Fuji (Japan) introduced a color reversal film in the 1970s.

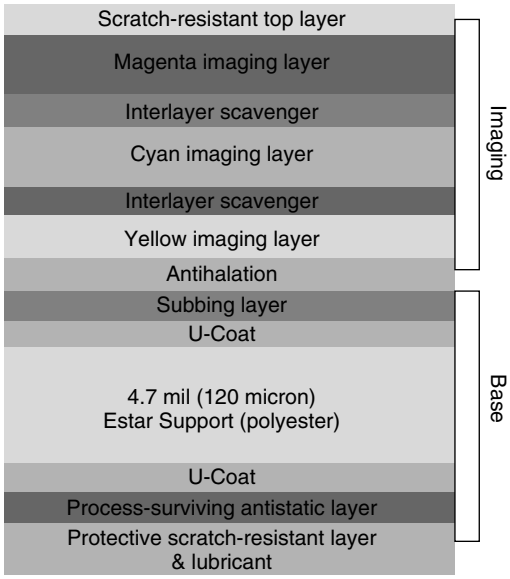


Figure 2. Cross section of Kodak Vision™ color print film (courtesy Eastman Kodak Co.) (5).

In 1950, Eastman Kodak introduced a negative/positive 35 mm color film system for the theatrical motion picture world. In the early 1970s, it was introduced in 16 mm, intended for nontheatrical markets. Eastman Color became the world standard. Fuji Photo Film of Japan brought out a film compatible with Eastman Color. Now, Agfa sells only color positive (print) and some recording films. General Aniline and Film no longer sells motion picture film.

Currently, the most popular 35 mm camera color negative film is 500T (tungsten). Fuji finds that its 250T film is its next most popular. Eastman Kodak reports that the 800T is the next most popular, followed by 320T. Today, very fast films whose speeds are 800T can record good pictures in candlelight. New emulsion technology by Eastman Kodak resulted in its patented T-grain in 1989, which makes the silver halide crystal a more efficient gatherer of light. Likewise, it is better as a storage medium.

16 mm and Safety Film

In 1912, Pathé Frères of France introduced 28 mm film that had a nitrate base, even though acetate safety base had been available for 6 years. Alexander F. Victor, a Swede who had moved to the United States, urged manufacturers to put this 28 mm film on the new acetate base instead of the volatile nitrate base. This would let motion pictures be used safely in schools, churches, homes, and other places outside of theaters. In April 1918, the Society of Motion Picture Engineers (SMPE) adopted 28 mm film on a safety base for portable projectors (6).

Some industry people continued to push for 17.5 mm film for portable projectors. George Eastman argued that unscrupulous manufacturers might split 35 mm nitrate-based film which could pose a severe danger from fire. In 1923, Eastman Kodak introduced 16 mm film on acetate base plus a camera and projector designed for home use. Victor quickly designed a 16 mm projector for the nontheatrical education and industrial markets. The 28 mm film never went any further.

PHYSICAL CHARACTERISTICS

Film Width

From the very beginning, film width has been measured in millimeters instead of inches for more precise measurement.

- *35 mm film* — Theatrical industry film standard for more than 100 years. This is for use in both cameras and projectors.
- *65 mm* — Used solely for photographing a motion picture for *wide screen*.

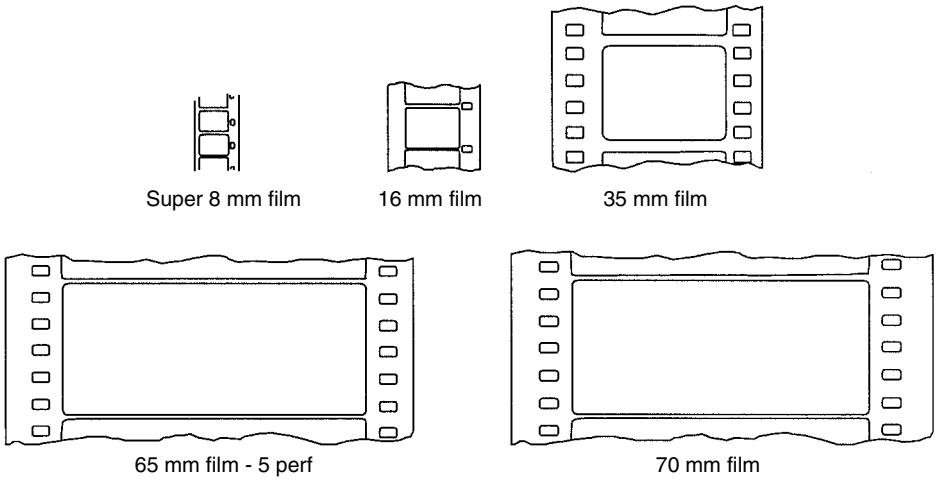


Figure 3. Five basic film sizes (width) (courtesy Fuji Photo Film).

70 mm — The resulting projection film, printed from 65 mm camera film, is 70 mm wide to allow for four to six sound tracks.

16 mm — For many years, 16 mm film was highly popular, second in annual volume to 35 mm film. Introduced in 1923, it quickly became the standard format for educational subjects and, subsequently, for industrial films and documentaries. In addition, by the early 1930s, an amateur market had developed for 16 mm film.

8 mm — In an effort to reduce the cost of filming for amateurs, several narrower widths were introduced. One, 8 mm, became the amateur standard worldwide. Another size, from France, was 9.5 mm. The French format never gained worldwide popularity, but it continues to be used.

Three New Formats

As a general rule for choosing a film format, the larger the original negative, the better the picture. The larger film format reduces projected image grain and increases sharpness. It also improves picture steadiness because less magnification is needed. This recognition led to newer formats.

Super 8. In 1965, 8 mm film was changed, reducing perforation size and thus enlarging the picture area and providing room for a future sound track. Then, in 1973, sound super 8, that had a magnetic stripe for the sound track, became available.

Super 16. 16 mm film was the next to have an enlarged picture area, using more of the width for a sound track.

Super 35. Eventually, a larger picture area of 35 mm was designed. The full width of the film inside of the perforations is now the picture frame, using space originally reserved for the sound track.

Length

Motion picture film used in cinematography is a continuous length on a spool or core. For camera use, the film normally has a specific length such as 100, 400, 1,000, or 2,000 feet, depending on the camera magazine capacity. For projection, the film can be any length. Typical lengths are 400, 800, 1,000 feet, and longer. In trade terminology, 1,000 feet of 35 mm and 400 feet of 16 mm film are referred to as a *reel*. Laboratories are supplied with color positive film in rolls up to 6,000 feet long.

Perforations

Motion picture film contains small rectangular or square holes along both or one edge called *perforations*. The perforations are used to move the film through the camera, printer, or projector. Most cameras use a pull-down claw that works in concert with the perforations to move the film. In the projector, a small roller that has sprockets engages the film perforations to move the film. These functions are described in the sections on cameras, printers, and projectors. Perforations, generally called *perfs*, vary in size and shape. They are measured by width and height in hundredths of an inch (or millimeters)

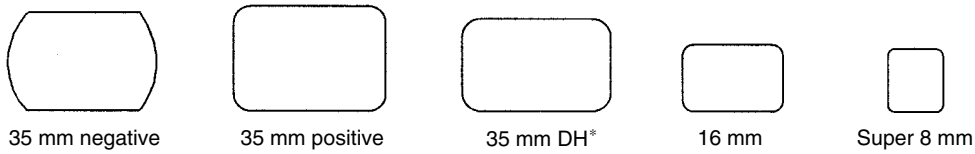


Figure 4. Enlargement of five perforation types. *35 mm DH (Dubray and Howell) perf is used for some intermediate film (courtesy: Fuji Photo Film).

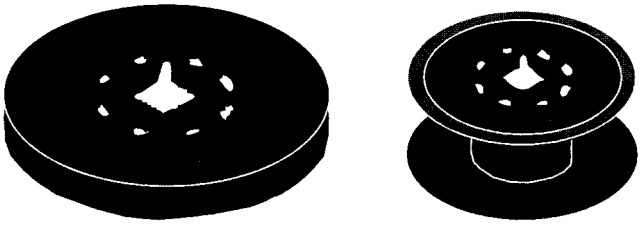


Figure 5. Film spools (7) (courtesy Eastman Kodak Co.).

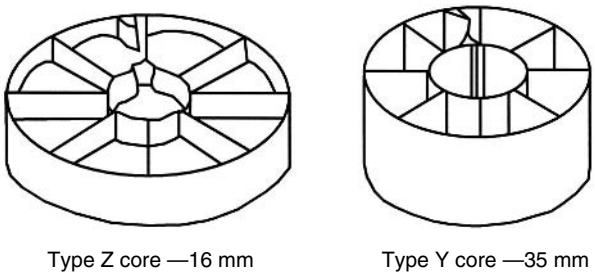


Figure 6. Film cores (7) (courtesy Eastman Kodak Co.).

designed precisely by international standards to ensure worldwide compatibility. Another critical measurement is *pitch*, the distance from the center of one perf to the center of the next perf. In reality, the measurement is made from the bottom edge of one perf to the bottom edge of the next perf.

> In the early days of 35 mm motion pictures, film perforations were round. Because these perforations were more subject to wear, the shape was changed to what is now known as the Bell & Howell (BH) or 'negative' perforation*. . . . The high shrinkage of older films on nitrate base made the negative perf a problem on projection films because of the excessive wear and noise during projection as the sprocket teeth ticked the hold-back side of the perf as they left the sprocket. The sharp corners also were weak points and projection life of the film was shortened. To compensate for this, a new perforation was designed with increased height and rounded corners to provide added strength. This perf, commonly known as the KS or "positive" perf, has since become the world standard for 35 mm projection prints (7).

To improve the printing of color print film, another perforation was designed by Dubray and Howell (DH perf) that had the same height as the negative perf to maintain exact registration but rounded corners for longer printing life. Because shrinkage of acetate film is greatly reduced, the DH perf is used for some *intermediate films* (7). A fourth perforation was designed for CinemaScope, an early wide-screen system. Square and smaller than the KS perf, it provided space on 70 mm film for as many as four magnetic sound stripes for surround sound. It is also called "Foxhole Perf," because 20th Century Fox invented the cinemascope format. Perforations for 16 mm and super 8 mm film have not changed over the years.

CAMERA FILM

Most shorter lengths of film, 50 feet of 8 mm, 100 feet of 16 mm, and 100 feet of 35 mm, are available on a spool. Longer lengths of 16 mm and 35 mm are wound on a core. Film on a spool is loaded directly into a camera in subdued light. Film on a core must be loaded in a darkroom into a lighttight metal container called a *magazine*. The magazine, similar to a layer cake tin, is mounted on the camera. Usually several magazines are loaded at one time, so that additional ones are available during shooting.

* Used for all camera films and most intermediate films.

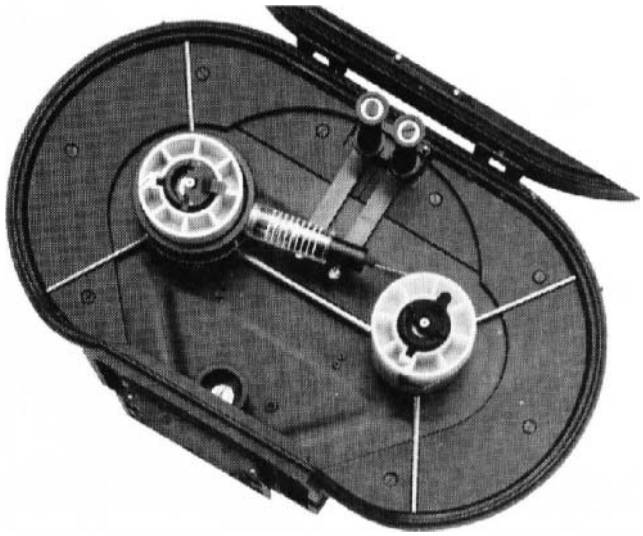


Figure 7. 16 mm film magazine (8) (courtesy Arriflex Camera Co.).

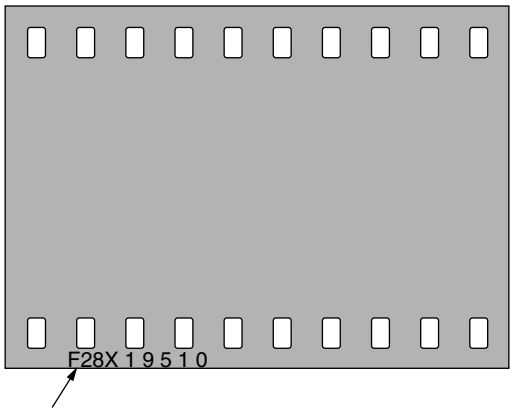


Figure 8. Traditional edge numbering (7).

Edge Numbering. All negative film has a series of sequential numbers printed along one edge of the film, a number every foot. This allows a film editor to identify specific scenes easily. Since the advent of video editing and the use of the computer and bar codes, in addition to edge numbers, film manufacturers preexpose numbered bar codes along the film edge that can be read electronically. Referred to as *edge codes*, they show the

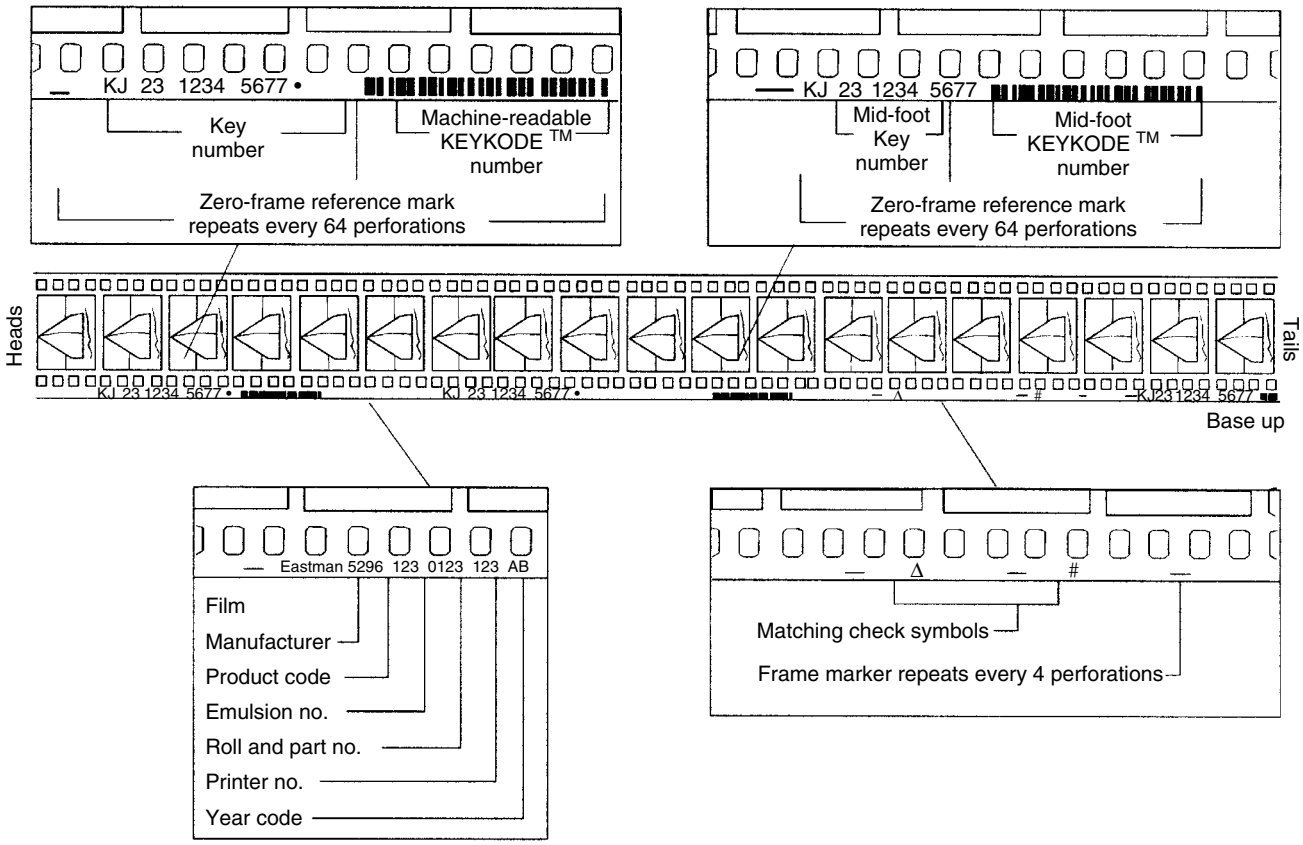


Figure 9. Eastman 35 mm edgeprint format with KEYKODE^{TM} numbers (17).

manufacturer's name, product code, emulsion number, roll and part number, printer number for positive film, and year code.

MOTION PICTURE CAMERA

A motion picture camera is a lighttight precision instrument. All professional cameras, with a few exceptions, are operated by an electric motor. The power can be supplied by a battery in the field or ac power in the studio. Early motion picture cameras were hand cranked. Subsequently, cranks were replaced by spring-driven motors. Today, most amateur cameras are spring-driven.

Pull-Down Claw

In the camera and in an optical printer, a *pull-down claw* moves the film through the film gate. Registration pins are retracted as the film moves, then protrude when the film pauses, so that the film is exactly aligned with the aperture opening.

Shutter

The camera has a disc, called a *shutter*, mounted perpendicularly to the axis of the lens. The shutter has an open portion of the disc which allows the image of the subject being photographed to pass from the lens through an aperture to the film. The shutter opening can be varied

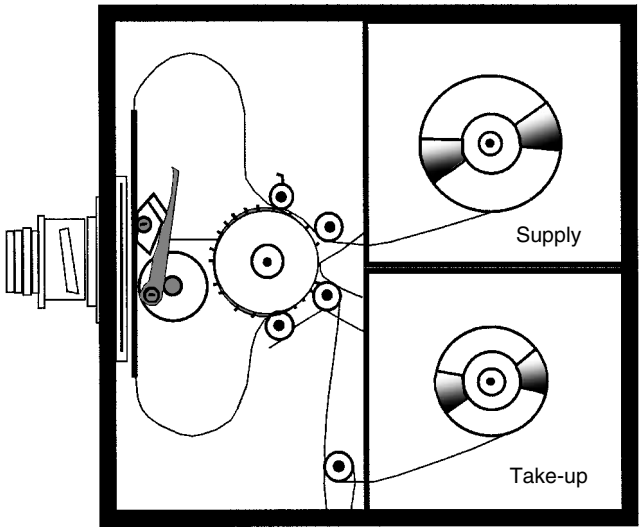


Figure 10. Simple schematic of a movie camera (1).

from a slit to an opening as wide as half the shutter. This is known as a *variable shutter*. The shutter moves at whatever speed the camera is set for, based on the number of revolutions per second. The pull-down claw is synchronized precisely with the moving shutter, so that the shutter opening is in direct line with the lens axis just as the film frame is momentarily still and can be exposed from the image coming through the lens.

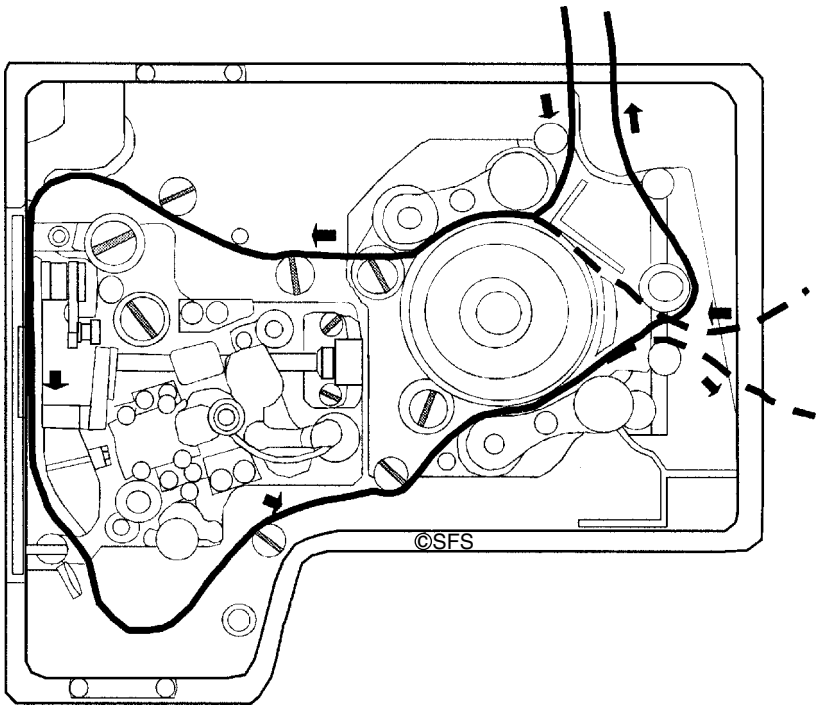


Figure 11. Schematic of Panavision camera. Solid line (film) goes up to regular magazine Dashed line (film) goes back to handheld camera magazine (courtesy Panavision).

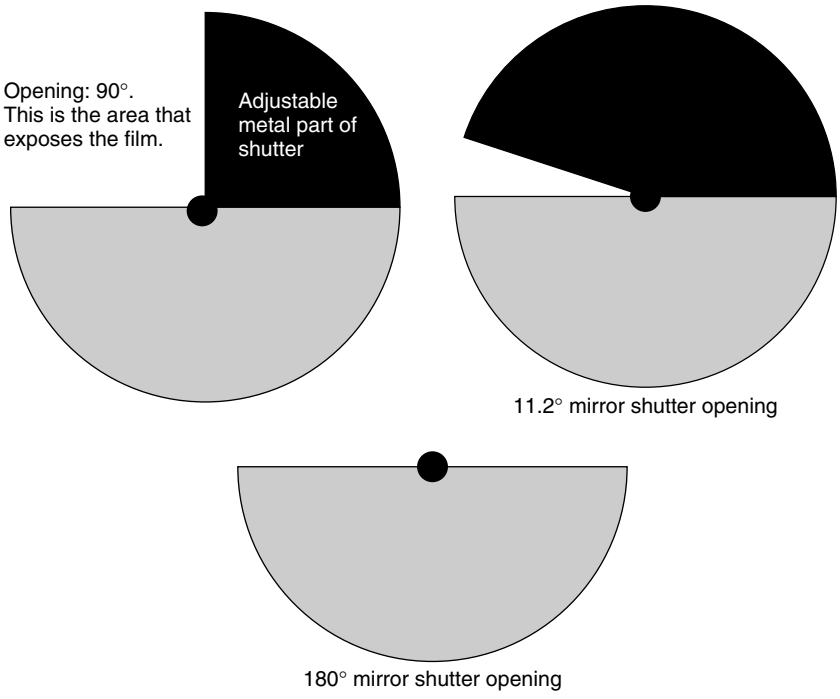


Figure 12. Variable Arri camera shutter (8) (courtesy Arriflex Camera Co.).

Frame and Film Gate with Aperture

As the film passes through the film gate and stops momentarily in front of the aperture and lens, a small portion of the film is exposed. The resulting picture is called a *frame*. The size of the aperture opening determines the size of the image. Aperture aspect ratio is covered in the section on Projection. Using 35 mm film as an example, the pull-down claw pulls an unexposed segment of film 22.5 mm in a fraction of a second to a position perfectly aligned with a lens and the aperture opening. The frame is exposed. Then, the pull-down claw pulls the exposed film (usually) down, and the next unexposed frame comes into alignment.

Camera Speed

When that action is repeated 24 times in one second, that is termed *sound speed*. For slow motion, the camera runs at a faster rate, such as 64 frames per second or much

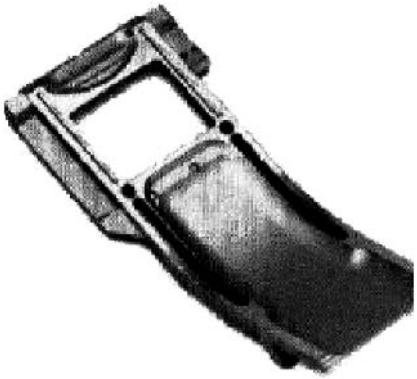


Figure 13. Film gate and aperture (courtesy Arriflex Camera Co.).

faster. High speed for extreme slow motion is a highly technical and specialized subject. For silent speed, such as in amateur cameras, the camera usually runs at 16 frames per second.

Lens

The *lens* is another essential element of a camera. It is an optical device manufactured to extremely close tolerances. Each professional lens is tested on a *lens-testing bench* to ensure that it conforms precisely to the specification. High quality lenses are expensive. A camera has a standard lens that produces a screen picture that has a 4×3 ratio (actually, 1.371 : 1). The focal length varies depending on the size film that the camera is designed to use. Another type of lens is a *zoom lens*. Its focal length can be changed from wide angle to standard to telephoto while the camera is filming.

Viewfinder

Motion picture cameras have two kinds of viewfinders. Some are built in and bring the image to the viewfinder through the lens; it is known as a *reflex finder*. Others are attached to the right side of the camera body and must be adjusted for parallax to avoid positional errors in framing, especially for close-up shots.

VIDEO

Many cameras have integrated video cameras attached, allowing a shot to be recorded on film and videotape simultaneously. This provides immediate evaluation of a shot before the film is processed.

Significant Cameras

Aaton — Made in France (1971). Noted for in-camera time code. Super16/16 mm and 35 mm models used for documentaries, TV, and independent feature production.

Arri — Made in Germany by the Arriflex Corporation, it is widely used in Hollywood as well as among contract producers who make TV commercials and industrial films. Available in 35 mm, 65 mm, and 16 mm (8).

Cine Special — Eastman Kodak's professional U.S.-made 16 mm camera can be said to have launched the nontheatrical industrial film business.

Mitchell — America's Mitchell camera was the workhorse of the Hollywood film industry for more than 50 years until the last quarter of the twentieth century. It was primarily a 35 mm camera, but there was also a 16 mm model.



Figure 14. Aaton XTR prod Super 16/16 mm camera (courtesy Aaton).



Figure 15. Arri camera available in 35 mm and 16 mm models (courtesy Arriflex Camera Co.).

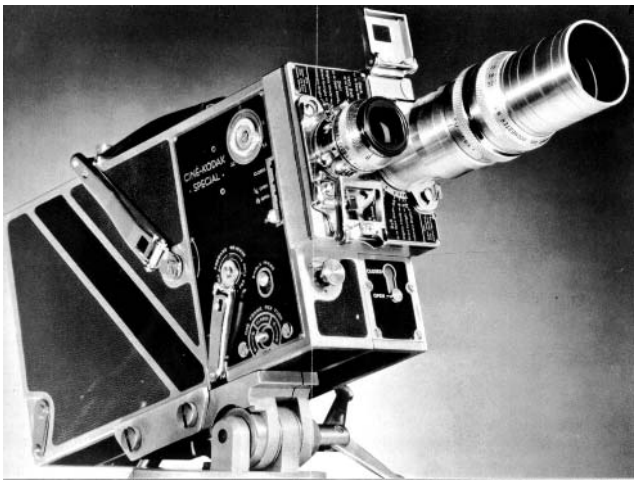


Figure 16. Kodak Cine Special (10) (courtesy Alan Kattelle, Home Movies.).



Figure 17. Mitchell camera (courtesy Grant Loucks, Alan Gordon Enterprises, Inc.).

Moviecam — The newest (1970s) professional camera from Austria is lightweight and well designed for all kinds of production work.

Panavision — Available in 16 mm, 35 mm and 65 mm. All cameras are owned by Panavision and are rented to producers worldwide. Panavision has been a major U.S. camera supplier for feature, television, and TV commercial production since 1972.

PRO8 — A professional U.S. camera from Super8Sound. The camera is adapted from a French Beaulieu. Super 8 film dates from 1965 (9).

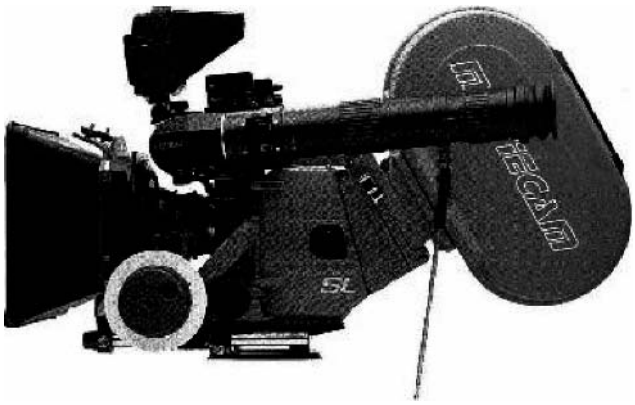


Figure 18. Moviecam (courtesy Moviecam F. G. Bauer GesmbH.).



Figure 19. Panavision Panaflex camera in 35, 65 and 16 mm models (courtesy Panavision).



Figure 20. Super Pro8 camera (courtesy Super8Sound).

There are other makes of 35 mm and 16 mm cameras. Wide-screen 65 mm cameras are discussed in the section on wide screens.

Accessories

Tripod — One of the most important accessories for the motion picture camera is a tripod. All heavy cameras need a tripod of some kind. The best tripods are those that have fluid heads. A good fluid head allows panning or tilting the camera so smoothly that the eventual viewer of the film will not be aware that the camera was moving.

Dolly — A special *truck* built to hold a large studio camera and camera operator. The *dolly* usually runs on tracks, so that its movement is absolutely smooth. The camera can move toward or away from action, as well as parallel to a scene.

Handheld — Motion pictures shot without a tripod convey a feeling of casualness and movement. There is a noticeable difference between filming crudely done without a tripod versus intentional handheld cinematography. News footage is often shot by using a shoulder pod.

Steadicam™ — A unique camera stabilizing system for handheld camera operation. Worn as a harness, the Steadicam allows the camera to glide smoothly in any direction while the camera operator moves around.

Video Assist — A new feature has been added at the shooting stage. A video camera is used with the film camera and photographs the same scene. This allows the director to look at a scene right after it has been shot and decide on the spot to accept or reshoot it. Videotaping along with filming does not replace the need to look at the *dailies* the first thing the next morning. See Dailies, at end of Production.

SOUND DEVELOPMENT

In 1923, Lee DeForest showed short synchronized sound films using his *phonofilm* (11). Warner Bros. of Hollywood introduced sound in its 1926 film *Don Juan*, using a musical score on 18-inch discs. The following year Warner again made history using its "Vitaphone" for *The Jazz Singer* which used spoken dialogue, the first *talking picture* (2).

Optical Sound

Within a few years, sound on 35 mm film had been perfected. The sound track was located between the perforations and the picture frames. The sound track is known as optical sound. Two different systems were introduced. One was *variable density*, called *push pull*, which used minute horizontal lines similar to today's bar code lines. The other, *variable area*, has the appearance of a jagged graph line running along the film track area. Today, all analog sound is variable area. Sound on 16 mm film came in 1930 when RCA introduced a 16 mm sound projector. Because 16 mm film is narrower than 35 mm, perforations on one side were eliminated to provide room for the sound track. Eight mm film did not get sound until 1973 when *Super 8 sound* was introduced. It likewise had perforations only on one side.

Digital Sound

In recent years, recorded sound has undergone a dramatic change as digital technology is used. Theater digital sound is available in three systems:

Dolby sound
Sony SDDS sound
DTS (digital theater system)

WIDE SCREEN

There are basically three kinds of wide-screen systems. The most common uses 35 mm film where the picture is masked at top and bottom to give a 1.85 : 1 aspect ratio wide-screen appearance. Theater screens for this system tend to be small, 23 to 47 feet wide. This wide-screen system in the industry is termed *flat*, and, at the beginning of the twenty-first century, is really not considered "wide-screen."

Another system uses 35 mm film which uses an anamorphic lens during filming, to squeeze the image, so that the images on the negative are extremely tall and slim. In projection, a 2× anamorphic lens spreads the image to produce a wide-screen picture whose aspect ratio is 2.39 : 1. This wide-screen system, originated by Cinemascope, is called *scope*.

The third wide-screen system is shot on 65 mm film and is projected using 70 mm film; both have five-perf pull-down. The aspect ratio is 2.21 : 1, and the screen can vary from 30 feet wide up to 60 feet or more. Because of the expense, this system is mainly used for spectacular films such as 3-D.

Though wide-screen cinema exploded in popularity in the last half of the twentieth century, it actually dates back to 1896–1910. In 1900 at the Paris Exposition, Raoul Grimoin-Sanson used his *Cineorama* to show a film on a 360° screen 300 × 30 ft. It involved ten 70 mm projectors to fill the giant screen (11).

The use of the wide screen in movie theaters has grown in popularity, especially since television has become a major entertainment venue. Modern use of the wide screen began with *Cinerama* in 1952.

In the following year, Cinemascope came out, followed by VistaVision, Todd-AO, and Technirama, all in the 1950s. In 1970, three more wide-screen systems were launched — Ultra Panavision, IMAX™, and Omnimax™. Three other wide-screen systems have been introduced since — iWERKS, Show Scan, and MegaSystems, using 8-, 10- and 15-perf pull-down systems.

IMAX has become one of the most widely used wide-screen systems worldwide. It is unique because the film is photographed and also projected on a horizontal plane. The camera is IMAX's own 65 mm design. They also have their own 70 mm projector. Because of the horizontal format, the frame is larger than that of any other system.

ShowScan is another unique technology. The film runs at 60 frames per second. Each film is short, 4 to 5 minutes. The seats move during screening so that the show is more like an amusement park experience.

PRODUCING A MOTION PICTURE

Making a professional motion picture can be relatively simple or very complex. The former can be done by a single person, although a crew of at least three or four is the more common minimum. In contrast, the production of a feature film involves hundreds of highly skilled individuals. The procedure essentially is the same for any film production.

Preproduction

Creating the story is obviously the starting point for any motion picture, be it a documentary, an educational subject or an entertainment feature film, television drama,

Table 1. Wide Screen Technologies[a]

Film	Year	Gauge	Cam or Proj	Aperture (Inches)	Perfs/ Frame	Aspect Ratio	Sound Format	Screen
Cinerama	1952	Triple 35 mm	c	0.996 × 1.116	6	2.65 : 1		
		Triple 35 mm	p	0.985 × 1.088	6	2.65 : 1	seven-track mag	Curved
Cinemascope	1953	35 mm	c	0.868 × 735	4	2.35 : 1[c]		
		35 mm	p	0.839 × 0.700	4	2.35 : 1[c]	four-track stereo mag	2× anamorphic squeeze
VistaVision	1954	35 mm	c	1.495 × 0.991	8	1.5 : 1		
		35 mm	p	0.825 × 0.446	4	1.66 : 1.85	Optical monophonic	Proj spherical
		35 mm	p	1.418 × 0.7228	8	1.85 : 2.1	Optical monophonic	Spherical
Todd-AO	1955	65 mm	c	2.072 × 0.905	5	2.21 : 1		
		70 mm	p	1.912 × 0.870	5	2.21 : 1	Mag six-track stereo	Spherical
		35 mm	p	0.839 × 0.700	4	2.35 : 1	Mag optical	2× anamorphic squeeze
Technirama	1957	35 mm	c	1.496 × 0.992	8	2.25 : 1		
		35 mm	p	0.839 × 0.700	4	2.35 : 1[c]	Mag optical four-track	2× anamorphic squeeze
		70 mm	p	1.912 × 0.870	5	2.21 : 1	Mag 6-track	
Ultra Panavision	1970	65 mm	c	2.072 × 0.906	5	2.76 : 1		
		70 mm	p	1.912 × 0.870	5	2.76 : 1	Mag six-track stereo	2× anamorphic squeeze
		35 mm	p	0.839 × 0.700	4	2.35 : 1[c]	Mag optical four-track	2× anamorphic squeeze
IMAX	1970	65 mm[e]	c	2.032 × 2.772	15	1.364 : 1		
		70 mm[e]	p	1.91 × 2.74	15	1.435 : 1	Digital six-channel	Curved high gain of 2
IWERKS	1989	65 mm[e]	c	2.032 × 2.772	15	1.364 : 1		
		35 mm	p[d]	0.920 × 0.690	4	1.33 : 1	DTS[f]	
		70 mm	p[d]	1.912 × 0.868	5	2.20 : 1	DTS[f]	
		70 mm	P[d]	1.913 × 1.417	8	1.35 : 1	DTS[f]	
		70 mm	p[d]	1.913 × 1.838	10	1.64 : 1	DTS[f]	
		35 mm	p[d]	1.485 × 0.991	8	1.498 : 1	DTS[f]	
ShowScan	1993	65 mm	c		5			Curved
		70 mm	p		5	2.2 : 1	DTS 5.1	60 f/s
MegaSystems	1998	65 & 35 mm	c	Varies	Varies	1.35 : 1		
		70 mm	p	1.417 × 1.413	8	1.35 : 1	Digital	Dual Geneva intermittent

[a]Ref. 12.
[b]C = camera aperture; P = projector aperture
[c]With squeeze.
[d]Linear loop projector.
[e]Horizontal.
[f]DTS = Digital theater sound
Note: None of the magnetic-strip sound formats is still produced except for archival purposes.

sitcom, or TV commercial. The script is normally the next step. Probably, 99% of all productions begin with a carefully written script. For the majority of motion pictures, locations must be determined and sets designed and constructed in the planning stage. Another major preproduction function is casting, choosing the actors or on-screen personalities if it is to be a documentary.

Production

There are a number of elementary principles in motion picture production. They are, however, not hard and fast. For example, a director may choose to violate one to achieve a certain effect or mood.

Pictorial Continuity (13)

The director, director of photography, and film editor are guided by a number of artistic principles developed over the years. Following are a few essentials presented in headline form.

Sequence — Long shot, medium shot, close-up. Extreme long shot. Extreme close-up.



Figure 21. IMAX 3-D camera. The film path of the 65 mm 3-D camera goes through a staircase where the film is twisted, so that the magazines can be mounted in their vertical configuration. The image from the lenses is actually folded by mirrors before it reaches the film; one eye folds up, and the other folds down. This allows a natural lens interocular for proper 3-D shooting (courtesy IMAX Corp.).

Panning — Planned panning. Panning ahead. Caution against unnecessary camera movement.

Reestablishing shot — Pulling back. Panning. Reverse angle.

Travel shot — Truck. Moving camera on a dolly.

Directional continuity — Screen direction of the action. Consistency to avoid distraction.

Overlap and matching action — Controlled and uncontrolled action. Avoid jumpy action.

Cut-ins, Cut-aways — Head-ons. Tail-aways.

Angles — Vary the angles.

Buildup — Good use of continuity principles. Time and space. Build suspense.

General Rule — When shooting a new scene, change the size of the image, change the angle, or both.

Sound Recording

Audio for a sound motion picture is usually recorded by using sound equipment that is not an integral part of the camera.

Double system sound. For Hollywood feature film production and practically all other professional production, the sound recording equipment is separate from the camera. This is known as *double-system sound*. Skilled engineers and highly trained audio technicians work in cooperation with the camera operator, also known as the cameraman.

Sync sound. Most feature films are shot with synchronous sound recording, called sync sound. All dialog, for example, is both photographed and sound recorded at the same time, using a camera and sound recorder running at *exactly* the same speed using crystal-controlled motors.

Single-system sound. A major exception was the super 8 mm sound camera. That type of camera recorded the sound on a magnetic stripe placed on 8 mm negative or reversal film opposite the side that is perforated. The sound was recorded in the camera as the film was shot. Today, very little is recorded single system.

Voice Over. The commentary for a film is done by a narrator and other voices speaking off camera in conjunction with the scenes pictured. The recording is normally done after the film work print has been edited into final form.

Sound for much news coverage is recorded in the camera, usually shot on videotape. Primarily amateur filmmakers using a video camera or people doing low-budget film production might shoot videotape that records the sound in the camera.

Digital Technology. Audio technology has made great strides in recent years. Digital audio, for one, has transformed sound quality. Digital is a technology whereby the audio signal is broken down into binary bits that represent a mathematical model of the original signal. This system is superior to and is superseding the *analog* electrical signal that is a continuous variable.

Digital sound has two principles.

1. Digital bits are recorded on the film (Dolby Digital and Sony SDDS).
2. Time code is recorded on the film that controls the playback from a separate CD-ROM player (DTS).

Lighting (14)

If any element of production is underrated, lighting (and shadows) is, but it is important to the success of cinematography. A *mood* can be created. Proper use of *backlighting* and a *key light* can emphasize something in a scene. It can help to *create depth*. *Shadows* can dramatize a scene. The *wrong shadows* on the face of an actor can be devastating. Lighting is a fine art. The lighting director has many *tools* today to create the proper lighting for a scene. This applies both indoors and outdoors.

Dailies

Whenever possible, all of the negative film shot during the day is processed each night. A *one light* work print (normally in color) is made. Color corrections are not made at this stage. *See Printing*. These are called *dailies*. In the morning, the director and cameraman (director of photography) look at the dailies to decide if they got what is needed. This includes the acting as well as proper camera angles and other technical aspects.

Video dailies are common. After the film negative is processed, it is put on a telecine machine that scans the

image and records it on videotape. It can be NTSC, PAL (standard definition), or high definition.

Post Production

When all *production* (shooting) is finished, the film goes into postproduction. Film is processed (developed) by a motion picture film laboratory.

Laboratory Services. *For purposes of this article on Motion Picture Photography, cinematography laboratory services and post-production functions are greatly simplified.* A motion picture laboratory provides many services. Not every lab offers the same services. Some smaller labs specialize in production support. Others do volume printing of release prints such as theatrical feature productions, television programs, and TV commercials. Other labs may cater to the industrial market and to contract producers who often work in 16 mm and end up making videotapes for distribution and video projection for use on television. For some of these labs, videotape work has become their major business, although it originated from the basic film processing service.

Processing. A primary service of most laboratories is film processing. Processing is the function of developing exposed film through a series of chemical baths and washes. It begins with the negative film. This same chemical process is also used to develop intermediate film. Print film uses a different process. Certain films require special handling and may use processing chemicals different from those normally used. The final stage is drying the film, after which it is wound onto a core or a reel.

Printing. An important service of most laboratories is *printing*. Printing is making a copy or print from a negative or, for quantity release printing, from a duplicate internegative (to protect the very valuable original negative). The *printer* is a sophisticated piece of equipment. The negative film is placed in a printer, which beams light through the aperture and projects the negative onto unexposed print film.

Printing Lights. In the early years, printers literally had 21 lamps. Depending on the density of the negative, a technician preparing a film for printing judged the film to decide how many lights would be needed for a given scene. The settings were punched into a tape, hence the term *tape setting*. Today printers use a single light which is split into three components (red, green, and blue) using filters. Each component's intensity can be varied. The term *printer light* has remained a common lab term in print exposure. In speaking about the density of a negative, they will refer to a certain scene as requiring so many lights or *trim adjustment*. Each change in setting is 0.025 log exposure.

Film Editing

The major operation in the postproduction phase is film editing, although the actual technique has changed greatly in recent years. For nearly a century, editing both picture and sound used a procedure barely changed over the years.

Moviola. For more than a half century, the *Moviola*, a unique machine designed as an editing tool, was the backbone of the editing process. Using a small rear projection screen and the capacity to play back several sound tracks, an editor could review portions of a film work print, as well as synchronize and interlock picture and sound. Eventually the Moviola was displaced by *flat bed editing tables* which are still used in editing today. Steenbeck and Kem are two popular makes. The flat bed has all of the features of the old Moviola and the advantage of holding several reels of film, such as A and B rolls, plus splicer.

Fade. Done either in the camera or in the laboratory. A *fade-in* is a scene that begins when the screen is totally black and gradually fades to full exposure. A *fade-out* is the opposite, going from full exposure to black. The *fade* is usually used to begin or end a sequence of scenes.

A and B Rolls. Original film, the negative or reversal film, is divided into two rolls by a break wherever an optical effect, such as a dissolve, is to be made. Using A and B rolls, titles can be printed over backgrounds. Using a 16 mm color negative, splices can be hidden. The A and B rolls are both put on the printer.

Dissolve. As one scene fades out, the next scene fades in over the fade-out. The *dissolve* is usually made by laboratory printing using A and B rolls. It can be made, however, in the camera by doing a fade-out and then winding the film back and finally a fade-in.

Special Effects. Scenes in a film that are not standard camera shots, used to achieve something unusual, are broadly termed *special effects*. They may be done in the laboratory or shot by a company or people specially skilled in this technical art. Optical techniques are still used, but many visual effects are done digitally after scanning the film at a very high resolution — 1980 pixels per line or higher.

Feature Film Editing

What follows is a simplified description of the traditional film editing operation. For nearly a century, all editing literally consisted of cutting the work print into pieces. Scenes would be regrouped to create the desired sequence. Each piece of film is then trimmed to the desired length to make the action flow smoothly. The individual pieces are cemented together (with a special glue called *film cement*) putting all of the pieces back into a single linear strip of film. Once the editing of the work print is completed and approved by the director and/or others, the negative is cut or assembled to match the approved work print. The camera film is timed using a color analyzer. At this point, depending on whether the original is 65 mm, 35 mm, or 16 mm, several options are usually available. They depend on the capability of the laboratory as well as cost factors. For purposes of this subject, let us assume that 35 mm color prints will eventually be made from a 35 mm negative. Once the negative is cut to match the work print, a *color master positive* (interpositive) is made. Next, using

optical printing, a *color duplicate negative* (also called an internegative) is made from which the final color release prints are made.

Editing On Videotape

Many feature films are edited on videotape. Once the edited videotape is completed, one of two steps is taken. For a regular feature film, the camera negative is matched to the edited videotape. That results in one of the procedures shown in Fig. 22. In small budget pictures destined to be released in videotape, the dupes are made directly from the edited videotape when the very highest quality is not required. Film shot in 65 mm or in 16 mm film goes through essentially the same process.

Nonlinear Editing

More recently, the editing process involves digital technology —the most recent state of the art. Most film is now edited in a computer using a *nonlinear editing system* developed by Avid, Media 100, and Lightworks. The original negative "image" is digitized and transferred into a computer in *computer bins*. Here, the editor is not confined to editing in typical linear fashion as outlined before. Instead, scenes can be called up and shifted around randomly, as the film is edited. Once the whole film, with sound, is in final form in the computer, an "edit decision list" (EDL) is produced from which a negative cut list is derived using the edge code. The film negative is matched, and the process follows the same procedure outlined for making film prints.

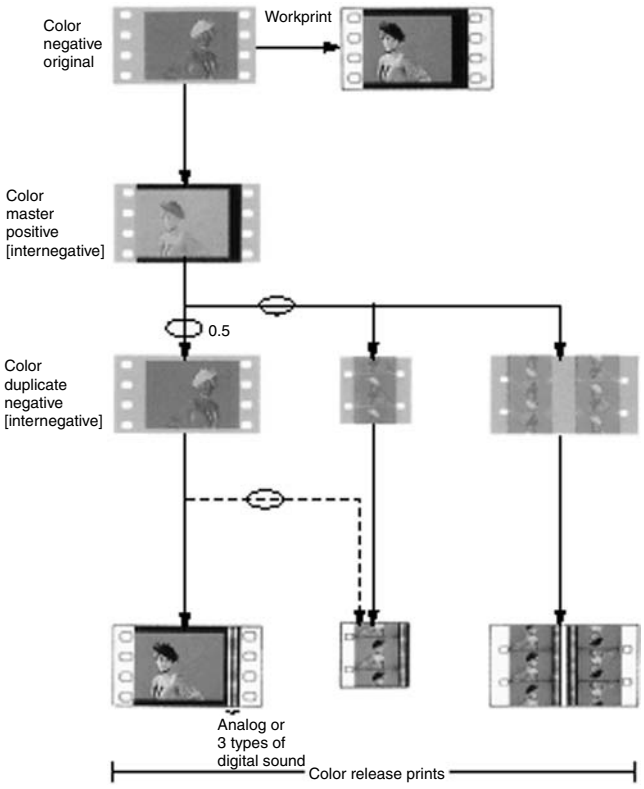


Figure 22. Steps in making color prints. 1. The choice of printing procedure depends on a number of factors, including the types of printing and processing facilities available and cost. Final print quality will depend on those factors. The dotted lines show alternate, less common, or less satisfactory methods. 2. The symbol ⬭ indicates optical printing. In general, if you want to reduce the print size, do it at the last practical stage to obtain the highest definition. 3. All film clips shown here have the base side up. 4. Optical or step contact. 5. Continuous contact or optical. Optical may produce the best quality (courtesy: Kodak Entertainment Imaging).

Sound Editing

Sound is critical to the success of a film. It can be routine, or it can have a powerful effect on the finished product. Sound, beginning in the production phase and carrying through the post production work, is an entire discipline that has its own digital technology. During the editing phase, sound effects are brought in. Again, this is an entire subject by itself that has its own skilled engineers and technicians.

Digital Intermediate. The entire negative footage is scanned digitally (1,920 pixels per line or higher). All of the editing and special effects are done digitally before recording the images back onto color intermediate film using a laser or cathode-ray tube recorder. This is expensive, but its use will increase, predictably, as digital costs decrease.

Animation

Animation, the art of making inanimate pictures or objects move, is increasingly used for motion pictures. This includes such subjects as highly acclaimed Walt Disney films, medical subjects, and industrial training films as well as educational subjects. For decades animation was accomplished by batteries of artists designing, drawing, and then hand coloring individual cels which are transparent triacetate sheets. Each frame would be slightly different where the elements of motion changed minutely. Background scenes can be on a different cel which would remain constant as the action element in front moved. A special animation camera would expose one frame of film at a time to photograph individual cels progressively. The result might be a cartoon film or some complex machine operation shown in slow motion. As computer software technology became more sophisticated, it was possible to develop individual cels much quicker than by hand. Animation art as a form of computer graphics is common.

PROJECTION

Motion Picture Projector

Essentially, there are two types of projectors. Both operate using the same basic configuration of picture head, optics, and shutter system. The major difference is in the film supply and take-up system.

1. All 16 mm, super and regular 8 mm plus many older style 35 mm projectors use film reels somewhat similar to the spool in a motion picture camera.

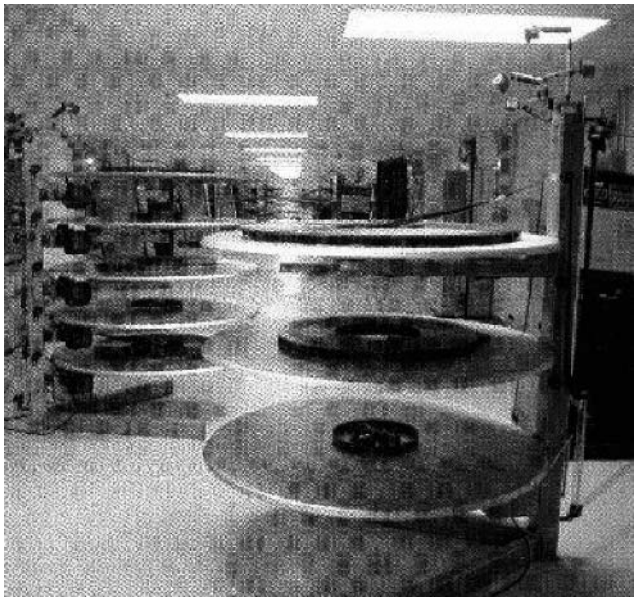


Figure 23. Platters for film in a modern multiplex theater (courtesy Strong International, Division of Ballantyne of Omaha, Inc.)

2. Most 35 mm and 70 mm projectors use two or more large flat platters; one has the film yet to be shown and another takes up the film as it leaves the central mechanism of the projector.

Theater projection operation is described following this technical portion.

All projectors have three essential elements:

1. picture head
2. lamphouse
3. sound reproducer

Picture Head. The picture head, the critical part of a projector, consists of the *film path*, the *picture aperture*, the *shutter*, and the *lens*. The film moves through the projector guided by two *constant-drive sprockets*. One feeds the film into the *film gate*, and the other pulls the film away from the gate and *intermittent*. The intermittent is a sprocket that does not turn continuously. It makes a quarter turn, moving a frame into the *trap* at the gate/aperture and holds the film for 3/96th of a second and then repeats the sequence 24 times in a second.

Sprockets. Sprockets are on small hubs, one mounted above the film gate to move the film from the supply reel or platter to the film gate. The second sprocket pulls the film from the gate to the take-up reel or platter.

Film Gate and Trap. The *film gate* consists of two small, rectangular indented plates, slightly wider than the film and nested together. The indentation is exactly the width of the frame, so that no metal portion touches the center of the film in the frame area. This prevents film scratches. Centered in both plates of the film gate is an aperture cut out to the exact size of the film frame. In the US most theater projectors have two aperture plates, one for an 1.85 : 1 aspect ratio and the other for a 2.39 : 1 scope ratio. In Europe, 1.66 : 1 is sometimes used.

Shutter. Like the camera, the projector has a shutter. The shutter revolution is timed so that the open portion of the shutter is in line with the opening (aperture) in the film gate, which then allows light to pass from the lamp through the film frame and on through the lens to the screen. As the shutter continues to revolve, the solid portion of the shutter blocks the beam of light while the film is moved to the next frame. Normally, 35 mm and 70 mm projectors have two blades. Each frame is projected twice. Some three-bladed shutters for 35 mm have reduced flicker but light transmission is reduced.

Shutters in 16 mm projectors have two, three, or five blades. A two-bladed projector gives maximum light. A three-bladed shutter has less flicker than the two blade but projects less light onto the screen. A five-bladed projector was used for television projection in old film chains.

Lens. Almost every theater projector has two lenses, some mounted on a turret. One is for the 1.85 : 1 image, and the other is usually an anamorphic lens for a 2.39 : 1 scope image.

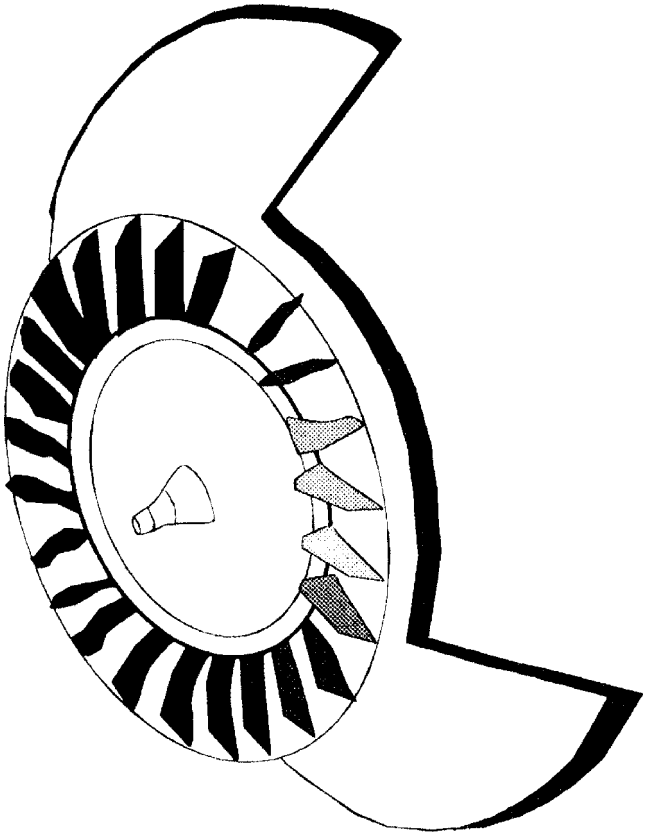


Figure 24. Two-bladed shutter used in all 35 mm and many 16 mm and Super 8 projectors for maximum light. Fins on shutter are used for balance in a 35 mm Simplex projector (courtesy Strong International, Division of Ballantyne of Omaha, Inc.).

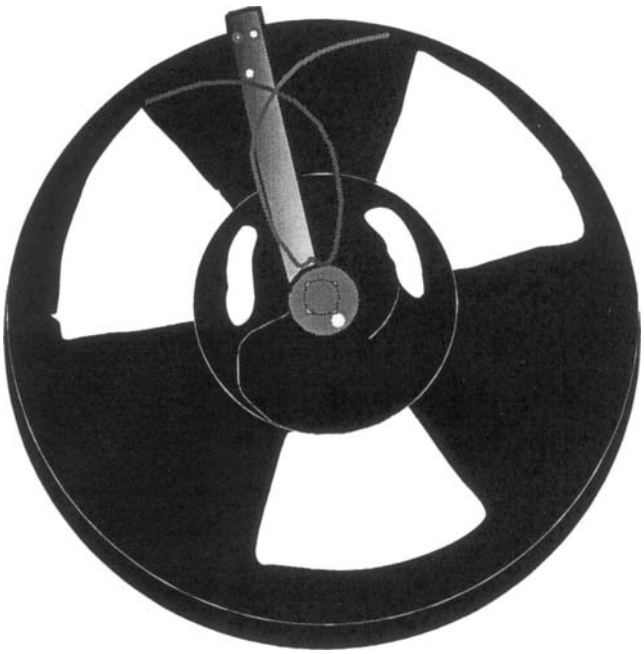


Figure 25. Three-bladed shutter used in Kodak Pageant 16 mm projector for minimum screen flicker (courtesy Eastman Kodak Co.).

Lens. There are a variety of different projection lenses. They range from quite simple lenses for standard 35 mm projectors, 16 mm and 8 mm, to highly sophisticated lenses such as those used for wide-screen shows. Some theater projectionists use a telescope or binoculars to focus the lens. *Sharpness* or *acutance* varies with viewing distance. The farther from the screen, the sharper the image. The projectionist is always far from the screen.

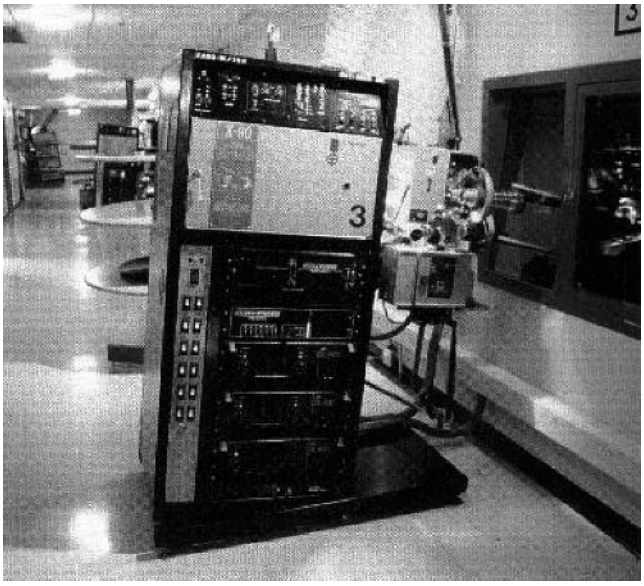


Figure 26. Simplex projector with Strong Console for xenon lamp (courtesy Strong International, division of Ballantyne of Omaha, Inc.).

Projection Lamps. There are hundreds of types of projection lamps. Each is designed for a specific projector model to provide maximum light output.

1. *Incandescent halogen lamps* are used in smaller projectors, principally 8 mm, some 16 mm projectors and some older 35 mm models.
2. *Xenon lamps* are in 95% of projectors used for large audiences in auditoriums and theaters. Essentially this includes all 35 mm and 70 mm projectors, many 16 mm and a few 8 mm.
3. A third light source is the *carbon arc*, which has been used in theater projectors from the early twentieth century up to the 1970s, when metal halide lamps came in. Carbon arcs are fast disappearing. Only the very oldest small theaters may still use the carbon arc.
4. Most video projectors use a *metal halide lamp*.
5. The newest high intensity lamp also used in video projectors for large audiences is the *ultra high pressure mercury lamp*.

Brightness. Screen luminance is measured in foot lamberts (fL). The Society of Motion Picture and Television Engineers (SMPTE) standard for screen luminance is 16 fL at the center of a uniformly lit screen when the projector is running with an *open gate* projecting white light. A range of 12 fL to 22 fL is within the tolerance. Checks of theaters reveals that many are below the range at 6 fL to 8 fL, thus projecting a *dull picture*.

Sound. Today, two basic types of sound are used for film projection.

Digital Sound. As described in Sound Development, there are three (Dolby, Sony SDDS, and DTS) proprietary digital systems in use in 2001.

Optical Sound. The older sound system used for films is called *optical sound*. A *sound track* is recorded photographically onto the film. All but some small theaters use digital sound. Analog sound is reserved only as a backup in case of print damage. In 35 mm and 70 mm film, the soundtrack lies between the perforations and the picture frame. In 16 mm and super 8 sound film prints, the perforations are omitted from one side of the film. The optical soundtrack is on the side of the picture frame opposite the perforations.

Magnetic Sound. A third system, *magnetic sound*, is nearly obsolete. A magnetic stripe was applied to the film in the same location on the print film as the optical track. A 35 mm print could have as many as four tracks or magnetic stripes for stereo sound. A 70 mm print would have as many as six stripes.

Sound Head

Every projector has a *sound head* that reproduces the sound from the *film sound track*. The head is always mounted below the film gate. The sound for analog is 19

frames ahead of the picture. This distance is standard for all sizes of film from 8 mm up to 70 mm. The offset for the different digital sound systems is proprietary. The sound head or sound scanner is opposite the *sound drum*, which is a flat roller designed to keep the film precisely positioned at the point where a scanning beam slit scans the sound track.

THEATERS

The movie theater has been part of our culture for nearly as long as the motion picture has existed. Theaters reflect the culture of their age. Movie palaces in the Roaring Twenties were extremely ornate, reflecting the architectural styles of the day. From the very beginning in the early part of the twentieth century, 35 mm film projectors have been the industry standard. At one point, an unsuccessful attempt was made to introduce 16 mm for small theater use. At the other extreme, the addition of the wide screen using 70 mm projectors represented a significant, successful move in the last half of the twentieth century. Until then, the typical theater used two 35 mm projectors that took 20-minute film reels. The projectionist had to be on his toes at all times. At nineteen minutes, a signal on the film would trigger a bell to alert the projectionist to be ready to switch projectors. He would light up the arc in the other projector and then watch for a dot in the corner of the picture. When he saw it, he would step on a pedal to move a *douser* that would put a screen in front of the lens and at the same time start the second projector. In addition to timing the projector change, the projectionist had to watch the arc lamp continually. As the arc burned, it became shorter. The projectionist periodically had to crank the arc unit so that it would be at the proper position to produce the brightest light. The arc is nearly a thing of the past.

Today's Movie Theater

Xenon lamp technology made great strides in the later half of the twentieth century. Probably 99% of the theaters now use a xenon lamp that emits light much brighter than the arc of the past without the nerve-racking anxiety. In addition to the advancement in projection lamps, three other major improvements have been made in movie theaters, two obvious to the audience, the other in the projection booth.

Platter System. The platter has eased the life of the projectionist. In modern theater design, one wide projection booth can serve as many as 18 small theaters in the same complex. Furthermore, several theaters can project the same film using the platter system. When the film has gone through one projector at theater A, it travels over series of rollers to theater B projector and then back to the take-up platter at theater A. The platter system is used for both 35 mm and 70 mm films.

Surround Sound. All newer theaters have sound speakers placed strategically around the theater. This allows sounds from different tracks on the film to come from different speakers around the theater.

Stadium seating. Most of the newest theaters no longer have seats sloping slightly toward the screen. Instead, each row of seats is on a step above the seats in front. In this way, even a child can see the screen easily without heads in front cutting the view line.

PRESERVATION — ARCHIVES (15)

Preserving films for posterity is vital. Many of the early Hollywood films have been lost due to improper storage that lead to deterioration, especially of the film base. Shrinkage is another problem caused by the loss of stabilizers and acetic acid (triacetate film). If that occurs, the distance lost between frames makes projection impossible. Adequate industry standards for film base degradation do not exist. Storage of completed motion picture films under controlled temperature and humidity is crucial. Properly stored (40°F at 50% relative humidity or lower) motion picture films can last at least 500 years,

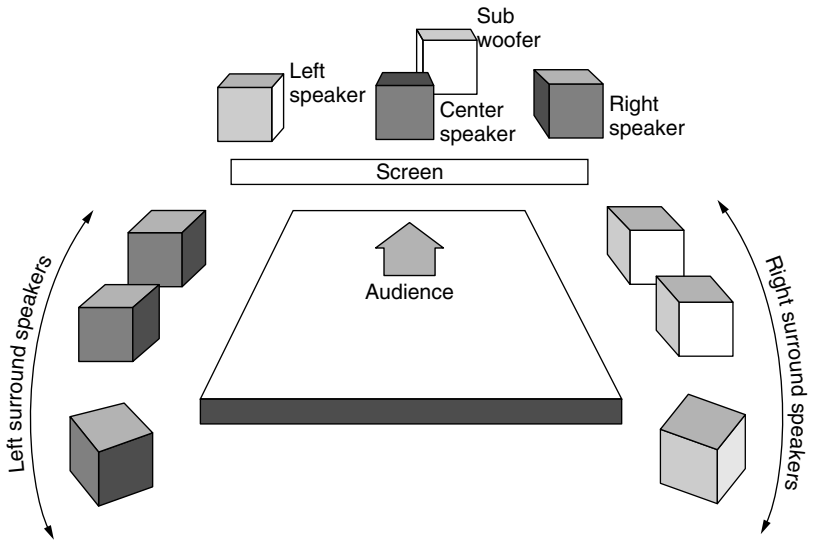


Figure 27. Surround sound using digital technology. The number of speakers varies depending on the individual theater.

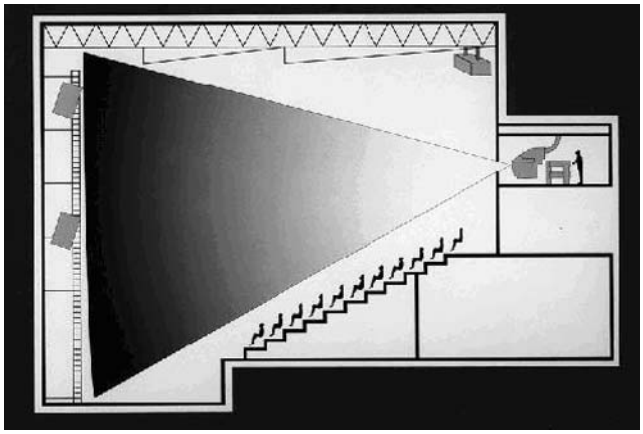


Figure 28. Stadium seating improves viewing from any seat in the theater (courtesy IMAX Corp.).

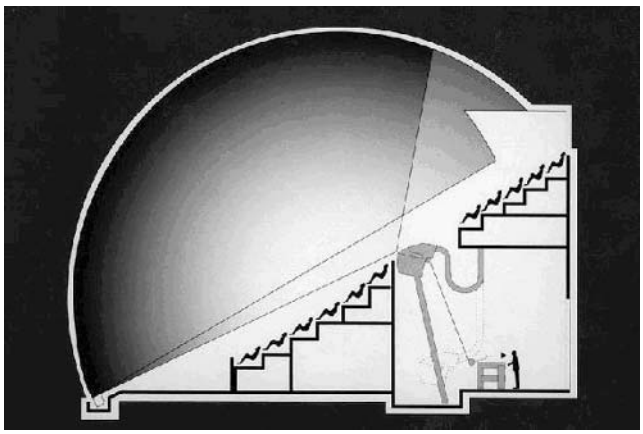


Figure 29. Stadium seating in a theater designed for wide-screen, 3-D, dome projection (courtesy IMAX Corp.).

and shrinkage is minimal. The colder the storage (not below freezing), the longer the film will last.

Nitrate base. Nitrate degradation occurs through a chemical process. Rusting of film cans is caused by gases emitted by the degrading nitrate and triacetate films. Nitrate film becomes highly flammable even at lower temperatures once the degradation has begun. The emulsion becomes adhesive, and layers stick together. The film contains gas bubbles and emits a noxious odor. As the film becomes soft, covered with a viscous froth, it eventually degrades into a brownish, acrid powder. Surprisingly, the 1903 classic, *The Great Train Robbery*, is still in perfect condition stored in the film vault of the Library of Congress.

Acetate Base. When acetate base begins to deteriorate, the chemical action causes a vinegar odor due to the evaporation of acetic acid from the film base. This is called the *vinegar syndrome*. There are acid detection strips (A-D strip) which can detect deterioration before the syndrome accelerates. An A-D strip placed in a film can for a day will change color. It can then be compared with colors on a chart to determine the stage of deterioration.

Polyester base. This rugged film base, when stored under the ideal conditions noted before, has a predicted life of at least 500 years.

Emulsion. The cold storage recommended for the various film bases will insure the stability of the chemicals and gelatin in the emulsions.

Containers. Films are stored in metal cans, plastic containers, and paper (cardboard) boxes. Many film archives tend to use cans for 35 mm and 70 mm films and plastic (if ventilated) for 16 mm and 8 mm films. Acid-free cardboard is good, but some archivists are concerned about potential water damage. The key is ventilation to remove gases. Molecular sieves in cans and unventilated vaults that can be replaced periodically are used.

DIGITAL CINEMA

Motion picture photography and projection in the Twenty-first century are undergoing a major evolution from film/chemical photography to electronic/digital cinematography, distribution, and exhibition. As of December 31, 2000, 30 movie theaters worldwide had been converted to digital cinema projection. Several feature films have been shot (photographed) using digital cameras.

The conversion to digital involves technical, procedural, and financial factors. Foremost are the *technical* aspects. The prime technology is in place. Digital cameras exist. Equipment for digital editing has been adapted from other computer systems used for high definition television, and better and more efficient equipment will undoubtedly be designed in the future. As mentioned, projectors for showing digital pictures are installed and being tested in theaters around the world.

Adaptation of the production process from systems developed over decades of film and HDTV production will have to be made to take advantage of digital technology. Editing requires a new system and, in some cases, the development of new equipment. Satellite and fiber optic delivery of a finished motion picture to theaters with proper security to prevent piracy is yet to be perfected. Product development and the manufacture of new types of equipment and installation are costly. Theater owners are extremely concerned about the huge expense of converting to digital technology which, in turn, could drive up the ticket price. Their direct competition is the home video market, where DVD media that have high resolution projection and displays are available.

During the first 75 years of the Hollywood movie business, changes and new product development were gradual. Toward the end of the twentieth century, the pace of change increased radically. As the introduction of new systems, products, and upgrades continues to increase, the rate of obsolescence becomes a major consideration in management planning. Given the chaotic situation of incompatible video formats, preserving electronic movies for the future is not assured any more than television programs. In the end, however, new technology prevails.

Acknowledgments

The author gratefully acknowledges the contributions by the following to the writing of Motion Picture Photography:

Grant Loucks, Alan Gordon Enterprises, Inc., Hollywood, CA

John Pytlak and Karen Dumont, Eastman Kodak Company, Rochester, NY

Robert Gibbons and Glenn Kimmel, Eastman Kodak Company, Hollywood, CA

Donald Roberts, Fox, WUHF-TV, Rochester, NY

Masaru Jibiki, Fuji Photo Film. Co., Los Angeles, CA

James Harte, film editor and projectionist, Rochester, NY

Kenneth Baker, IMAX Ltd., Toronto, ON, Canada

Philip Hill, iWERKS Moving Entertainment, Burbank, CA

Kathy Nifield, MegaSystems, Wayne, PA.

F. G. Bauer, Moviecam F. G. Bauer GesmbH, Vienna, Austria

John Walsh, NBC, WHEC-TV, Rochester, NY

Frank Kay, Panavision Corp., Woodland Hills, CA

Michael Ellis, Showscan, Culver City, CA

Phil Vigeant, Super8 Sound, Burbank, CA

Ray F. Boegner, Strong, Div. of Ballantyne of Omaha, Inc. Omaha, NB

Dr. Richard Goldman, Technicolor Inc., North Hollywood, CA

BIBLIOGRAPHY

1. Hope Reports, Inc.
2. *Encyclopedia Britannica.* (1940).
3. *Encyclopedia Britannica.* (2001).
4. J. Belton, *SMPTE J.*, p. 652–(August 1990).
5. *http://www.kodak/US/en/motion/support/types*
6. S. G. Rose, *J. SMPTE*, (August 1963).
7. *Student Filmmaker's Handbook*, Eastman Kodak Co. Rochester (1991).
8. J. Fauer, *Arriflex 435 Book*, Arri USA, (2000).
9. M. Mikolas and G. Hoos, *Handbook of Super 8 Production.* United Business, NY, (1976).
10. A. Kattelle, *Home Movies*. (2001).
11. F. E. Riley, Univ. of Kansas Master of Arts dissertation, (1996).
12. *http://www.widescreenmuseum.com//cinemaaspecs.htm*
13. A. L. Gaskill and D. A. Englander, *Pictorial Continuity*, Duell, Sloan and Pearce, NY, (1947).
14. Ross Lowell, *Matters of Light & Depth*, Lowell-Light Manufacturing, Inc., (1999).
15. James M. Reilly, *Film Decay And How To Slow It*, Image Permanence Institute, Rochester Institute of Technology, Rochester, NY, (1993).

GLOSSARY OF MOTION-PICTURE TERMS

Student Filmmakers Handbook, 1993
Courtesy of Eastman Kodak

A & B Cutting. A method of assembling original material in two separate rolls, allowing optical effects to be made by double printing (A and B printing).

A or B Wind. When a roll of 16 mm film, perforated along one edge, is held so that the outside end of the film leaves the roll at the top and toward the right, winding "A" should have the perforations on the edge of the film toward the observer, and winding "B" should have the perforations on the edge away from the observer. In both cases, the emulsion surface should face inward on the roll.

Academy Aperture. In projection, the aperture cutout, designed as specified by the American Academy of Motion Picture Arts and Sciences, that provides for a screen-image aspect ratio of approximately 1.37 : 1; also called "sound aperture."

Academy Leader. A nonprojected identification and timing countdown film leader designed to specifications of the American Academy of Motion Picture Arts and Sciences that is placed at the head end of a print reel. The countdown cuing information is related to "feet" which, in the silent days, meant projection at 16 frames per second, or 1 foot per second.

Acetate. A slow-burning base material frequently used for motion picture films. Also, in sheet form for overlay cels.

Acetate-base film. Any film that has a support made of cellulose triacetate; safety film.

Additive printing. The use of three separate colored sources — red, green, blue — combined to form the light source that exposes the film. Modem additive printers separate white light from a tungsten-halogen bulb into its red, green, and blue components by using a set of dichroic mirrors.

Advance. The separation between a point on the sound track of a film and the corresponding picture image.

Aerial image or virtual image. An image focused by a projection lens near a field or relay lens. A camera lens is then used to form a real image on the film from the aerial image. A cell or another material can be placed at the aerial-image location to combine it with the aerial image on film.

Analog. An electrical signal that is continuously variable.

Analytical density. Measurement of the amount of yellow, cyan, and magenta dye in an image.

Anamorphic image. An image that has been squeezed in one direction, usually horizontally, by an anamorphic lens.

Anamorphic lens. A lens that produces a "squeezed" image on film in the camera. When the film is projected on a screen, an appropriate lens reverses the effect, and the image spreads out to lifelike proportions. Designed for wide-screen movie photography and projection.

Anamorphic release print. A print in which the images are compressed horizontally.

Angle. With reference to the subject, the direction from which a picture is taken. The camera–subject relationship in terms of the immediate surroundings.

Animation. The making of inanimate objects to appear mobile. This can be done by exposing one or two frames

of movie film and then moving the objects slightly and exposing one or two more frames, etc. When the movie is projected, the objects will appear to have moved by themselves.

Animation camera. A motion picture camera that has special capability for animation work, which usually includes frame and footage counters, the ability to expose a single frame at a time, reverse-filming capability, and parallax-free viewing.

Ansi. American National Standards Institute.

Answer print. The first print (combining picture and sound, if a sound picture), in release form, offered by the laboratory to the producer for acceptance. It is usually studied carefully to determine whether changes are required before printing the balance of the order.

Antihalation backing (coating). A dark layer coated on or in the film to absorb light that would otherwise be reflected back into the emulsion from the base.

Aperture. (1) Lens: The orifice, usually an adjustable iris, which limits the amount of light passing through a lens. (2) Camera: In motion picture cameras, the mask opening that defines the area of each frame exposed. (3) Projector: In motion picture projectors, the mask opening that defines the area of each frame projected.

Aperture plate. A metal plate containing the aperture that is inserted into a projector or camera. (Note: In some cameras, the aperture plate cannot be removed.)

Arc lamp. A lamp whose light source consists of an open carbon arc or a closed xenon arc. The light is generated in a gas ball between two electrodes.

ASA. Exposure Index or speed rating that denotes the film sensitivity, defined by the American National Standards Institute. Actually defined only for black-and-white films, but also used in the trade for color films.

Aspect ratio. Proportion of picture width to height such as 1.37 : 1, 1.85 : 1, or 2.35 : 1.

Autoassemble. An operation in which a computer performs editing unaided, working from a previously edit decision list.

Backing. (1) Antihalation backing: A temporary, dark-colored, gelatin coating that is sometimes applied to the rear side of a photographic plate or film to reduce halation by absorbing any light that may pass through the emulsion. (2) Noncurl backing: A transparent, gelatin coating, sometimes applied to the side of photographic film opposite from the emulsion to prevent curling by balancing the forces that tend to curl the film as it is wet and dried during processing (3) Coating: (e.g., antiabrasion coating or rem-jet backing) applied to the base side of the film to improve characteristics and performance.

Balance stripe. A magnetic stripe on the edge of the film opposite from the magnetic track. Although the purpose of the stripe is to keep the film level on the reel, some projectors also use it for recording.

Barn door. A frame that has adjustable flaps attached to the studio light to control unwanted spill light or the spread of the light beam.

Barney. A lightweight padded covering that generally performs the same function as a blimp. Heated barneys are sometimes used to facilitate shooting under extremely cold outdoor conditions.

Base. The transparent, flexible support, commonly cellulose acetate, on which photographic emulsions are coated to make photographic film.

Bipack filming. The running of two films simultaneously through a camera or optical printer either to expose both or expose one through the other, using the one nearest to the lens as a mask. Often used in special-effects work to combine live action with animated images.

Black-and-white film. A film that produces a monochromatic picture in shades of gray (usually a metallic silver image).

Black light. Ultraviolet light.

Blimp. A soundproof enclosure that completely covers the camera to prevent camera-operating noise from being recorded on the sound track.

Blooping. The technique of applying a special opaque ink or tape in a wide triangular or circular pattern across the sound track at the splice to prevent soundtrack clicks and other annoying sounds caused by splices.

Blow up (part of frame). In transferring an image by an optical printer, it is possible to enlarge a properly proportioned fraction of the original image to full frame size in the copy or to enlarge an original 16 mm image to 35 mm size.

Blue-screen. The filming or videotaping of actors, props, or objects in front of a blue screen (or green screen). In postproduction, the blue or green is replaced by another element, such as background, using digital or optical special-effects techniques.

Boom. A long, adjustable arm used to position a microphone during production.

Bounce light. Light that is reflected off ceilings and walls to illuminate the subject indirectly.

Breakdown. The separation of a roll of camera original negative into its individual scenes.

Broad light. Soft, floodlight type of illumination unit; usually not focusable.

Butt splice. Film splice in which the ends come together without overlapping; ends are held together by splicing tape.

Camera axis. Any imaginary line running exactly through the optical center of the camera lens.

Camera log. A record sheet giving details of the scenes photographed on a roll of original negative.

Camera operator. In a feature production, the person who operates the camera under the direction of the DP-Director of Photography.

Camera operator-animator. The person responsible for translating instructions on the exposure sheet into camera moves and photographing the artwork.

Candela (cd). International unit of luminance measurement (1 candela per square meter = 0.2919 footlamberts).

Cel animation. An animation technique in which the figures to be animated are drawn and painted on cels, placed over a background, and photographed frame by frame. Cel animation has been the standard technique for studio animation since its invention in 1915.

Cellulose triacetate. Transparent, flexible material used as a base support for photographic emulsions.

Cement splice. Film splice made by using a liquid solvent cement to weld the overlapping ends together.

Changeover. In projection, the act of changing from one projector to another, preferably without interrupting the continuity of projection, or the points in the picture at which such a change is made.

Character animation. The art of making an animated figure move like a unique individual; sometimes described as acting through drawings. The animator must understand how the character's personality and body structure will be reflected in its movements.

Chromakey. A method of electronically matting or inserting an image from one camera into the picture produced by another. Also called "keying", the system uses a solid color background behind the subject to be inserted and signal processing through a special-effects generator.

Chrominance. The color portion of a video signal.

Cinch marks. Short scratches on the surface of a motion picture film, running parallel to its length; these are caused by dust or other abrasive particles between film coils or improper winding of the roll that permits one coil of film to slide against another.

Cinemascope. Trade name of a system of anamorphic wide-screen presentation. The first commercially successful anamorphic system for the presentation of wide-screen pictures combined with stereophonic sound. The 35 mm negative camera image is compressed horizontally by 50% using a special anamorphic camera lens. Upon projection, the 35 mm print image is expanded horizontally by the same amount using a similar anamorphic projection lens. Depending on the type of sound used in the print, the screen image has an aspect ratio of 2 : 35 : 1 (optical sound) or 2 : 55 : 1 (four-track magnetic sound).

Cinemiracle. A wide-screen presentation, as in Cinerama, that used three separate 35 mm film strips projected on a large, deeply curved screen. One of the main differences, however, was the consolidation of the three projectors in a single booth away from the audience. This was accomplished by using mirrors on the two outer projectors to maintain picture orientation.

Cineon digital film system. A new Kodak system that transfers images originated on film to a digital format for electronic compositing, manipulation and enhancement, and outputs back to film without losing image quality.

Cinerama. Originally, a wide-screen presentation using three separate 35 mm films, each containing one third of the total image (six perforations high) and projected on a deeply curved and vertically slotted screen from three projectors located in booths on the main floor of the auditorium. The sense of involvement was extraordinary, but the ever-present seams between the separate projected images were quite distracting. Current Cinerama presentations use 70 mm film containing a single image that is purposely distorted. During projection, the image distortion is corrected by the deeply curved screen, and the original Cinerama sensation is re-created.

Cinex strip. A short test print in which each frame has been printed at a different exposure level.

Circarama. A special presentation system used at Disneyland. The spectators stand in the middle of a circle viewing a 360° panorama on a surround screen 8 feet high and 40 feet in diameter made of eleven panels. The original negatives are made on eleven 16 mm cameras arranged in a concentric circle. The prints are projected by a ring of interlocked 16 mm projectors.

Clapsticks. Two boards hinged at one end that are slapped together to indicate the start of a filming session (take). Used by editors in conjunction with a slate, which provides the corresponding visual cue, to synchronize sound and image.

Claw. Mechanism used in most camera and projectors to move the film intermittently.

Clay animation. An animation technique involving the use of pliable clay figures that are manipulated before each exposure.

Close-up. A detail photographed from such a distance that only a small portion of the subject fills a frame of film.

Coated lens. A lens covered with a very thin layer of transparent material that reduces the amount of light reflected by the surface or the lens. A coated lens usually transmits more light than an uncoated lens at the same *f*-stop because flare is less.

Collimated. A beam of light is said to be collimated when all of its rays have been made parallel.

Color analyzer. A device for determining the correct printing light ratios for printing color negatives.

Color balance. The perceptual appearance of a color image of film as a function of the ratio of exposures of each of these primary color records on the film.

Color correction. Altering the color balance by modifying the ratio of the printing light values.

Color duplicate (dupe) negative. Duplicate of a negative color image made from a negative color original. Typically used for making release prints.

Color internegative. Negative-image color duplicate made from a positive color original. Typically used for making release prints.

Color negative. A negative (opposite) record of the original scene. Colors are the complementaries of the colors in the scene; light areas are dark, and dark areas are light.

Color positive. A positive record of the original scene.

Color print film. Film designed for making positive prints from color originals and color duplicates.

Color reversal film. Film that has a color positive image after processing. Can be an original camera film or a film on which other positive films are printed.

Color reversal intermediate. Color duplicate negative made directly from an original color negative by the reversal process.

Color saturation. The brilliance or purity of a color. When colors in a film image are projected at the proper screen brightness and without interference from stray light, the colors that appear bright, deep, rich, and undiluted are said to be saturated.

Color separation negative. Black-and-white negative made from red, green, or blue light from an original subject or from positive color film.

Color temperature. The color quality of the light source expressed in Kelvin (K). The higher the color temperature, the bluer the light; the lower the temperature, the redder the light.

Combined negative. Negative film containing the picture and the sound track.

Complementary color. Color that is minus one of the primary colors. Cyan is minus red — cyan and red are complementary colors; yellow is minus blue — yellow and blue are complementary colors; magenta is minus green — magenta and green are complementary colors. Produces white when mixed in equal parts with the primary color to which it is complementary.

Composite print. A print of a film that contains both picture and sound track. Films regularly shown in theaters are composite prints. Also called release print.

Composite video. A video signal in which the luminance and chrominance elements have been combined, as is NTSC, PAL, and SECAM.

Compound. The flat, tablelike part of an animation stand, on which the artwork rests while it is being photographed.

Conform. Match the original film to the final edited work print.

Contact print. Print made by exposing the receiving material in contact with the original. Images are the same size as the original images but have a reversed left-to-right orientation.

Continuity. The smooth flow of action or events from one shot or sequence to the next.

Continuous contact printer. A printing machine where the emulsion of the negative film is in direct physical contact with the positive raw stock emulsion, and the two films move continuously across the printing aperture.

Contrast. (1) The general term for describing the tone separation in a print in relation to a given difference in the light-and-shade of the negative or subject from which it was made. Thus, "contrast" is the general term for the property called "gamma" (Y), which is measured by making an H & D Curve for the process under study. (2) The range of tones in a photographic negative or positive expressed as the ratio of the extreme opacities or transparencies or as the difference between the extreme densities. This range is more properly described as "scale" or "latitude." (3) The ability of a photographic material, developer, or process as a whole to differentiate among small graduations in the tones of the subject.

Control strip. A short length of film containing a series of densities to check on laboratory procedures.

Core (film). A plastic cylinder on which film is wound, shipped, and stored.

Correction filter. A medium enabling a color change.

Coupler. A chemical incorporated in the emulsion of color film stocks that produces a dye image associated with the developed silver image.

Crane. The mounting that supports the camera over the compound.

Credits. Titles of acknowledgment for the production.

CRI. Color reversal intermediate, a duplicate color negative prepared by reversal processing.

Cropping. To change, delete, or otherwise alter the size of an image being projected or viewed as a print. In theatrical projection, it is usually the result of "homemade" aperture plates, improper screen masking, wrong focal length lenses, etc.

Cut. (1) The instantaneous change from one scene to another. Successive frames contain the last frame of one scene and the first frame of the following scene. (2) To stop operation of camera, action, and/or sound recording equipment. (3) To sever or splice film in the editing process.

Cutting. The selection and assembly of the various scenes or sequences of a reel of film.

Cyan. Blue-green; the complement of red or the minus-red subtractive used in three-color processes.

***D* log *H* curve.** The curve showing the relation between the logarithm of the exposure and the resultant density on processed film.

***D* log *E* (Density vs. the log of exposure).** The graph made by plotting the density of a film sample against the

log of the exposure that made that density. Also known as D log H and H and D curve. D log H (H for exposure) is the technically correct term.

Dailies. Picture and sound work prints of a day's shooting; usually an untimed one-light print, made without regard to color balance. Produced so that the action can be checked and the best takes selected; usually shown before the next day's shooting begins.

Daylight. Light consisting of a natural combination of sunlight and skylight (approximately 6,500 K).

Decibel (dB). Unit of loudness measured on a logarithmic scale. The human ear can perceive 1 dB changes in loudness in the aural range.

Definition. The clarity or distinctness with which detail of an image is rendered; fidelity of reproduction of sound or image.

Delrama. A wide-screen process compatible with CinemaScope-type presentations

Densitometer. Instrument used to measure the optical density of an area in a processed image by transmittance or by reflectance.

Density. Light-stopping characteristics of a film or a filter. The negative logarithm to the base 10 of the transmittance (or reflectance) of the sample. A sample which transmits 2 of the incident light has a transmittance of 0.50, or 50%, and a density of 0.30.

Depth of field. The range of object distances within which objects are in satisfactorily sharp focus in a photograph.

Depth of focus. The range through which a photographic film or plate can be moved forward and backward with respect to the lens while retaining satisfactorily sharp focus on an object at a given distance.

Designer. In studio animation, the person responsible for the overall look and style of the film.

Developer. A solution used to turn the latent image on exposed films into a visible image.

Dialogue. The portion of the sound track that is recorded by the voice artists and spoken by the characters on the screen.

Diaphragm. An adjustable opening mounted behind or between the elements or a lens used to control the amount of light that reaches the film. Openings are usually calibrated in f-numbers.

Dichroic. A type of coating that, when applied to glass, can produce a so-called "cold" mirror for use in projector lamphouses that permits greater screen brightness without the risk of radiant energy (heat) problems. Usually, the rear surface of the mirror is treated by depositing very thin layers of a special coating designed to transmit infrared (IR) radiation effectively and reflect visible radiation. Alternatively, by selecting certain other materials for the deposit, IR radiation can be reflected and visible radiation transmitted, thus providing an efficient heat filter for arc radiation devices.

Diffraction. The spreading of light as it passes the edges of opaque objects or through narrow slits. Light also is diffracted when passing through a lens. The effects of this distortion on images is greater as the aperture becomes smaller.

Diffusion. The spreading of light rays from a rough reflecting surface or by transmission of light through a translucent material.

Digital. A system whereby a continuously variable (analog) signal is broken down and encoded into discrete binary bits that represent a mathematical model of the original signal.

Digital effects. Special effects, such as picture compression, rotation, reversal, performed by a digital effects system.

Digital recording. Sound-recording process in which sound waves are recorded as digital bits. During playback, a digital-to-analog conversion occurs that changes the digital bits back into sound waves. Digital recording produces high-quality true sound that does not contain any system noise.

Dimmer. An electrical device, normally in the form of variable resistance or load, that reduces electrical energy to a lamp, usually by reducing voltage.

Director. The person who interprets a written book or script. The director oversees all aspects of the production.

Dissolve. An optical or camera effect in which one scene gradually fades out when a second scene fades in. There is an apparent double exposure during the center portion of a dissolve sequence where the two scenes overlap.

Distributor. Firm that sells, leases, and rents films.

Dolby system. Trade name for an audio noise-reduction system.

Dolly. (1) A truck built to carry camera and camera operator to facilitate movement of the camera while shooting scenes. (2) To move the camera toward or away from the subject while shooting a scene.

Double-frame. Identical views photographed twice (two frames) instead of once. This technique cuts in half either the speed of a movement or the number of drawings required for a complete action. Sometimes called "on twos."

Double-system recording. Synchronous sound recording on a recorder that is separate from the camera. Recorders are typically magnetic and have sync-pulse capability.

Dubbing. The combination of several sound components into a single recording.

Dupe, dupe negative. A duplicate negative made from a master positive by printing and development or from an original negative by printing followed by reversal development.

Dye. In photography, the result of color processing in which the silver grains or incorporated color couplers have been converted into the appropriate dye to form part of the color image.

Edge numbers (key numbers/footage numbers). Sequential numbers printed along the edge of a strip of film by the manufacturer to designate identification.

Edit. To arrange the various shots, scenes, and sequences, or the elements of the sound track, in the order desired to create the finished film.

Editor. The individual who decides what scenes and takes are to be used, how, where, in what sequence, and at what length they will appear.

(Edit decision list EDL). List of edits prepared during off-line editing.

Emulsion, emulsion layer. (1) Broadly, any light-sensitive photographic material consisting of a gelatin emulsion containing silver halides together with the base and any other layers or ingredients that may be required to produce a film that has desirable mechanical and photographic properties. (2) In discussions of the anatomy of a photographic film, the emulsion layer is any coating that contains light-sensitive silver halide grains, as distinguished from the backing, base, substratum, or filter layers.

Emulsion side. The side of a film coated by emulsion.

Emulsion speed. The photosensitivity of a film, usually expressed as an index number based on the film manufacturer's recommendations for the use of the film under typical conditions of exposure and development.

Encoder. A circuit that combines the primary red, green, and blue signals into a composite video signal.

***Estar* base.** The trade name applied to the polyethylene terephthalate film base manufactured by Eastman Kodak Company.

Exchange. A depository and inspection/distribution center for theatrical release prints. Exchanges are located in approximately 35 regional areas within the United States dependent roughly on theater and population density.

Existing light. Available light; strictly speaking, existing light covers all natural lighting from moonlight to sunshine. For photographic purposes, existing light represents the light that is already on the scene or project and includes room lamps, fluorescent lamps, spotlights, neon signs, candles, daylight through windows, and outdoor scenes at twilight or in moonlight.

Exposure. Amount of light that acts on a photographic material; product of illumination intensity (controlled by the lens opening) and duration (controlled by the shutter opening and the frame rate).

Exposure latitude. Degree to which film can be underexposed or overexposed and still yield satisfactory results.

Exposure setting. The lens opening selected to expose the film.

***f*-Number.** A symbol that expresses the relative aperture of a lens. For example, a lens having a relative aperture of 1.7 would be marked $f/1.7$. The smaller the f-number, the more light the lens transmits.

Fade. Exposure of motion picture film either in the camera or during subsequent operations, so that, for a fade-in, starting with no exposure and extending for a predetermined number of frames, each successive frame receives a systematically greater exposure than the frame preceding it, until full normal exposure for the scene has been attained. From this frame on, successive frames receive identical exposure for the remainder of the take.

Fast. (1) Having a high photographic speed. The term may be applied to a photographic process as a whole, or it may refer to any element in the process, such as the optical system, emulsion, or developer. (2) Resistant to the action of destructive agents. For example, a dye image may be fast to light, fast to heat, or fast to diffusion.

Field of view. The portion of the scene in front of the camera represented within the limits of the camera aperture at the focal plane. Area of field thus varies with the focal length of the lens and the camera-to-subject distance.

Fill light. Light used to fill in shadows.

Film (motion picture film). A thin, flexible, transparent ribbon perforated along one or both edges; it bears either a succession of images or a sensitive layer capable of producing photographic images.

Film base. Flexible, usually transparent, support on which photographic emulsions are coated.

Film can. Metal container designed to hold rolls, spools, or reels of motion picture films.

Film cement. A special combination of solvents and solids used to make overlap splices on motion picture film by its solvent action and subsequent welding of the film at the junction.

Film gate. Components that make up the pressure and aperture plates in a camera, printer, or projector.

Film gauge. Width of the standard sizes of motion picture films.

Film identification code. Letter that identifies film type.

Film-to-tape transfer. The process of transferring an image captured on film to videotape.

Film perforation. Holes punched at regular intervals for the length of film, intended to be engaged by pins, pegs, and sprockets as the film is transported through the camera, projector, or other equipment.

Filter. A piece of glass, gelatin, or other transparent material used over the lens or light source to emphasize, eliminate, or change the color or density of the entire scene or certain elements in the scene.

Final cut. Last editing of a work print before conforming is done or before sound work prints are mixed.

Fine grain. Emulsion in which silver particles are very small.

First print. The first trial composite (married) print containing both picture and sound for checking picture and sound quality.

Fixing bath (hypo). A solution that removes any unexposed silver halide crystals in the film. In addition, in color films, the silver is removed from the exposed area, leaving only the image-forming dyes.

Flaking. The removal (chipping away) of emulsion particles from the edges of the film that tend to redeposit in the image area while the film is going through the projector gate. Flaking is eased by proper edgewax lubrication.

Flange. The rim on a roller used for guiding the film. Also, a large disc used on a rewind to take up film on a core. A pair of flanges (disc) that screw together is called a split reel.

Flashing. Technique for lowering contrast by giving a slight uniform exposure to film before processing.

Flat. An image is said to be "flat" if its contrast is too low. Flatness is a defect that does not necessarily affect the entire density scale of a reproduction to the same degree. Thus, a picture may be "flat" in the highlight areas or "flat" in the shadow regions, or both.

Flutter. In sound, rapid periodic variation of frequency caused by unsteadiness of the film or tape drive.

Focal plane. The area in space on which parallel rays of light refracted through a lens focus to form sharp images.

Fog. Darkening or discoloring of a negative or print, or lightening or discoloring of a reversal material. Causes include accidental exposure to light or X rays, overdevelopment, using outdated film, and storing film in a hot, humid place.

Foley. Background sounds added during audio sweetening to heighten realism, for example, footsteps, bird calls, heavy breathing, and short gasps.

Follow focus. To change the focus setting of a lens as a scene is photographed to keep a moving subject in sharp focus.

Footage. A method of measuring film length and, therefore, screen time. Because 90 feet of 35 mm film equal one minute of screen time, 35 mm footage is used in many studios as a measure of an animator's weekly output. Animators also refer to the length of scenes in feet, rather than in seconds or minutes — a 30-foot scene, rather than a 20-second one.

Footlambert. US luminance measurement unit (1 footlambert = 3.425 candelas per square meter).

Force-process. Develop film for longer than the normal time to compensate for underexposure.

Foreground. The part of the scene in front of the camera, represented within the limits of the camera aperture, occupied by the object(s) nearest to the camera.

Format. The size or aspect ratio of a motion picture frame.

FPS. Frames per second, indicating the number or images exposed per second.

Frame. The individual picture image on a strip of motion picture film.

Frame-by-frame. Filming in which each frame is exposed separately, as the object being photographed must be altered before each exposure to create the illusion of movement in the finished film; as opposed to the more usual method of filming in which the film runs through the camera at a steady, prescribed rate to record action taking place before it.

Frame line. The separation between adjacent image frames on motion picture film.

Frame (video). A complete television picture made up of two fields, produced at the rate of approximately 29.97 Hz (color) or 30 Hz (black-and-white).

Freeze frame. An optical printing effect in which a single frame image is repeated so as to appear stationary when projected.

Frequency response. Ability of the photographic sound track to reproduce the full spectral range of sounds.

Gain, screen. The measure of a screen's ability to reflect the light incident on it. A perfect screen reflects back all of the light that is incident on it at all angles. Such a screen would have a gain of 1.0. In practical use, however, most matte screens that allow wide viewing angles have a gain of about 0.85. Special metallized or directional screens can provide up to about 15 times more reflected light than a common matte screen, but their viewing angles are generally very limited, making them unsuitable for most theatrical applications.

Gamma. Measurement of the contrast of an image, representing the slope of the straight-line portion of the characteristic curve.

Gate. The aperture assembly at which the film is exposed in a camera, printer, or projector.

Gauge. Refers to the format of the film stock, that is, super 8, 16 mm, or 35 mm.

Gelatin filter (gel). A light filter consisting of a gelatin sheet in which light-absorbing pigment or dye is incorporated.

Geneva movement. A mechanical device that produces intermittent film movement in a projector. The principle behind the movement involves a rotating cam and pin that intermittently engage in a four-slotted star wheel, also known as a Geneva cross or Maltese cross. During the pin/slot engagement, the star wheel shaft containing the intermittent sprocket rotates 90°, or one frame. At

normal projection speed, this intermittent rotation occurs 24 times per second.

Glove. A white, lintless, cotton glove used when handling motion picture raw stock and new release prints in the laboratory. Should be used in all film handling and changed frequently.

Gobo. Panel of opaque material on a footed stand that has an adjustable arm. Used to confine the area that a light illuminates or to keep light from shining directly into the camera lens.

Graininess. The character of a photographic image when, under normal viewing conditions, it appears to be made up of distinguishable particles or grains. This is due to the grouping together, or "clumping" of the individual silver grains, which are far too small to be perceived under normal viewing conditions by themselves.

Gray card. A commercially prepared card that reflects 18% of the light that hits it. Visually, it appears neutral or a middle gray halfway between black and white.

"Green" print. A newly processed print on which the emulsion may still be a little soft. If projected the first time without proper edgewax lubrication, perforation damage can result.

Guide rails. Vertical rails located on both sides of the projector trap that restrict lateral movement of the film as it passes through the projector gate.

Guide roller. Any flanged roller that is used to guide or restrict the position of motion picture film as it moves through a camera, projector, or printer.

Guillotine splicer. Device used for butt-splicing film using splicing tape.

Halation. A defect of photographic films and plates. Light forming an image on the film is scattered by passing through the emulsion or by reflection at the emulsion or base surfaces. This scattered light causes a local fog which is especially noticeable around images of light sources or sharply defined highlight areas.

Halide. Compound of a halogen, such as chlorine, bromine, or iodine.

Hard. (1) As applied to a photographic emulsion or developer, having a high contrast. (2) As applied to the lighting of a set, specular or harsh, giving sharp dense shadows and glaring highlights.

Hard light. Light made up of directional rays that creates strong, hard, well-defined shadows; sometimes called specular light.

Haze filter. These filter provide varying degrees of blue-light and green-light absorption.

HDTV. High definition television, a video format whose resolution is approximately twice that of standard television.

Head end, heads. The beginning of a reel where the film image is upside down when the film is threaded into a projector for showing.

Head-recording. On a tape recorder, printer, or projector, an electromagnet across which the tape or film is drawn and which magnetizes the coating on the tape base during recording.

Hertz (Hz). Unit of frequency; $1\ \text{Hz} = 1$ cycle per second.

High-speed camera. A camera designed to expose film at rates faster than 24 frames per second. Used to obtain slow-motion effects.

Highlights. Visually, the brightest, or photometrically the most luminant areas of a subject. In a negative image, areas of greatest density; in a positive image, areas of least density.

HMI lights. Metal halide lamps are fundamentally mercury arcs that have metal halide additives to adjust the color balance. Usually rated at approximately 5,400 K. For daylight-balanced films.

Hold. To freeze or stop the action. To achieve a hold in animation, the same cel or position of an object is photographed for several frames.

Hot. Referring to too much light in an area or to an excessively bright highlight.

Hue. Sensation of the color itself measured by the dominant wavelength.

Humidity. The moisture content of the air. For instance, low humidity describes conditions in a desert. Conversely, high humidity is related to tropical rain forest conditions.

Hypo (fixer). The name for a fixing bath made from ammonium or sodium thiosulfate, other chemicals, and water; often used as a synonym for fixing bath.

Idle roller. Free turning nonsprocketed rollers for guiding film through its appropriate path.

Image, latent image. The invisible image formed in a camera or printer by the action of light on a photographic emulsion.

Imbibition print. A color print produced by the transfer of magenta, cyan, and yellow dyed matrix films in register on a specially prepared clear film base or paper — Technicolor process and *Kodak* Dye Transfer process.

Intensity, light. The power (strength) of a light source; the total visible radiation produced by the light source.

Interlock. A system that electronically links a projector to a sound recorder; used during postproduction to view the edited film and sound track, to check timing, pacing, synchronization, etc.

Intermediate. Film used only for making duplicates from which other duplicates or prints are made. Does not include camera films.

Intermediate sprocket. An intermittently rotated sprocket that positions the film in the aperture of a projector and moves it after the exposure cycle.

Intermittent. Not continuous but equally spaced (sometimes random) motion, as the intermittent (24 fps) motion of film through a projector.

Internegative (dupe negative). Color negative made from a color negative. For making release prints.

Interpositive. A color master positive print.

In the can. Describes a scene or program which has been completed. Also, "That's a wrap."

Interlace. The manner in which a television picture is composed, scanning alternate lines to produce one field, approximately every 1/60 of a second in NTSC. Two fields comprise one television frame. Therefore, the NTSC television frame rate is approximately 30 fps.

Intermittent movement. The mechanism or a camera, printer, or projector by which each frame is held stationary when exposed and then advanced to the next.

ISO. International Standards Organization. The international counter part of ANSI.

K. Degrees kelvin, the unit of the color temperature scale.

Keykode number. Kodak's machine-readable key numbers, Includes 10-digit key number, manufacturer identification code, film code, and offset in perforations.

Keystoning. A geometric image distortion resulting when a projected image strikes a plane surface at an angle other than perpendicular to the axis of throw or when a plane surface is photographed at an angle other that perpendicular to the axis of the lens.

Kinescope. A film of a videotape made by shooting the picture on a specially designed television monitor. Also referred to as Kine.

Laboratory film. Film products, not intended for original photography, but necessary to complete the production process.

Latent image. Invisible image in exposed, undeveloped film; results from exposure to light.

Latent image edge numbering. Images placed on the edge of film products in manufacturing that become visible after development.

Latitude. In a photographic process, the range of exposure across which substantially correct reproduction is obtained. When the process is represented by an H and D curve, the latitude is the projection on the exposure axis of that part of the curve which approximates a straight line within the tolerance permitted for the purpose at hand.

Layout. A detailed drawing of a shot in which background elements, staging of the action, and camera moves are carefully worked out and plotted; the stage of production in which these are determined.

Leader. Any film or strip of material used for threading a motion picture machine. A leader may consist of short lengths of blank film attached to the ends of a print to protect the print from damage while threading a projector, or it may be a long length of any kind of film that is used to establish the film path in a processing machine before using the machine for processing film.

Lens. An optical device designed to produce an image on a screen, on a camera film, and in a variety of optical instruments. Also used to converge, diverge, or otherwise control light rays in applications not involving images.

Light axis. An imaginary line running exactly through the center of intensity of a light.

Light filter. A light-absorbing transparent sheet, commonly consisting of colored glass or dyed gelatin that is placed in an optical system to control the spectral quality, color, or intensity of the light passing a given plane.

Light intensity. Degree of light per unit falling on subject; usually expressed in footcandles.

Light meter. An electrical exposure meter for measuring light intensity.

Light output. The maximum power or energy delivered by a given light, concentrated by a spotlight, or spread out by a floodlight.

Lighting ratio. The ratio of the intensity of key and fill lights to fill light alone.

Light valve. Device for controlling the intensity and color quality of light on additive prints.

Lip sync. Simultaneous precise recording of image and sound so that the sound appears to be accurately superimposed on the image, especially if a person is speaking toward the camera.

Liquid gate. A printing system in which the original is immersed in a suitable liquid at the moment of exposure to reduce the effect of surface scratches and abrasions.

Live-action. The filming or videotaping of staged or documentary scenes of people, props, and locations.

Long shot (ls). The photographing of a scene or action from a distance or a wide angle of view, so that a large area of the setting appears on a frame of film and the scene or objects appear quite small.

Loop (continuous film). A section of film spliced end-to-end for use in printing, testing, dubbing, etc.

Low key. A scene is reproduced in a low key if the tone range of the reproduction is largely in the high density portion of the H and D scale of the process.

Lubrication. To reduce friction, required on processed print film for optimum transport and projection life.

Lumen. The measure of luminous flux (the rate at which light pulses are emitted or received). For instance, one candela of light covers a square foot of surface.

Luminance. The measured value of brightness; reflected light measured on motion picture screens as footlamberts or candelas per square meter.

Lux. Metric measure of illumination approximately equal to 10 footcandles (1 lux = 10.764 fc).

Magazine (projector). Enclosures on a motion picture projector which hold the reels of film.

Magenta. Purplish color; complementary to green or the minus-green subtractive primary used in the three-color process. Magenta light results when red and blue light overlap.

Magnetic sound. Sound derived from an electronic audio signal recorded on a magnetic oxide stripe or on full-coated magnetic tape.

Magnetic striping. The application of magnetic material on motion picture film intended for recording sound.

Magnetic tape/magnetic film. Usually (1/4)-inch plastic audiotape that has been coated with particles that can be magnetized. In film use, it is also used in various formats compatible with Super 8, 16 mm, 35 mm, and 70 mm films.

Magnetic track. Linear path of magnetically recorded audio signal on a magnetic film stripe or tape. The number of "mag tracks" can vary from one to six, depending on the picture format.

Magoptical. Sound track that has an optical track and one or more magnetic tracks.

Makeup table. A film handling unit that is one component of a platter system. It is used to assemble (makeup) the individual shipping reels into one large film roll on a platter for uninterrupted projection.

Masking. Restricting the size of a projected image on a screen by using black borders around the screen. Also, the restriction in size of any projected image or photographic print by using undercut aperture plates or masks and borders.

Master positive. Timed print made from a negative original and from which a duplicate negative is made.

Master. The final negative-reversal positive or intermediate film from which subsequent prints are made.

Matte. An opaque outline that limits the exposed area of a picture, either as a cutout object in front of the camera or as a silhouette on another strip of film.

Medium shot. A scene that is photographed from a medium distance, so that the full figure of the subject fills an entire frame.

Meter-candle. Unit of illuminance. The light received at a point one meter away from a point light source whose intensity is one candela (formerly candle).

Mid-foot key number. Full key number plus bar code, including 32-perforation (35 mm) offset, positioned halfway between each footage number. Will help identify short scenes without a key number. Uses a smaller type size to distinguish from 1-foot key numbers. Use a magnifying glass to read it easily.

Minimum density (D-min). Constant-density area in the tone of the characteristic curve where less exposure on negative film or more exposure on reversal film will produce no reduction in density. Sometimes called base plus fog in black-and-white film.

Mix. To combine the various sound tracks — dialogue, music, sound effects — into a single track.

Motorboating. The distracting sound heard when the film becomes misaligned over the sound drum and causes the sound scanning beam to "read" the film perforations instead of the sound track.

Moviola. A trademarked name for a machine that has a small rear-projection screen and the capacity to play back several sound tracks. Used in editing and for reviewing portions of a film during production. Also used to synchronize or interlock picture and sound track in editing. Newer devices called "flat-bed viewers" are slowly replacing upright Moviolas.

Multiplexer. Device or circuit used for mixing television signals to a single video recorder.

Narration. The off-screen commentary for a film; often referred to as "voice-over."

Negative film. Produces a negative image (black is white, white is black, and colors appear as complementaries).

Negative image. A photographic image in which the values of light and shade of the original photographed subject are represented in inverse order. Note: In a negative image, light objects of the original subject are represented by high densities, and dark objects are represented by low densities. In a color negative, colors are represented by their complementary colors.

Negative-positive process. Photographic process in which a positive image is obtained by developing a latent image made by printing a negative.

Neutral-density filters. Used to reduce the intensity of light reading the film without affecting colors.

Neutral test card. A commercially prepared card: One side has a neutral 18% reflection that has the appearance of medium gray. The other side has a neutral reflection of 90% and the visual appearance of stark white.

Newton's rings. Fuzzy, faintly colored lines in a projected image caused by high or uneven printer gate pressure.

Nitrate film. A highly flammable motion picture film that has not been domestically manufactured since around 1950. It is still present in large quantities in storage vaults and archives and must be very carefully stored to prevent spontaneous combustion, explosions, or other forms of destruction.

Noise. Unwanted sound in an audio pickup.

Notching. Practice of making a "V" cut to remove damaged perforations rather than removing the damage and making a splice. Not recommended because it weakens the film even more.

NTSC (National Television Systems Committee). Committee that established the color transmission system used in

the U.S. and some other countries. Also used to indicate the system itself consisting of 525 lines of information, scanned at a rate of approximately 30 frames per second.

Off-line editing. The process of creatively assembling the elements of a production, to communicate the appropriate message or story, and/or calculating the order, timing, and pace for user-friendly equipment such as film, (3/4) inch videotape, or nonlinear computer editing systems.

On-line editing. Final editing or assembly using the original master tapes to produce a finished program ready for distribution. Usually preceded by off-line editing. In some cases, programs go directly to the on-line editing suite. Usually associated with high-quality computer editing and digital effects.

Opaque. Of sufficient density that all incident light is completely absorbed (the opposite of transparent).

Optical effects. Trick shots prepared by using an optical printer in the laboratory, especially fades and dissolves.

Optical printer. Used when the image size of the print film is different from the image size of the preprint film. Also used when effects such as skip frames, blowups, zooms, and mattes are included.

Optical sound. System in which the photographic (optical) sound track on a film is scanned by a horizontal slit beam of light that modulates a photoelectric cell. The voltages generated by the cell produce audio signals that are amplified to operate screen speakers.

Original. An initial photographic image or sound recording — whether photographic or magnetic — as opposed to some stage of duplication thereof.

Original negative. The negative originally exposed in a camera.

Orthochromatic (ortho) film. Film that is sensitive only to blue and green light.

Out-take. A take of a scene which is not used for printing or final assembly in editing.

Overlap splice. Any film splice in which one film end overlaps the other film end.

Overlay. A technique in cel animation in which foreground elements of the setting are painted on a cel and placed over the characters to give an illusion of depth to the scene.

PAL (phase alternation by line). Color television system developed in Germany and used by many European and other countries. PAL consists of 625 lines scanned at a rate of 25 frames per second.

Pan. A camera move in which the camera appears to move horizontally or vertically, usually to follow the action or scan a scene. In animation, the effect is achieved by moving the artwork under the camera.

Panavision 35. A 35 mm process using 35 mm negative film and photographed through a Panavision anamorphic lens with a compression of 2×. Contact 35 mm prints are compatible with anamorphic systems such as Cinema-Scope.

Panchromatic (PAN) film. Black-and-white film which is sensitive to all colors in tones of about the same relative brightness as the human sees in the original scene. Film is sensitive to all visible wavelengths.

Parallax. In camera work, the viewfinder often is mounted so that its optical axis is at an appreciable distance from the optical axis of the camera lens, commonly resulting in inadvertent positional errors in framing.

Perforations. Regularly spaced and accurately shaped holes that are punched throughout the length of a motion picture film. These holes engage the teeth of various sprockets and pins by which the film is advanced and positioned as it travels through cameras, processing machines, and projectors.

Persistence of vision. The ability of the eye to perceive a series of rapid still images as a single moving image by retaining each impression on the retina for a fraction of a second, thus overlapping the images. This phenomenon makes it possible to see the sequential projected images of a motion picture as lifelike continuous movement.

Perspecta sound. A system of recording that produces a form of stereophonic reproduction by using a single optical sound track carrying three subaudible control tones that can shift the one-track sound source to the left, center, or right speakers by using the appropriate reproducing equipment. The system is compatible with normal single-track sound reproducers.

Photometer. An electro-optical device used to measure light intensity (a light meter).

Pin registration. A film term relating to the steadiness of the image. For optical and film-to-tape transfers, a pin-registered device holds each frame in position for a perfectly registered image, critical for creating multilayered special effects.

Pitch. (1) The property of sound that is determined by the frequency of the sound waves. (2) Distance from the center of one perforation on a film to the next, or from one thread of a screw to the next, or from one curve of a spiral to the next.

Pixel ("picture element"). The digital representation of the smallest area of a television picture, appearing as a tiny dot on the television screen. In a full color image, each pixel contains three components — a combination of red, green, and blue signals — reflecting the trichromatic nature of human vision. The number of pixels in a complete picture differs from one system to another; the more pixels, the greater the resolution.

Polarizing filter. Transparent material used to subdue reflections and control the brightness of the sky.

Polyester. A name for polyethylene terephthalate developed by E.I. Dupont de Nemours & Co. (Inc.) A film base

material exhibiting superior strength and tear characteristics. Cronar is the trade name for Dupont motion picture products; *Estar* Base is the tradename for Kodak products.

Positive film. Motion picture film designed and used primarily for making master positives or release prints.

Postproduction. The work done on a film such as editing, developing and printing, and looping, once photography has been completed.

Primary color. One of the colors of light — blue, red, or green — that can be mixed to form almost any color.

Printer lights. Incremental steps on an additive printer.

Print film. Film designed to carry positive images and sound tracks for projection.

Printed edge numbers. Edge numbers (usually yellow) placed on film at the laboratory by a printing machine.

Printing. Copying motion picture images by exposure to light.

Processing. Procedure during which exposed film is developed, fixed, and washed to produce either a negative or a positive image.

Process screen photography. The filming or videotaping of actors, props, or objects in front of a blue screen (or green screen). In postproduction, the blue or green is replaced by another element, such as a background, using digital or optical special-effects techniques.

Producer. The administrative head of the film, usually responsible for budget, staff, legal contracts, distribution, scheduling, etc.

Production. The general term used to describe the process involved in making all of the original material that is the basis for a finished motion picture. Loosely, a completed film.

Production supervisor. An assistant to the producer, who is in charge of routine administration.

Projection. The process of presenting a film by optical means and transmitted light for either visual or aural review, or both.

Projection speed. The rate at which the film moves through a projector; twenty-four frames per second is the standard for all sound films.

Pull-down claw. The metallic finger that advances film one frame between exposure cycles.

Push processing. A means of increasing the exposure index of film.

R-190 Spool. A 4.940-inch outside diameter metal camera spool. Square hole has single keyway, two offset round drive holes, one elliptical hole in both flanges. Side one and side two markings. For 200-foot, 16 mm film loads.

R-90 Spool. A 3.615-inch outside diameter metal camera spool. Square hole has single keyway in both flanges, and center hole is aligned on both flanges for 100-foot film loads. Used in 16 mm spool loading cameras.

Raster. The lines forming the scanning pattern of a television system. Scanned lines comprise the active portion of a video signal displayed on a cathode-ray tube (CRT).

Raw stock. Unexposed and unprocessed motion picture film; includes camera original, laboratory intermediate, duplicating, and release-print stocks.

Real time. The instantaneous response of a computer or device to instructions: the normal viewing time of any film or videotape program.

Reciprocity law. Expressed by $(H) = Et$, where E is light intensity, and t is time. When E or t are varied to the extreme, an unsatisfactory exposure can result.

Recommended practice. An SMPTE recommendation specifying good technical practice for some aspect of film or television.

Reduction printing. Making a copy of a film original on smaller format raw stock by optical printing; for example, printing a 35 mm original onto 16 mm stock for use in libraries, etc.

Reel band. A stiff paper strip that has a string loop tie that contains the release print number, title, and reel number and is used to keep the film snug on the shipping reel.

Refraction. The change of direction (deflection) of a light ray or energy wave from a straight line as it passes obliquely from one medium (such as air) to another (such as glass) in which its velocity is different.

Release print. In a motion picture processing laboratory, any of numerous duplicate prints of a subject made for general theater distribution.

Rem-jet backing. Antihalation backing used on certain films. Rem jet is softened and removed when processing starts.

Rendering. The simulation of light on three-dimensional objects; determining an object's surface characteristics, such as color and texture.

Resolution. The capacity of a medium to capture and playback distinctly fine details. 35 mm film is a high resolution storage medium; current videotape formats are low-to-moderate resolution media. Computers can perform at a wide range of resolutions, from the lowest to the highest, depending on hardware and software capabilities, and therefore are considered resolution independent.

Resolving power. Ability of a photographic emulsion or an optical system to reproduce fine detail in a film image and on a screen.

Reticulation. The formation of a coarse, crackled surface on the emulsion coating of a film during improper processing. If some process solution is too hot or too alkaline, it may cause excessive swelling of the emulsion, and this swollen gelatin may fail to dry down as a smooth homogeneous layer.

Reversal film. Film that processes to a positive image after exposure in a camera or in a printer to produce another positive film.

Reversal process. Any photographic process in which an image is produced by secondary development of the silver halide grains that remain after the latent image has been changed to silver by primary development and destroyed by a chemical bleach. In film exposed in a camera, the first developer changes the latent image to a negative silver image. This is destroyed by a bleach, and the remaining silver halide is converted to a positive image by a second developer. The bleached silver and any traces of halides may now be removed using hypo.

RGB. Red, green and blue, the primary color components of the additive color system used in color television.

Rotation. A camera move in which the camera is moved in a complete circle to give a spinning effect in the film. Partial rotation is called a tilt.

Rough cut. Preliminary stage in film editing, in which shots, scenes, and sequences are laid out in an approximate relationship without detailed attention to the individual cutting points.

Safety film. A photographic film whose base is fire-resistant or slow burning by various fire codes. At the present time, the terms "safety base film," "acetate base film" and "polyester base film" are synonymous with "safety film."

Saturation. Describes color brilliance or purity. When color film images are projected at the proper brightness and without interference from stray light, colors that appear bright, deep, rich, and undiluted are said to be "saturated."

Scene. A segment of a film that depicts a single situation or incident.

"Scope". A diminutive term used to describe any anamorphic projection system or film.

Scrim. A translucent material that makes hard light appear more diffuse or reduces, like a screen, the intensity of the light without changing its character.

Script. The text of a film that gives dialogue, action, staging, camera moves, etc.

Secam (System Electronique pour Colour Avec Memoire). The color television system developed in France and used there and in most of the former Communist-bloc countries and a few other areas, including parts of Africa.

Sensitivity. Degree of responsiveness of a film to light.

Sensitometer. An instrument with which a photographic emulsion is given a graduated series of exposures to light of controlled spectral quality, intensity, and duration. Depending upon whether the exposures vary in brightness or duration, the instrument may be called an intensity scale or a timescale sensitometer.

Separation light. A light that helps define the outline of a subject, thereby separating it from the background. Also called edge light, top light, rim light, backlight, hair light, skimmer, or kicker.

Separation masters. Three separate black-and-white master positives made from one color negative; one contains the red record, another the green record, and the third the blue record.

Sequence. A group of related scenes in a film that combine to tell a particular portion of the story and are usually set in the same location or time span.

Set. Derived from "setting." The prepared stage on which the action for three-dimensional animation takes place. A set may be as simple as a plain tabletop or as elaborate as props and decoration can make it.

Sharpness. Visual sensation of the abruptness of an edge. Clarity.

Short. The term usually refers to the cartoons made in the Hollywood studios during the 1930s, 1940s, and 1950s, which ran between 6 and 7 minutes long. Today, shorts range from $1\frac{1}{2}$ to more than 20 minutes in length and cover a variety of styles and subjects.

Shot. An unbroken filmed segment; the basic component of a scene.

Shutter. In theatrical projection, a two-bladed rotating device used to interrupt the light source while the film is being pulled down into the projector gate. One blade masks the pull-down while the other blade causes an additional light interruption, increasing the flicker frequency to 48 cycles per second, a rate that is not objectionable to the viewer at the recommended screen brightness of 16 footlamberts (55 candelas per square meter).

Sibilance. Excessive amount of vocal hiss when consonants such as "s" are spoken.

Single-frame exposure. The exposure of one frame of motion picture film at a time, like still photography. Commonly used in animation and time-lapse.

Silver halide. Light-sensitive compound used in film emulsions.

Single-perforation film. Film perforated along one edge.

Single-system sound. Sound on a magnetic or optical track that was recorded on the same strip of film on which the action was recorded.

16 mm film. Film 16 mm wide. May have single or double perforations.

Skip frame. An optical printing effect eliminating selected frames of the original scene to speed up the action.

Slitting. The act of cutting apart a film into narrower sizes of the required width.

Slow motion. The process of photographing a subject at a faster frame rate than used in projection to expand the time element.

Smpte. The Society of Motion Picture and Television Engineers.

Soft. The opposite of "hard." (1) As applied to a photographic emulsion or developer that has low contrast. (2) As applied to the lighting of a set, diffuse, giving a flat scene in which the brightness difference between highlights and shadows is small.

Soft light. Light made up of soft, scattered rays resulting in soft, less clearly defined shadows; also called diffuse light.

Sound drum. A flat roller in the sound head designed to keep the film precisely positioned at the point where the scanning beam slit scans the sound track. Also called the scanning drum.

Sound effects (foley). Sound from a source, other than the tracks, that bears synchronized dialogue, narration, or music: sound effects commonly introduced into a master track in the rerecording step, usually to enhance the illusion of reality.

Sound gate. The gate used in an optical sound head, instead of a sound drum, to keep the film sound track precisely aligned on the scanning beam slit during sound reproduction.

Sound head. The optical sound reproducer mounted beneath the projector head but above the take-up reel support arm or magazine.

Sound recorder. Device that may use audiotape, magnetic film, or motion picture film to record sound.

Sound sprocket. Any sprocket that pulls the film past the sound scanning beam slit.

Sound track. Photographic/optical sound track running lengthwise on 35 mm film adjacent to the edges of the picture frames and inside the perforations.

Special effect. A term broadly applied to any of numerous results obtained in the laboratory by combining and manipulating one or more camera records to produce an imaginatively creative scene different from that in front of the main camera. Making special effects may involve techniques such as double printing, fades, mattes, and vignetting.

Spectrum. Range of radiant energy within which the visible spectrum that has wavelengths from 400 to 700 mm exists.

Speed. (1) Inherent sensitivity of an emulsion to light. Represented by a number derived from a film's characteristic curve. (2) The largest lens opening (smallest *f*-number) at which a lens can be set. A "fast" lens transmits more light and has a larger opening and better optics than a "slow" lens.

Splice. Any type of cement or mechanical fastening by which two separate lengths of film are united end to end, so that they function as a single piece of film when passing through a camera, film processing machine, or projector.

Splicer. A mechanical device arranged for holding film in alignment and at the correct sprocket hole interval during the various operations required to join two pieces of film. It often includes a device for removing emulsion.

Splicing tape. Tape designed to make overlap or butt splices without the need for film cement or a mechanical fastener. Available in a variety of sizes, with or without perforations, and can be clear, translucent, or opaque orange.

Spool. A roll that has flanges on which film is wound for general handling.

Spotlight. A lighting unit; usually has a lens and shiny metal reflector that can be focused; produces hard light.

Sprocket. A toothed wheel used to transport perforated motion picture film.

Static electricity. Electric field that is present primarily due to the presence of electrical charges on materials.

Stereophonic. Sound recording and reproduction using multiple microphones and speakers, each of which has its own separate track; designed to simulate the actual sound and to achieve a three-dimensional effect.

Stock. General term for motion picture film, particularly before exposure.

Stop. The relationship between the focal length of a lens and the effective diameter of its aperture. An adjustable iris diaphragm permits using any ordinary photographic lens at any stop within its range. Sometimes used synonymously with *f*-number as in "*f*-stop". A unit of exposure change.

Stop motion. An animation method whereby apparent motion of objects is obtained on the film by exposing single frames and moving the object to simulate continuous motion.

Storyboard. A series of small consecutive drawings accompanied by captionlike descriptions of the action and sound, which are arranged comic strip fashion and used to plan a film. The drawings are frequently tacked to corkboards so that individual drawings can be added or changed in the course of development. The technique, invented at the Disney studio, is now widely used for live action films and commercials, as well as animation.

Stripe, magnetic. Narrow band(s) of magnetic oxide usually coated toward the edges of the base side of motion picture film for accepting audio signal recordings in the form of magnetic impulses.

Subtractive color. The formation of colors by removing selected portions of the white light spectrum by transparent filters or dye images.

Super panavision. Similar to Panavision 35, but photographed flat in 65 mm. The 70 mm prints produce an aspect ratio of 2.25 : 1 using four-channel sound and a ratio of 2 : 1 using six channel sound.

Superscope. A 35 mm anamorphic release print system adopted by RKO Radio Pictures that produced a screen image whose aspect ratio was 2 : 1 or 2.35 : 1 when projected by a normal anamorphic lens. The original camera negative was photographed flat, but special printing produced the anamorphic print.

Surround speakers. Speakers placed at the sides and at the rear of an auditorium to increase the realism of a stereophonic presentation or to produce other special effects.

Sweetening. Audio postproduction, at which time audio problems are corrected. Music, narration, and sound effects are mixed with original sound elements.

Swell. The increase in motion picture film dimensions caused by the absorption of moisture during storage and use in high humidity. Extreme humidity and subsequent swelling of the film increase the susceptibility of the film surfaces to abrasion.

Synchronizer. A mechanism employing a common rotary shaft that has sprockets which, by engaging perforations in the film, pass corresponding lengths of picture and sound films simultaneously, thus effectively keeping the two (or more) films in synchrony during editing.

Sync pulse. Inaudible timing reference recorded on the magnetic tape used in double-system recording. The source can be a generator in the camera cabled to the tape recorder or an oscillating crystal in the recorder when the camera also has a crystal. When the sound is transferred to magnetic film for editing, a resolver reads the reference and ensures that the tape runs at the same speed as that during shooting. In this way, the magnetic work print can be placed in sync with the images for which the original sound was recorded.

Synchronization. A picture record and a sound record are said to be "in sync" when they are placed relative to each other on a release print, so that, the action will coincide precisely with the accompanying sound when they are projected.

T-grain emulsion. Emulsion made up of tabletlike crystals, rather than conventional silver halide crystals; produces fine-grain high-speed films.

Tail ends, tails. The end of a film. The film must be rewound before projection if it is tails out.

Take. When a particular scene is repeated and photographed more than once to get a perfect recording of some special action, each photographic record of the scene or of a repetition of the scene is known as a "take." For example, the seventh scene of a particular sequence might be photographed three times, and the resulting records would be called Scene 7, Take 1; Scene 7, Take 2; and Scene 7, Take 3.

Tape splice. Film splice made using special splicing tape applied to both sides of the film.

Technirama. A 70 mm release print technique developed by Technicolor and printed from a horizontal double-frame 35 mm negative using 1.5 : 1 horizontal compression. The 35 mm reduction print image had a 1.33 : 1 compression ratio and produced a 2 : 1 aspect ratio when projected on equipment designed for CinemaScope. Some 70 mm prints were also available for showing at a 2.2 : 1 ratio.

Techniscope. A system designed to produce 35 mm anamorphic prints from a 35 mm negative whose images are approximately one-half the height of regular negative images and are produced by using a special one-half frame (two-perforation) pull-down camera. During printing, the negative image was blown up to normal height and squeezed to normal print image width to produce a regular anamorphic print that provided a projected aspect ratio of 2.35 : 1. The system was designed primarily to conserve negative raw stock.

Telecine. A device for scanning motion picture film images and converting them to standard videotape.

35 mm film. Film 35 mm wide that has four perforations on both edges of each frame. Image frame and sound-track area lie inside the perforations.

Thread. To place a length of film through an assigned path in a projector, camera, or other film handling device. Also called lacing.

3-D. The common term applied to three-dimensional (stereoscopic) images projected on a screen or viewed as a print. There have been several systems shown in theaters, but the discomfort attributed to the necessary eyewear, along with other equipment limitations has, more or less, relegated the present systems to the status of novelties.

Throw. In theatrical projection, the distance from the projector aperture to the center of the screen.

Tight wind. Relating to film wound tightly on a core or reel to form a firm roll that can be handled and shipped safely without danger of cinch marks or other damage to the film.

Time-lapse movie. A movie that shows in a few minutes or a few seconds, events that take hours or even days to occur; accomplished by exposing single frames of film at fixed intervals.

Timing. A laboratory process that involves balancing the color of a film to achieve consistency from scene to scene. Also includes adjusting exposure settings in duplication.

Todd-AO. A flat, 70 mm print system developed by Magna Pictures Corporation and American Optical Company to produce a 2.2 : 1 screen image of high resolution, sharpness, and brightness. The print was made from a 65 mm negative exposed in a specially designed camera. The extra width of the print film was intended to provide room for the six magnetic sound tracks contained on four magnetic oxide stripes. Todd-AO, considered the first commercially successful 70 mm film system, was introduced in 1955 with the release of *Oklahoma*.

Toe. Bottom portion of the characteristic curve, where slope increases gradually with constant changes in exposure.

Tone. That degree of lightness or darkness in any given area of a print; also referred to a value. Cold tones (bluish) and warm tones (reddish) refer to the color of the image in both black-and-white and color photographs.

Trailer. A length of film usually found on the end of each release print reel identifying subject, part, or reel number and containing several feet of projection leader. Also, a short roll of film containing coming attractions or other messages of interest.

Transition. The passage from one episodic part to another. Usually, film transitions are accomplished rapidly and smoothly, without loss of audience orientation, and are consistent with the established mood of the film.

Travelling matte. A process shot in which foreground action is superimposed on a separately photographed background by optical printing.

Triangle. A three-sided framework of wood or metal, designed to hold the three points of a tripod to limit their spread.

Trims. Manual printer controls used for overall color correction. Also, unused portions of shots taken for a film; usually kept until the production is complete.

Truck. A camera move in which the camera seems to move toward (truck in) or away from (truck out) the subject. The same effect in live action filmmaking is called zoom-in.

Tungsten light. Light produced by an electrically heated filament that has a continuous spectral distribution.

Type C. The SMPTE standard for the 1-inch nonsegmented helical videotape recording format.

Type K Core. 3-inch outside diameter, 1-inch inside diameter plastic core. Used with 1,000, 2,000, 3,000, and 4,000-foot lengths of negative, sound, and print films.

Type T Core. 2-inch outside diameter, 1-inch inside diameter plastic core. Used with most 16 mm films up to 400 feet.

Type U Core. 2-inch outside diameter, 1-inch inside diameter plastic core. Used with various length camera negative, sound, and print films.

Type Y Core. Similar to K but of heavier construction. Used for most color print films.

Type Z Core. 3-inch outside diameter, 1-inch inside diameter plastic core. Used for camera and print films in rolls longer than 400 feet.

Ultrasonic cleaner. Device that transfers ultrasonic sound waves to a cleaning liquid or solvent that dislodges embedded dirt on objects immersed in it.

Ultraviolet light. Energy produced by the (invisible) part of the electromagnetic spectrum whose wavelengths are 100 to 400 nanometers. Popularly known as "black light." UV radiation produces fluorescence in many materials.

Universal leader. A film projection leader, designed for the current projection rate of 24 frames per second (1 1/2 feet per second) and recommended for use on all release prints. It was designed to replace the Academy leader originally conceived when the motion picture projection rate was 16 frames per second.

Unsqueezed print. A print in which the distorted image of an anamorphic negative has not been corrected for normal projection.

Valve rollers. A cluster of three or four small rollers located at the entrances of the film magazines and designed to prevent fire from reaching the film reels. Because nitrate prints are now quite rare and are actually unlawful to use in some areas of the United States, the use of valve rollers, or fire rollers, is no longer essential on domestic projection equipment.

Variable-area sound track. Photographic sound track consisting of one or more variable-width transparent lines that run the length of a motion picture film within the prescribed sound track area. The most common type of track.

Variable-density sound track. Photographic sound track that has a constant width but varies in density along the length of a motion picture film within the prescribed sound track area. No longer used in motion picture productions.

Viewer. A mechanical and optical device designed to permit examining an enlarged image of a motion picture film during editing.

Vignetting. The partial masking or blocking of peripheral light rays either by intent or by accident. In theatrical projection, the blockage of peripheral light rays in a projection lens due to a lens barrel that is too long or to a lamphouse optical system that is not correctly matched to the limiting aperture of the projection lens. In photography, the intentional masking of peripheral light rays to soften and enhance a photograph.

VITC (vertical interval time code). Time code that is recorded in the vertical blanking interval above the active picture area. Can be read from videotape in the "still" mode.

Voice-over-narration. A sound and picture shot relationship in which a narrator's voice accompanies picture action.

Weave. Periodic sideways movement of the image as a result of mechanical faults in camera, printer, or projector.

Wet-gate printer. Printer in which the film passes through fluid-filled pads just before exposure. Released fluid temporarily fills film scratches with a solution that has the same refractive index as the film base, thereby eliminating scratch refraction and ensuring that the scratches will not appear on the printed film.

Wide screen. General term for form of film presentation in which the picture shown has an aspect ratio greater than 1.33 : 1.

Wild. Picture or sound shot without synchronous relationship to the other.

Winding. Designation of the relationship of perforation and emulsion position for film as it leaves a spool or core.

Wipe. Optical transition effect in which one image is replaced by another at a boundary edge moving in a selected pattern across the frame.

Work print. Any picture or sound track print, usually a positive, intended for use in editing to establish the finished version of a film through a series of trial cuttings. The purpose is to preserve the original intact (and undamaged) until the cutting points have been established.

Xenon arc. A short arc contained in a quartz envelope in which dc current, flowing from the cathode to the anode, forms an arc in a positive (high-pressure) atmosphere of xenon gas. The spectral distribution in the visible range closely resembles that of natural daylight.

Yellow. Minus-blue subtractive primary used in the three-color process.

Zero-frame reference mark. Dot that identifies the frame directly below as the zero-frame specified by both the human-readable key number and the machine-readable bar code.

Zoom-in. Continuous change of the camera lens focal length, which gradually narrows down the area of the picture being photographed and gives the effect of continuously enlarging the subject.

Zoom-out. Continuous change of the camera lens focal length, which gradually enlarges the area being photographed, and gives the effect of a continuously diminishing subject.

N

NEUTRON IMAGING, RADIOGRAPHY, AND CT

GRAHAM C. SMITH
Instrumentation Division
Brookhaven National Laboratory
Upton, NY

INTRODUCTION

The discovery of the neutron occurred in 1932, and the first neutron diffraction experiments were performed in the late 1940s. As usage progressively increased, the neutron has become a key probe in many fields of science, and sources around the world are now oversubscribed by experimenters. The European Spallation Source (ESS), and the Spallation Neutron Source (SNS), in the United State are new, high-intensity sources under development that attest to the importance of neutrons as a probe. Because they are uncharged, neutrons are one of the most difficult particles to detect. Only a few elements in nature have a high affinity for neutron interaction. This article is organized as follows. In the next section, the fundamental reactions upon which all neutron detection relies are described. Then a series of sections describe the main techniques of neutron detection and imaging: gas-based, scintillator based, semiconductor-based, film and image plates, and microchannel plates. In these sections, a few applications are noted. The section next to last focuses on neutron radiography/tomography, for which scintillator detectors and image plates are the primary imaging techniques. Some important reviews from recent literature for imaging techniques are: (1–5) and for radiography (6,7). Techniques for (non imaging) neutron detection have also recently been reviewed (8), and an excellent treatise on all neutron detectors is given in 9).

Neutron Interaction Mechanisms

The creation of free charge carriers is the primary process upon which most radiation detectors rely for their operation. Neutrons have no net charge and can be detected either by direct collision with, and displacement of, nuclei, or by nuclear interactions. For thermal neutrons, there is insufficient energy available for collision displacements, and nuclear reactions are the dominant process in neutron detection methods. Giving definitions for cold, thermal and epithermal neutrons, Fig. 1 shows the neutron absorption cross section versus energy for the most important capture reactions, which can be loosely categorized in three groups:

1. Reactions whose cross sections decrease as the square root of the neutron energy:

$$^{3}\mathrm{He} + {}^{1}\mathrm{n} \rightarrow {}^{3}\mathrm{H} + {}^{1}\mathrm{H} + 0.764\ \mathrm{MeV}$$
$$^{6}\mathrm{Li} + {}^{1}\mathrm{n} \rightarrow {}^{4}\mathrm{He} + {}^{3}\mathrm{H} + 4.8\ \mathrm{MeV}$$
$$^{10}\mathrm{B} + {}^{1}\mathrm{n} \rightarrow {}^{7}\mathrm{Li} + {}^{4}\mathrm{He} + 2.3\ \mathrm{MeV} + 0.48\ \mathrm{MeV}(\gamma)$$

2. Reactions that are (n, γ) resonances in which γ-ray emission is inhibited and energy is transferred to the orbital electrons. This class of reactions gives rise to complex γ-ray and conversion electron spectra (10):

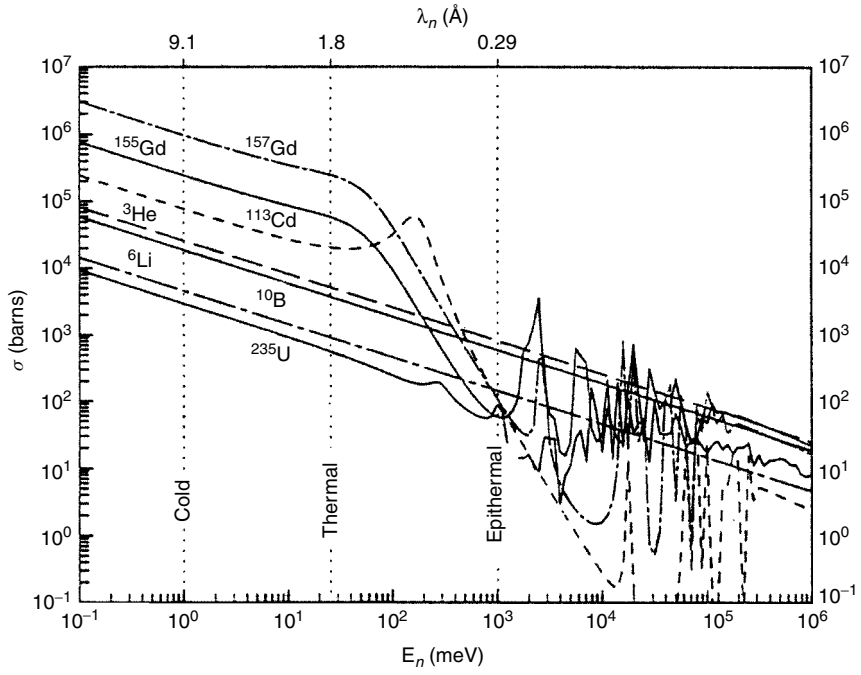


Figure 1. Cross sections for the most important neutron capture reactions versus neutron energy.

$^{155}\mathrm{Gd} + {}^{1}\mathrm{n} \rightarrow {}^{156}\mathrm{Gd}^{*}$

$\rightarrow {}^{156}\mathrm{Gd} + (\gamma + \text{conversion e's; 7.9 MeV})$

Main electron energies are 39 and 81 keV.

$^{157}\mathrm{Gd} + {}^{1}\mathrm{n} \rightarrow {}^{158}\mathrm{Gd}^{*}$

$\rightarrow {}^{158}\mathrm{Gd} + (\gamma\text{'s} + \text{conversion e's; 8.5 MeV})$

Main electron energies are 29, 71, 78, and 131 keV.

3. Fission of ^{235}U or ^{239}Pu:

$$^{235}\mathrm{U} + {}^{1}\mathrm{n} \rightarrow \text{ fission fragments} + 80 \text{ MeV}$$
$$^{239}\mathrm{Pu} + {}^{1}\mathrm{n} \rightarrow \text{fission fragments} + 80 \text{ MeV}.$$

A common feature of the reactions in (1) is that the products are ejected colinearly, and generally give rise to ionizations within several micrometers (solid) or millimeters (gas) of the reaction point; the range in lithium is about 100 μm. The (n, γ) reactions in (2) have a high probability of internal conversion, the conversion electrons ejected are much more easily stopped than γ rays and provide much more accurate positional information. Although ^{113}Cd has a high neutron capture cross section (Fig. 1), it has not been included in (2) because it has a low probability of internal conversion (about 4% of that for Gd) and accurate positional information is difficult to obtain. The reactions in (3) are based on nuclear fission. Their energies are very high, but the relatively low cross sections and difficulties in isotope enrichment have resulted in little use of these reactions. A compendium of neutron cross sections for many elements in the periodic table has been published in (11).

Two of the most important characteristics of a neutron imaging device are positional resolution and efficiency. The (useful) isotopes in (1) and (2) before are shown in Table 1 in their most common physical forms. Information pertaining to reaction product energies and ranges, and absorption depth for neutrons of energy 0.025 eV (1.8 Å) is also given. These numerical values are most valuable in determining the isotope's use in position-sensitive detectors.

DETECTORS USING GAS FILLING

Before gas-filled imaging detectors were developed, the usual component in single-wire proportional counters for neutron detection was boron trifluoride (BF_3). In recent years, ^{3}He has become the converter gas of choice because BF_3 is not an ideal gas in which to multiply electrons, and even enriched in ^{10}B, the absorption cross section is significantly smaller than that of ^{3}He.

The gas depth in a thermal neutron proportional chamber is typically 10–15 mm. This is a compromise between achieving high efficiency and minimizing positional error due to parallax for neutrons incident away from the normal to the detector. Figure 2 shows the calculated detection efficiency at 15 mm for a range of gas pressures. Extremely high detection efficiencies can be obtained for cold neutrons.

Ultimately, the major factor limiting positional resolution for neutron conversion in ^{3}He is the range of the proton and triton. Figure 3 shows the relative dE/dx profile of the proton and triton as a function of distance from the interaction point. The proton takes three-quarters of the reaction energy and also deposits a disproportionate quantity of ionization at the end of its track. The center of gravity of the resulting primary ionization is weighted part way along the proton track, about 0.4 R_p from the interaction point, where R_p is the range of the proton. A key parameter in obtaining high resolution is to ensure that the proton and triton ranges are kept small. This is normally achieved by using a heavier gas such as propane, (C_3H_8), or carbon tetra-fluoride, (CF_4), as an admixture with the ^{3}He. These gases also act as a quench for the electron avalanche.

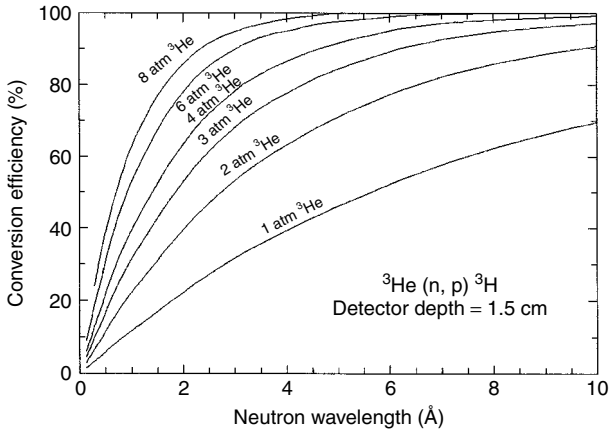


Figure 2. Calculated thermal neutron conversion efficiency for a range of ^{3}He partial pressures.

Table 1. The Most Useful Neutron Absorbing Isotopes and the Key Features of the Neutron Interactions at 25 meV

Isotope	State	σ_n[a] (barns)	Neutron abs^n. Length	Particle energies (keV)		Approximate Particle Range
^{3}He	Gas	5,330	70 mm.atm	p: 573	t: 191	3.8 mm.atm C_3H_8
^{6}Li	Solid	940	230 μm	t: 2,727	α: 2,055	130 μm
^{10}B	Solid	3,840	20 μm	α: 1,472	^{7}Li: 840	3 μm
$^{10}BF_3$	Gas	3,840	97 mm.atm	α: 1,472	^{7}Li: 840	4.2 mm.atm
Nat. Gd	Solid	49,000	6.7 μm	Conv^n electrons: ~30-200		12 μm
^{157}Gd	Solid	254,000	1.3 μm	Conv^n electrons: ~30-200		12 μm

[a] Cross sections, σ_n, are tabulated to the nearest 10 barns, and to the nearest 1,000 barns for Gd. This helps emphasize the huge stopping power of Gd relative to the other isotopes.

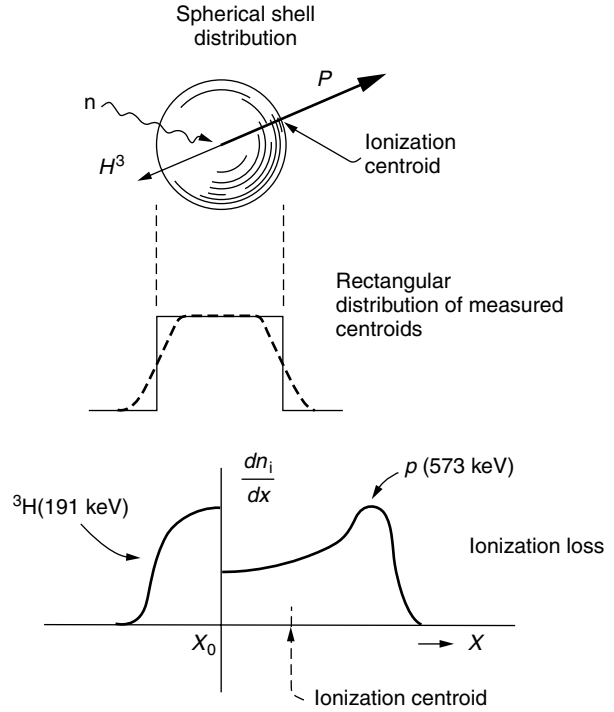


Figure 3. Representation of the sphere of centroids from many conversions. The ionization centroid is displaced from the conversion point by about $0.4 \times R_p$, the proton range.

Figure 4 shows the manner in which propane improves the positional resolution as a function of its pressure (12).

Multiwire Proportional Chambers (MWPCs)

The concept of the multiwire chamber was born over 30 years ago (13), but its application in a variety of radiation detection applications is ubiquitous even today; neutron detection is no exception. Figure 5 shows schematically a typical ^{3}He filled proportional chamber. Neutrons convert in the gas in the absorption region. The resulting primary ionization then drifts to the anode plane to undergo multiplication. Positional information in X and

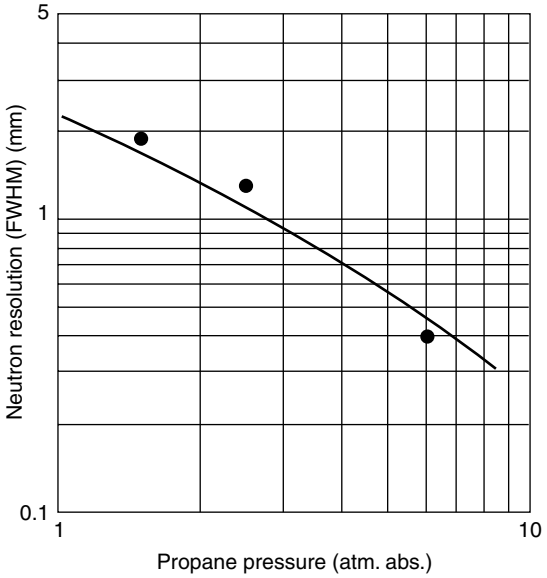


Figure 4. Positional resolution in $^3He + C_3H_8$ as a function of absolute propane pressure.

Y coordinates can be found from the cross-wire cathodes by resistive charge division. Refinement of charge division schemes has led to very good signal-to-noise ratios (14,15), so that the limit to positional resolution is solely due to the range of the reaction products shown in Fig. 3.

Gas proportional chambers can be fabricated in a wide range of sizes and configurations, a feature that has made them particularly popular in imaging experiments at the world's reactor and spallation sources. A report in (16) detailing detectors at these installations indicated that nearly all were some form of gas detector. Figures 6 and 7 show two detectors from Brookhaven National Laboratory at opposite extremes of the size scale. The first is a planar detector whose sensitive area is 5×5 cm, and the second is a curved detector whose sensitive area is 1.5×20 cm.

Gas-filled detectors are extraordinarily stable for long periods of time when fabricated correctly, and possess very good absolute count rate repeatability. These

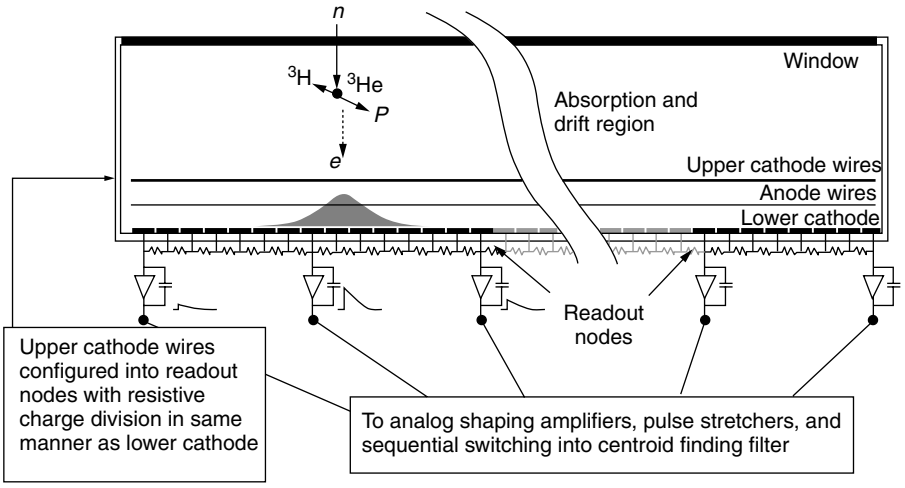


Figure 5. Side view of proportional chamber that has position readout by resistive charge division.



Figure 6. This 5 × 5 cm detector has the world's highest neutron resolution in gas, <400 μm FWHM.

features permit experiments that would otherwise be almost impossible to carry out. In a study of crystalline domains (17), a nonionic detergent–water mixture was investigated at constant temperature by time-resolved neutron scattering. Structural oscillations were observed at 5-min intervals by recording the Bragg peaks that speckle the Debye–Scherrer ring. One of the many time frames recorded during several days is shown in Fig. 8. It is the gas detector's ability to record the time dependence of the intensity and position that makes a study of this nature unique. Recent examples of gas detector results in structural biology are reported in (18).

Discrete Element Counters

Single-wire proportional counters (from which the MWPC ultimately was derived) constitute one of the most basic types of neutron detector. They comprise a cylindrical tube (of the order 1 cm in diameter) and have a central anode (which can be 1 m or longer), filled with the same gas mixture as the MWPC. An event's position is determined from resistive charge division along the complete length of the anode, using charge amplifiers at each end to measure the signal level (19). Although the readout is relatively simple, good positional resolution cannot be obtained from long counters (in contrast to the subdivided readout described for MWPCs), the counters have variable efficiency due to their circular cross section, and generally they have modest counting rate capability. In arrays, they can be used to cover large areas but generally are less superior to a MWPC that has a uniformly thick gas volume.

Microstrip Gas Chambers (MSGCs)

These were developed originally at the Institute Laue Langevin (ILL), Grenoble (20). In the MSGC, the anode and cathode wires of a MWPC are replaced by metallic strips on a glass or plastic substrate. These electrodes can be registered with high precision by photolithography, potentially leading to more automated fabrication processes, better positional resolution across the anodes, and higher count rates.

A section of an anode bounded on each side by a cathode is shown in Fig. 9. The anode, typically about 10 μm wide, has a field in excess of 3×10^5 V/cm at its surface.

Detectors based on the MSGC technology have been developed extensively for neutron detection at ILL (21). Figure 10 shows one arrangement of a detector. The Micro-Strip (MS) plate is made from Schott S8900 glass, a material that offers long-term resistance to charging problems. The front side, made of a 1,500 Å chromium layer, has a pitch of 1 mm, anodes of 12 μm, and cathodes of 500 μm. Induced signals on these cathodes determine one positional coordinate of an event. On the rear side, 86 chromium electrodes, 960 μm wide, are placed perpendicularly to the front structure to determine the second positional coordinate.

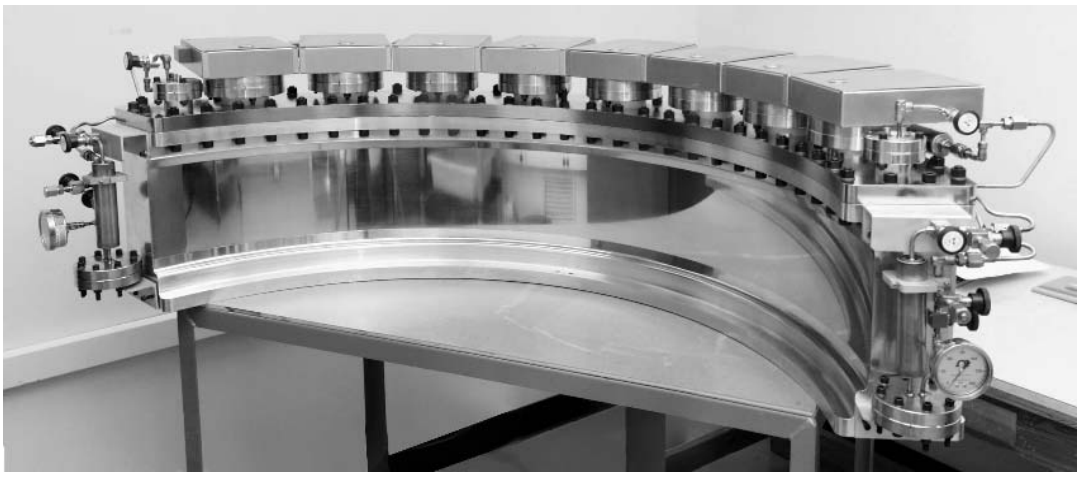


Figure 7. Curved 120° detector for protein crystallography; sensitive area is 1.5 × 20 cm. See color insert.

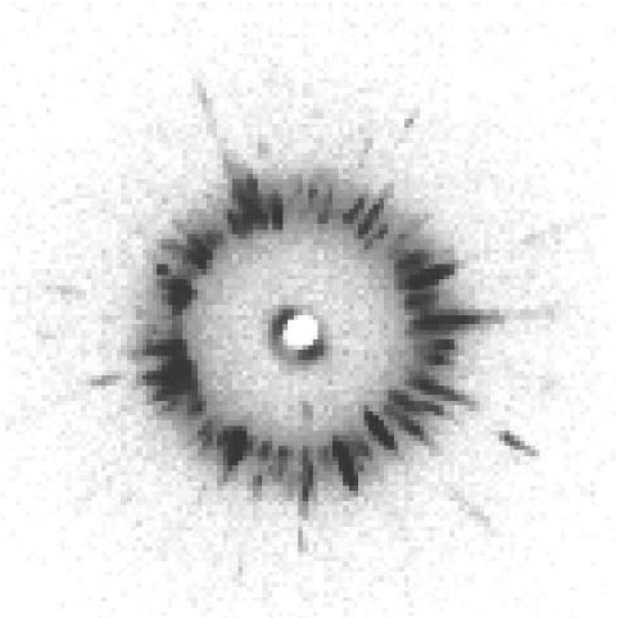


Figure 8. Small-angle scattering image, detailing many tiny Bragg peaks in a Debye–Scherrer ring.

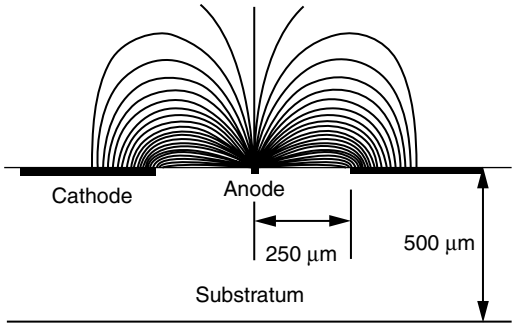


Figure 9. Cathode and anode strips in an MSGC. Field lines indicate an increased electric field at the anode surface.

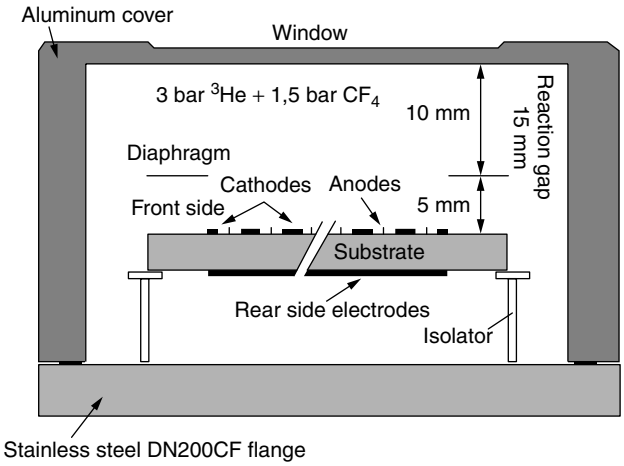


Figure 10. Schematic arrangement of the MSGC detector.

Detectors from ILL have used CF_4 as the stopping gas, and the positional resolution achievable is comparable

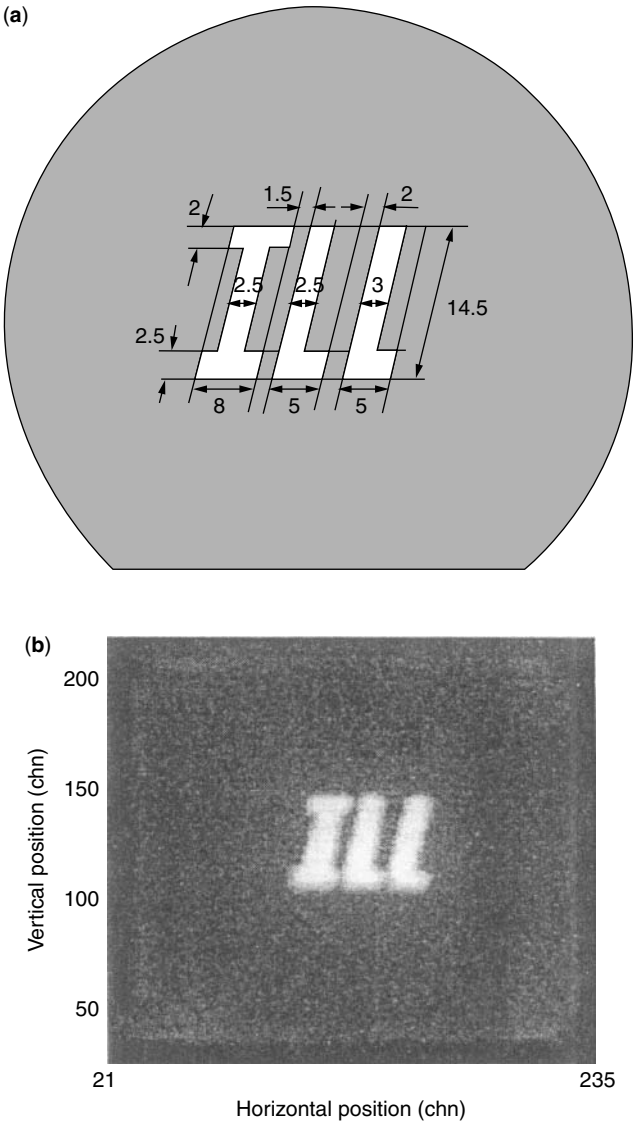


Figure 11. (**a**) Cadmium mask, dimensions in mm; (**b**) detector image of mask using thermal neutrons.

to that using C_3H_8. Figure 11 is the image of a cadmium mask.

The timing resolution of single wire counters, MWPCs, and MSGCs is typically of the order of 1 μs. This is primarily due to the uncertainty in the neutron conversion point in several millimeters of gas, combined with the slow neutron velocity.

MSGCs using Gd/CsI Converters

This principle is based on the original work of (22). The technique is being applied in a device that has a Gd foil in the center, a layer of CsI on each side of the foil, and a low-pressure MSGC on both sides to detect the secondary electrons emitted from the CsI (23,24). The main features of the device, illustrated in Fig. 12, are as follows. Neutrons convert in a thin Gd foil in the center of the device. Conversion electrons escape into a layer of CsI on both sides of the foil, and further secondary

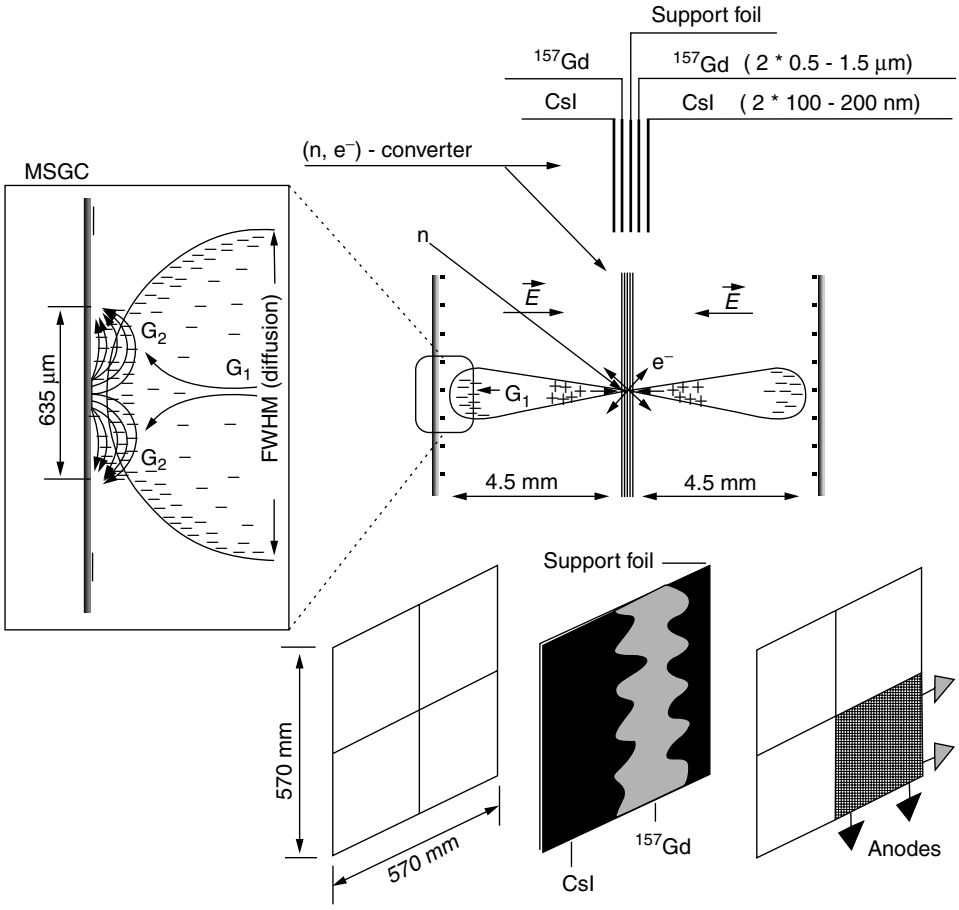


Figure 12. Principle of proposal for neutron imaging using Gd/CsI and low-pressure operation of MSGCs.

Table 2. Characteristics of Solid-State Scintillators

Material	Li Concentration (cm^{-3})	Photons/ Absorbed neutron	Decay Time (ns)	λ (nm)
^{6}Li-glass : Ce^{3+}	2×10^{22}	6,000–7,000	75	395
^{6}LiI : Eu^{2+}	1.8×10^{23}	51,000	1.4×10^3	470
^{6}LiF/ZnS : Ag^+	1.2×10^{22}	160,000	10^3	450
^{6}Li glass : Tb^{3+}	3×10^{22}	45,000	3×10^6	550
Gd_2SiO_5 : Ce^{3+}	–	800	56/600	430

electrons are emitted from this because a high electric field is present. These electrons drift across a low-pressure gas region to two-dimensional MSGCs on each side. This device has inherent resolution of less than 100 μm, high count rate capability, and excellent timing. No working devices have yet been reported. The timing resolution of this device is expected to be at least an order of magnitude smaller than the conventional gas detector because neutron conversion takes place in a very thin foil.

SCINTILLATOR DETECTORS

Solid-state scintillators are about three orders of magnitude higher in density than gases at atmospheric pressure and generally have much greater stopping power per unit volume. ^{6}Li based materials are by far the most common. Table 2 gives the characteristics of the main solid-state scintillators (25).

Anger Cameras

Originally conceived for X rays, the Anger camera (26) became an attractive option for neutrons through developments at Argonne National Laboratory (27). It consists of a single sheet of scintillator material, an intermediate layer to disperse the light, and an array of photomultiplier (PM) tubes (Fig. 13). Neutron conversion in the scintillator creates light that is seen by a group of neighboring PM tubes, and a center of gravity calculation is performed from the amplitude information from the respective tubes. The most commonly used scintillators

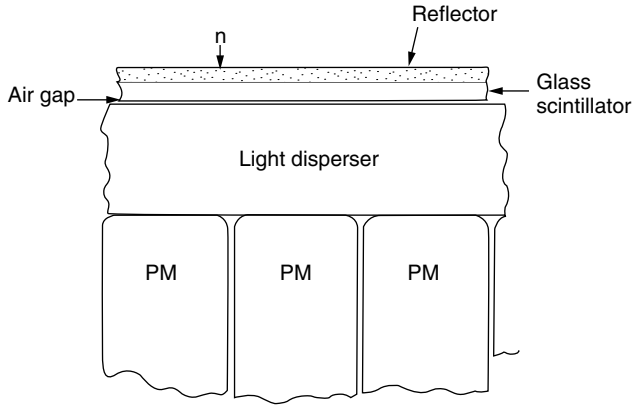


Figure 13. Schematic of the Anger camera.

are those based on ^{6}Li-glass : Ce^{3+}, available from vendors such as Bicron.

Kurz (28) describes further work on Anger cameras, and Schafer (29) describes linear and area detectors read out both as Anger cameras and using position-sensitive photmultipliers. These devices achieve relatively high counting rates and have good timing because of the thin layers. A major disadvantage is a relatively high gamma sensitivity.

A device using ^{6}Li glass and a microchannel plate will be described later.

Scintillators Using Fiber-Optic Coding

This technique was first demonstrated about 20 years ago (30). The principle is shown in Fig. 14. The light produced by neutron capture in the scintillator is divided among three fiber-optic channels and coupled uniquely to a combination of three PM tubes. The major advantage is a reduced number of PM tubes required, compared with the number of resolution elements. This technique has been further developed at the spallation facility ISIS, located at the Rutherford Appleton Laboratory. Initially used with linear detector arrays using $^{2}C_{n}$ coding, devices have now been developed using $^{4}C_{n}$ coding that yield two-dimensional information.

A recent example at ISIS of a two-dimensional device based on this coding principle (31) is a detector designed for a new diffractometer, where the requirement was for a square detector of dimensions 500 × 500 mm and a pixel size of the order of 10 × 10 mm. This detector was designed as four modules; each module has an active surface area of 348 × 144 mm.

A single module consists of a scintillator of ZnS/^{6}Li coupled by fiber optics to an array of photomultiplier tubes. The scintillator is located in front of an aluminized Mylar^{TM}, light guide grid that divides the module into 348 pixels, arranged on a 29 × 12 array, and pixel size is 12 × 12 mm. Figure 15 shows a single module from which the lighttight cover has been removed, and in which the scintillator has been removed (it is lying on the floor) to reveal the pixel array. Each pixel is viewed by sixteen

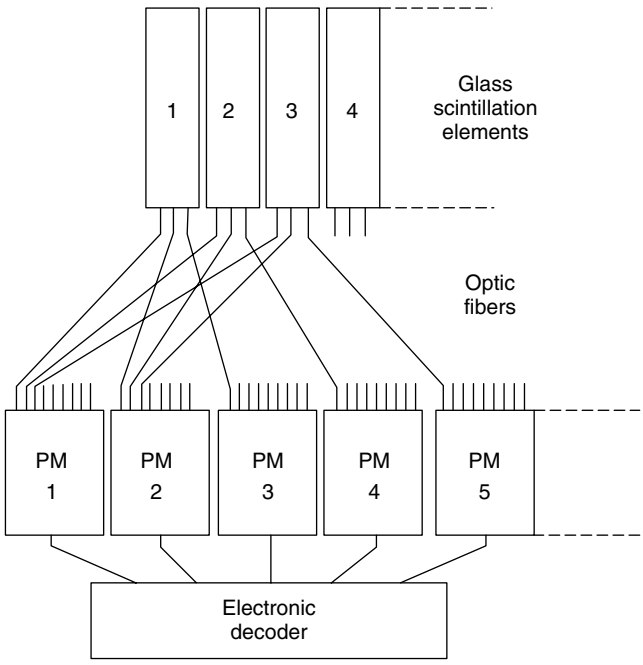


Figure 14. Fiber-optic coding of a discrete-element scintillator detector.

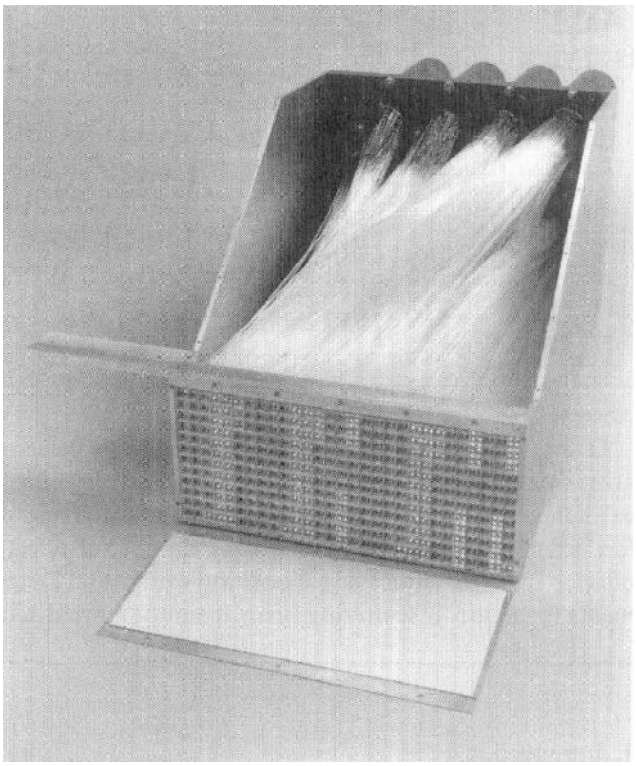


Figure 15. One module of the coded scintillator/fiber-optic device. Pixel array is 29 × 12. One PM has been removed, and its fibers illuminated (from rear). Beginning at the bottom of the right-hand column, the first pixel (lit) contains fibers from PMs 1, 2, 3 and 4; the next one up (dark) from PMs 1, 2, 3 and 5, all the way (all dark) to the ninth pixel up, PMs 1, 2, 3 and 12. The tenth pixel up (lit) contains fibers from PMs 1, 2, 4 and 5, the eleventh (lit) PMs 1, 2, 4, and 6; and so on. Thus, PM 4 is illuminated.

Table 3. Detector Characteristics of Two-Dimensional Fiber-Coded Scintillator Detector

Active surface	348×144 mm^2	# of photomultipliers	12
# of pixels	348	Diameter	51 mm
Pixel dimension	12×12 mm^2	Neutron efficiency at 1 Å	20%
# of fibers/module	5,568	Gamma sensitivity/^{60}Co	$<10^{-8}$
Fiber diameter	1.5 mm	Background cts (cm^{-2} h^{-1})	0.7
Fiber code	^4C_n	Pulse pair resolution	2.5 μs

1.5-mm diameter fiber-optic light guides. These fibers are optically bonded to PM tubes at the rear of the module, using a ^4C_n code. (Every pixel is actually a 2×2 array of square cells, each 6×6 mm^2. Centered on each cell are four fibers, and corresponding fibers from each cell join at the appropriate PM tube). For example, the first pixel is coupled to PMs 1, 2, 3 and 4; the second pixel is coupled to PMs 1, 2, 3, and 5; and so on. In this way, all 348 pixels of a module are coded by just 12 PMs. In Fig. 15, one photomultiplier tube has been removed, and its fiber optic bundle illuminated, thereby highlighting those fibers that feed that particular PM.

Outputs from the photomultiplier tubes are amplified and registered by discriminators, whose outputs are fed to a decoder. This EPROM generates the appropriate binary address for each unique set of four coincident inputs. After time stamping, a histogramming memory is updated. Table 3 shows the major characteristics of this device.

There is an interesting development from a new material based on Li/Gd borate (32). Initial measurements on small samples indicate a pulse height distribution for neutrons that is much more spectroscopic than existing scintillators and offers better n,γ discrimination. No imaging developments have yet been reported.

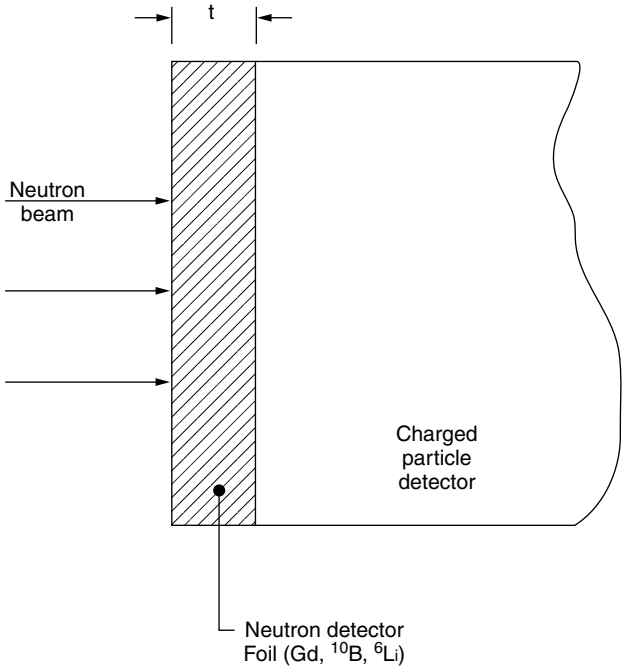


Figure 16. Principle of foil converters in transmission mode.

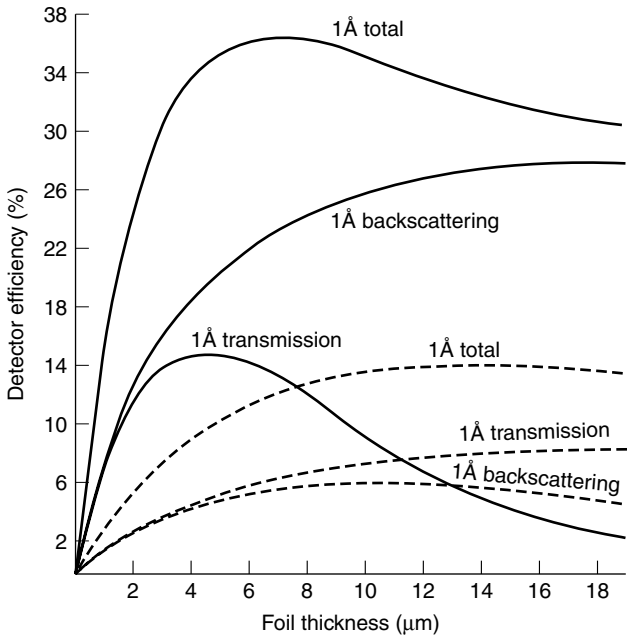


Figure 17. Detection efficiency as a function of foil thickness (solid line: ^{157}Gd; dashed line: ^{nat}Gd).

SEMICONDUCTOR DETECTORS

The use of semiconductors in neutron detection generally requires converting the neutron in a foil. The charged particle that may escape from the foil is sensed by the semiconductor. Foils can be used in transmission, as illustrated in Fig. 16, in backscatter mode, or both. However, this technique is a delicate and somewhat nature-dependent compromise: the foil should be sufficiently thick for high conversion efficiency, yet thin enough to allow secondary particles to escape.

Figure 17 illustrates the efficiencies of ^{157}Gd and ^{nat}Gd for 1 Å neutrons, as a function of foil thickness. Efficiency in this instance means that the probability of an escaping electron is unity. It should be noted that the foils are very thin and the efficiencies are not very high. (It is clear from this diagram why the Gd foil in the MSGC/Gd detector described previously is so thin.). These detectors generally have very good timing resolution because the secondary emission and its solid-state detection is extremely prompt.

Silicon Semiconductor Using ^{6}LiF converter

In this approach (33), a layer of LiF is attached to a reverse-biased Si diode. The reaction products are an

alpha, which has a range of only a few micrometers, and a triton, which has several times the alpha range. Even if only half of its 2.7 MeV is dissipated in the silicon, 3.8×10^5 electron/hole pairs are created, a sizeable signal. Such a LiF/Si device was made into a linear positional-sensitive detector by fabricating one of the two main electrodes as 24 strips of 1.25-mm spacing, and connecting them to an RC line. This application of charge division permits positional determination, and a measured spatial resolution of 1.25 mm has been obtained (33). The relatively thin converter layers (16 μm) yield less than 10% neutron absorption probability; further work is necessary to render this technique more practical.

Silicon/Gd Microstrip

A Si microstrip detector, a thin wafer of fully depleted silicon using readout strips on one face that permit positional determination, has been coupled to a Gd converter (34), as shown in Fig. 18.

In recent work on this type of instrument, a double-sided microstrip silicon sensor of area 20×20 mm^2 was used, as shown in Fig. 19. Signals from the 400-μm pitch strips are fed to low-noise, charge-sensitive preamplifier/shaping amplifier (VA) chips, the channel that has the largest signal is representative of the positional of a specific event. The neutron converter was a 250-μm thick natural Gd foil mounted in contact with the back-plane of the sensor. The less energetic line of the conversion electron spectra following neutron capture in Gd is observed at about 70 keV, creating about 2×10^4 electrons in the silicon. This is relatively easily detectable by the electronics.

Figure 20 shows the positional response, which was measured by scanning a beam of neutrons from the TRIGA reactor, Italy, across the detector. By selecting monochromatic neutrons, positional response as a function of energy was also determined, and a small dependence upon neutron energy is observed. It is likely that γ-ray background plays a limiting role in these measurements.

The timing resolution of semiconductor-based detectors should be of the order of a few nanoseconds.

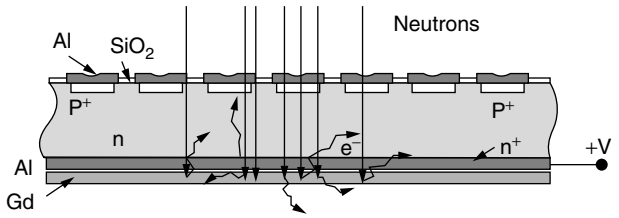


Figure 18. Principle of the Gd converter and silicon strip detector.

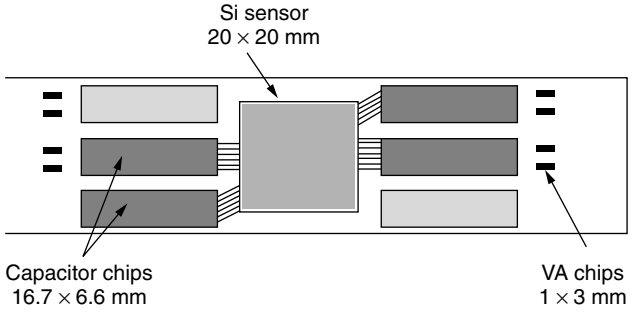


Figure 19. Silicon microstrip detector readout setup.

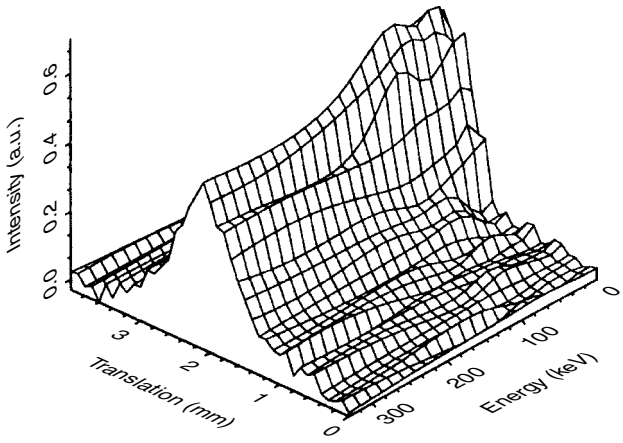


Figure 20. Positional response as a function of neutron energy.

FILM AND IMAGE PLATES

Film

Thermal neutrons can be imaged at high resolution by photographic film. Diffraction experiments using film were carried out many years ago (35), where indium was the neutron-absorbing medium. Now, the secondary foil is usually Gd. An alternative approach is to use a scintillator, such as ^{6}LiF/ZnS. A system has been described (36) in which film is sandwiched between two such screens, and a resolution of 100 μm was obtained. Although film consists of silver halide grains of the order of 5 μm, this intrinsic resolution is rarely achieved because the range of the electron from the foil is typically longer. Film, such as the image plate (next section), has the advantage of low cost, ability to cover large areas, flexibility, and good resolution. The disadvantages are no timing information, no sensitivity to single neutrons, and the inconvenience of a two-step readout process. Photographic film has a much smaller dynamic range, about 100 : 1, compared to image plates.

Image Plates

Image plates (IPs) were developed for diagnostic radiology during the early 1980s (37). They were soon used widely in X-ray protein crystallography (38). An IP is about 0.5 mm thick and is composed of a flexible plastic backing coated with fine storage phosphor crystals (BaFBr : Eu^{2+}) combined with an organic binder. The phosphor can store part of the energy released when ionizing radiation converts in the plate. When the plate is later stimulated by visible light, photostimulated luminescence (PSL) is emitted, whose intensity is proportional to the neutron dose. Essentially the IP is a form of electronic film, in which

incoming radiation creates an integrated latent image that is subsequently read out. The image is retrieved by illuminating the plate by laser light (HeNe); this produces PSL, or characteristic fluorescent radiation, from Eu^{2+}. The number of photons emitted is proportional to the converted incident radiation. A convenient method for the readout is a raster scan using a laser beam, simultaneously feeding the fluorescent radiation through a fiber-optic cable to a photomultiplier. IPs have a resolution less than 100 μm and a dynamic range much greater than that of film, typically $10^5 : 1$. These are impressive characteristics. Figure 21 illustrates the readout process for an IP.

Gadolinium is a good candidate to make an IP sensitive to neutrons. The strong (n, γ) resonance produces a wide range of keV γ rays and a cascade of conversion electrons. Both can stimulate the phosphor. Conversion can be done either by placing a thin plate containing gadolinium in front of the IP or by having the gadolinium compound inside the phosphor. It has been found that the second approach is better from the viewpoint of both resolution and signal-to-noise, (39,40). A key advantage of image plates is their mechanical flexibility. For example, they can be used to surround objects cylindrically as shown in Fig. 22. In this example, the image plate is mounted on a 30-cm drum, and the sample is inside, on-axis. Using a pixel size 0.25×0.25 mm^2, about 6 Mpixels were read out in 5 min. A diffraction pattern from a partially deuterated tetragonal lysosyme crystal taken by using this instrument is shown in Fig. 23 (39). Recently, it has been shown (41) that mixing ^{6}LiF into the image plate phosphor may significantly improve the neutron to γ sensitivity by an order of magnitude.

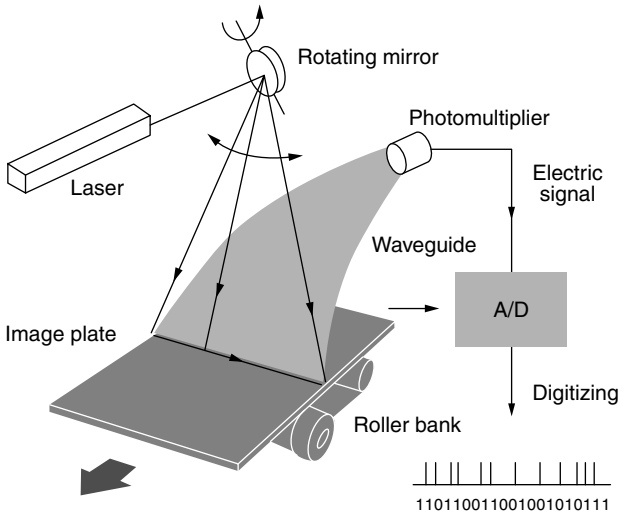


Figure 21. Signal readout for the image plate.

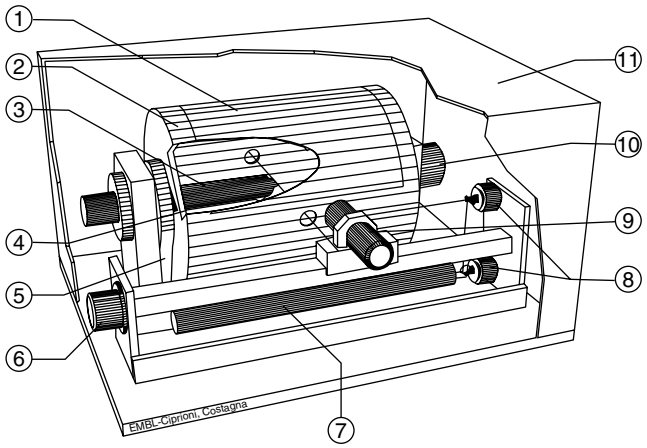


Figure 22. Arrangement for more rapid readout of an image plate. 1: Image Plate mounted on drum. 2: Drum. 3: Sample holder. 4: Crystal. 5: Belt to drive drum. 6: Carrier for read head with photomultiplier. 7: Laser. 8: Mirror to bring laser light to read head. 9: Read head with photomultiplier. 10: Encoder. 11: Cover.

MICROCHANNEL PLATES

Microchannel plates (MCPs) are vacuum imaging devices for UV/X-ray photons and charged particles (42). A limited amount of work has been carried out to determine their viability as direct detectors of neutrons. A first approach involved measurements using MCPs made from lead oxide glass containing lithium (43), where ionization from the emitted α particle is detected. Only a tiny efficiency for neutron detection was measured and, furthermore, it was realized that incorporating large mole fractions of lithium oxide into MCP lead glass was a severe challenge. A

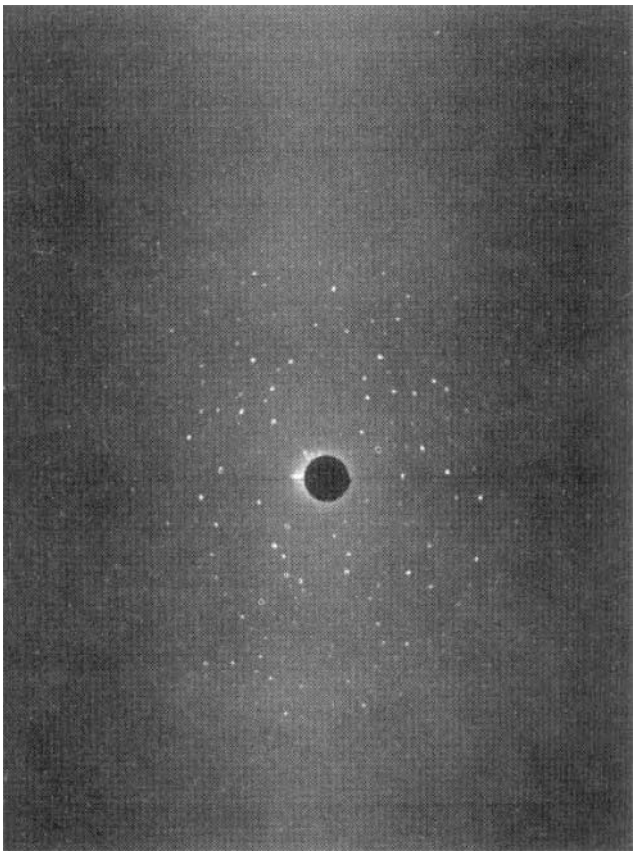


Figure 23. A diffraction pattern from a partially deuterated tetragonal lysosyme crystal.

more practical approach is to consider incorporating boron oxide into the MCP glass (3), but no results have been published so far. A recent development of the microchannel plate, called the microsphere detector, is described (44). The outer geometry and appearance of the two devices are essentially identical, but the newer device comprises irregularly packed, sintered glass beads in which an electron avalanche propagates in the gaps between the spheres. So far, however, there is no development of this plate specifically for neutrons.

A combination of scintillator, ^{6}Li-glass : Ce^{3+}, and MCP has been examined as a positional-sensitive detector (45), where the readout was by a two-dimensional resistive sheet. Figure 24 illustrates the principle. In a further development of this approach, a positional resolution of 40 μm was reported on a small device (46).

RADIOGRAPHY AND TOMOGRAPHY

Neutrons have unique scattering and absorption properties for studying structures and flaws in nondestructive testing of materials. X rays have a cross section for absorption that increases approximately as Z^4, and hence they are strongly absorbed by heavy elements, and little absorbed by light elements. Low energy neutrons, on the other hand, can interact strongly with light elements and minimally with heavy elements, and can have interaction properties that differ by several orders of magnitude between nuclei whose atomic numbers are close.

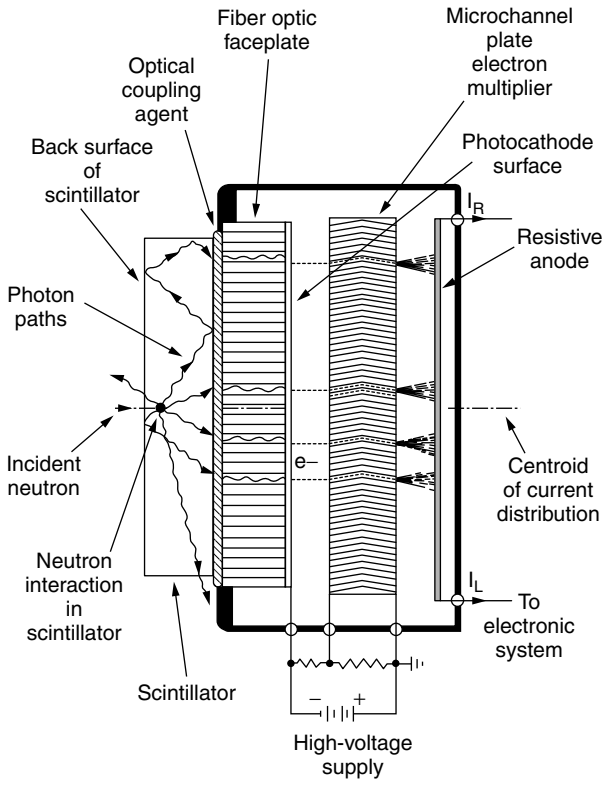


Figure 24. Combination of ^{6}Li glass, a microchannel plate, and a resistive sheet readout.

An excellent selection of papers relating to radiography and tomography using neutrons can be found in (6,7).

Planar neutron radiography was developed to take advantage of these unique neutron properties. It is ideally suited to studies of thin layers. In aluminum, for example, early corrosive changes that are rich in hydrogen-containing compounds can be detected because of the large interactive cross section with hydrogen nuclei. Neutron radiography, even though expensive compared with other nondestructive testing methods, is an indispensable part of inspection and examination in fields such as nuclear engineering, aerospace industry, pyrotechnics, the oil industry, and archeology. Specific examples are

- welding seams,
- cooling channels in turbine blades,
- gluing of sandwich structures used in aviation,
- homogeneity of new materials or enclosures inside them, and
- detection of hydrocarbonates inside metal or ceramic housings.

It has also been used in life science applications such as root activity in growing plants (47), insect research (48), and in the study of microorganisms (49).

However, the image in radiography is produced from the integrated interactive cross section across the entire path of radiation through the object. Local density or material changes can be hidden because of this essential averaging. Image reconstruction from a multitude of projections permits tomography to generate three-dimensional detail of the object's internal structure. Some early advantageous features of tomography were realized in test measurements on nuclear fuel (50), but its full potential has been thwarted by lack of efficient, high resolution (fraction of a mm) detectors.

Detectors for radiography have been largely based around image plates (51–54). However, image plates (and film) are not practical for tomography because a large number of projections need to be taken in a reasonable time and need to be available in digital form for further processing. Most tomographic work has been based on neutron scintillators coupled to CCDs (55–58); two are described here.

(a) McFarland et al. (55) at the Massachusetts Institute of Technology pioneered the use of a scintillator screen coupled to a CCD to demonstrate that such a system had the positional resolution and efficiency capable of yielding useful tomographs. Figure 25 shows a schematic of the experimental setup.

Neutrons that have passed through the test object convert in the screen, a ^{6}LiF–ZnS scintillator, whose light output is imaged onto a cooled CCD by a front-surfaced mirror and lens. This geometry allows shielding the CCD from the incident neutron beam, preventing radiation damage to the CCD and possible neutron activation. An efficiency of 15% for thermal neutrons and a resolution of about 100 μm for the screen was claimed by the manufacturer but this specific setup had significantly lower efficiency because the blue ZnS phosphor and the lens aperture were mismatched. However, acceptable images

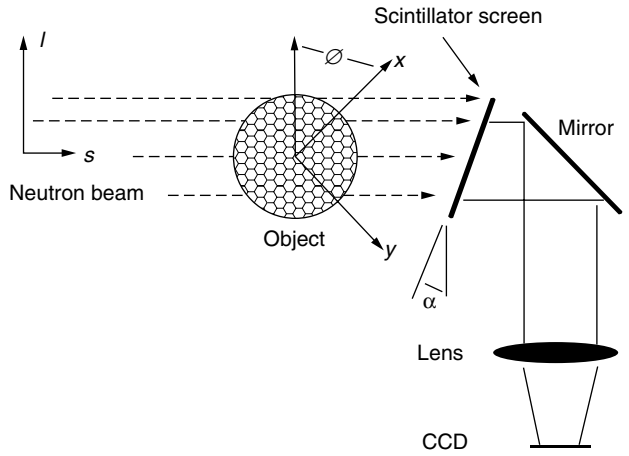


Figure 25. One of the first tomographic measurements using a CCD detector. Neutrons pass through the object and are converted in a scintillator screen mounted on the inside of a lighttight box. Optical photons reflected from the mirror are focused by the lens onto a CCD.

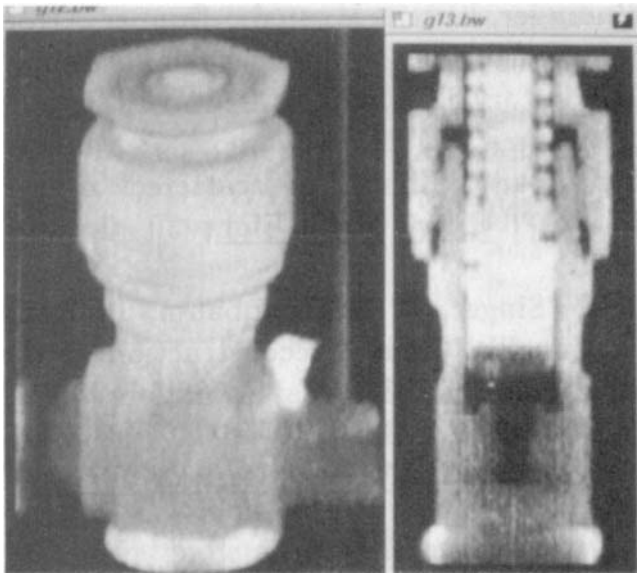


Figure 26. Three-dimensional view of a neutron tomographic reconstruction of a brass control valve.

were obtained. Figure 26 shows a neutron tomogram of a small brass control valve (about 3 cm outside dimension) reconstructed from 90 projections at 2° increments. The total imaging time was a little under 5 minutes, and the neutron fluence was 7×10^8 n/cm^2 per projection. Intrinsic resolution is close to the 100 μm of the scintillator.

(b) A group at the Technical University, Munich developed a high sensitivity CCD camera for detecting the faint light of a neutron scintillator (56,57). A prototype that had an active area of up to 28×28 cm^2 was developed. It is expected that after some further improvements, this detector will be sensitive enough to detect even a few single neutrons on an area as large as 0.5×0.5 m^2.

As in (a), a commercially available granular scintillator consisting of ZnS(Ag) +^{6}LiF was used. Thermal neutrons are absorbed at an efficiency of about 15%, and the ionizing reaction products produce approximately 1.8×10^5 photons in the ZnS(Ag) crystals at an average wavelength of 470 nm. The product particles, an α and triton, have average ranges of 10 μm and 26 μm, respectively, within the scintillator. The granular compositional of the scintillator limits the resolution to about 50 um.

The efficiency is improved for subthermal neutrons because the absorption cross section is inversely proportional to the neutron velocity. The scintillator material itself is somewhat opaque to its own radiation, so the scintillator layer must not be thicker than 0.5 mm for optimum light output; a thicker layer would be required for optimum neutron absorption. Light output for gamma rays is 450 to 4,500 times smaller than for one detected neutron.

Figure 27 shows the setup of the Munich instrument. The detector consists of a 28×28 cm^2 scintillator screen, an aluminum-coated surface mirror, and a Peltier cooled CCD camera system that has a 512×512 pixel Thomson CCD (TH7895 that has a 19×19 μm^2 pixel size). The driving electronics are very flexible and compact due to the use of highly integrated chips. Control of exposure, readout, and data transfer is performed by a standard PC. As in (a), the scintillator screen is viewed via the mirror by the CCD chip to prevent radiation damage to the latter. The mirror and the back side of the detector housing are transparent for neutrons, so that the nonabsorbed part of the beam leaves the detector. A zoom lens of $f = 1.0$ permits coverage of the whole active area of 28×28 cm^2.

Although the CCD is not in the direct line of neutrons, some problems arise from stray gamma rays that are generated by neutron capture in the beam catcher or in the sample. The light output of the scintillator is negligible

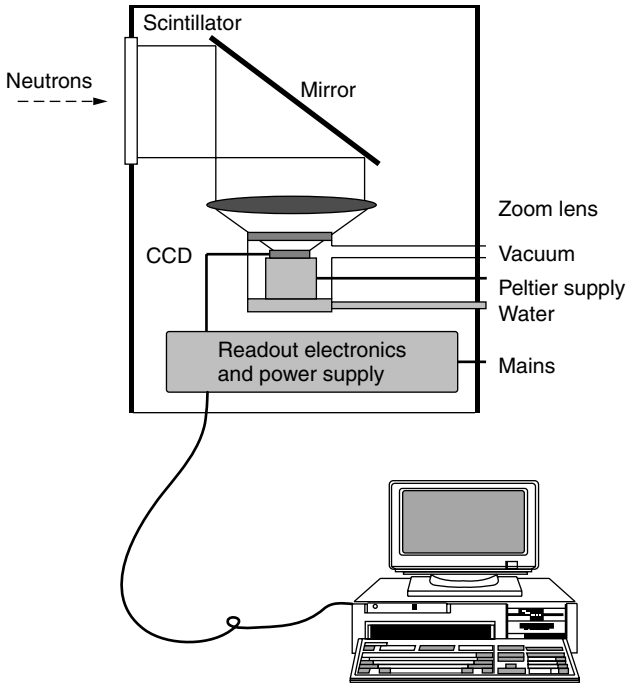


Figure 27. Schematic setup of the detector.

for gammas compared to neutrons, but the CCD chip itself is highly sensitive to gammas. A gamma event in a CCD generates several thousand photoelectrons and forms a nearly perfect point source. These gamma events appear as isolated bright spots in the picture and can be eliminated by electronic image processing by comparing two identical exposures. In the future, the CCD will be shielded with lead glass.

Neutron exposures have been taken at the Garching reactor FRM-I in a flux of the order of 3.0×10^5 n/s/cm^2 and an average neutron energy of 6.5 meV. Because the beam was only about 2.5 cm wide and 16 cm high, large-area pictures were obtained by mounting the detector and sample on a translation table, which was scanned across the beam. A compete radiograph was taken in 10 minutes, exposing each part of the sample to the beam for 2 minutes.

Figure 28 shows a radiograph of a car ignition distributor. On top is the low pressure box for ignition time adjustment that has a steel spring and membrane, and below this is the plastic housing of the capacitor, nearly dark because of its hydrogen content.

An example of the transformation from projection to tomograph is shown in Fig. 29. The object is a small electromotor about 20 mm in diameter. The CCD instrument was used to record many projections in 0.9° steps using a three-minute exposure each time. One such projection is shown in Fig. 29(a). From all projections, a three-dimensional reconstruction can be put together from a stack of two-dimensional reconstructions; the latter in turn is obtained from Fourier analysis of the projection raw data (59). The three-dimensional reconstruction was put together from 268 separately reconstructed layers of 201×201 pixels, forming a data cube of $201 \times 201 \times 268$ pixels (Fig. 29b). Each 3-D pixel corresponds to a cube whose side is about 120 μm.

Some upgrades of the detector system that will improve both throughput and image quality are a reduction of dark current noise by liquid nitrogen cooling, improvement of the sensitivity of the CCD for the 470-nm scintillation light by coating it with laser dyes, which act as wavelength shifters, or usage of backlit CCDs. Better shielding against gamma rays and better light collection by high-aperture lenses ($f = 0.75$) and fiber-optic tapers are also being planned.

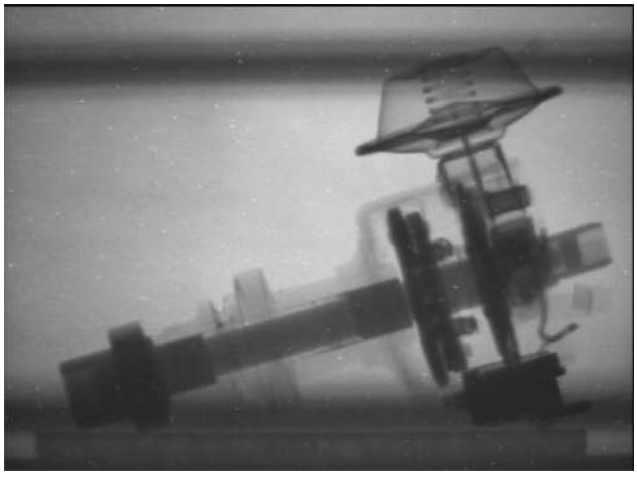


Figure 28. The raw data from a car ignition distributor.

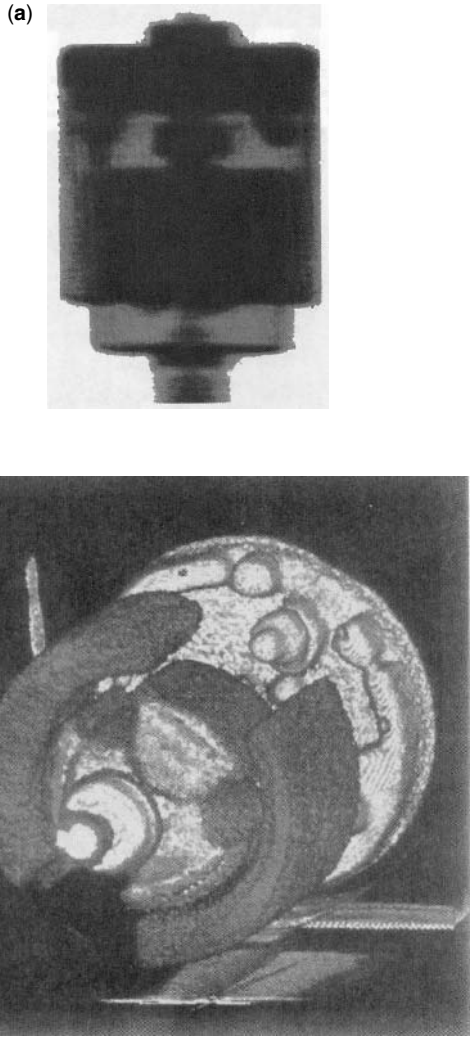


Figure 29. Neutron radiography/tomography of a small motor. (**a**) One single projection; (**b**) front view of motor (268 reconstruction layers).

Fast neutron radiography is an attractive nondestructive inspection technique because of the excellent penetrating characteristics of fast neutrons into matter. The difficulty of detecting fast neutrons, however, reduces its attractiveness. Some techniques have been developed to help overcome this difficulty. Two image plates and a sheet of polyethylene as a proton emitter were examined (60); the results showed that the method is applicable to fast neutron radiography and effectively discriminates γ rays.

CLOSING REMARKS

In this article, we described the principal reactions in nature that are useful for detecting of neutrons, highlighted the salient features of a range of detector technologies that have been developed for imaging

purposes, and outlined an important application, namely, radiography and tomography. However, in covering quite a range of disparate technologies, it is not possible to delve too deeply into some fundamental aspects of radiation detection. For example, in describing each imaging technology, little distinction was made between single neutron counting and the integrating mode. In fact, there are only three detector types described in this article that are truly integrating, namely, film, the image plate, and the combination of scintillator and CCD. These are most useful in very high rate applications where an image can be recorded quickly and where no time resolution is required; hence, they are commonly used in radiography and tomography. All other detector types described register neutrons electronically, one event at a time. This offers the possibility of time tagging events and, in experiments at spallation sources, each neutron energy can then be determined. Thermal neutrons, particularly, move very slowly; 1-meV and 100-meV neutrons traverse 10 m in 23 ms and 2.3 ms, respectively. Flight paths between source and detector can easily be 10 m, and therefore, time differences of several milliseconds can occur between the arrival of the least, and most, energetic neutron from one spallation pulse. Timing resolution, nonexistent for integrating detectors, has been indicated for counting detectors where appropriate.

Development of improved neutron imaging techniques will continue to be an area of challenging research. Detectors that have sub millimeter resolution, counting rates of megahertz per square centimeter, and area coverage of several square meters are essential if the world community is to make effective use of the newest spallation sources now under construction and consideration throughout the world.

Acknowledgments

The author acknowledges many informative discussions with his Brookhaven colleagues, particularly Drs. Veljko Radeka and Bo Yu, and many other colleagues at neutron facilities in the United State and abroad.

This work was supported by the U.S. Department of Energy Contract No. DE-AC02-98CH10886.

ABBREVIATIONS AND ACRONYMS

CCD	charge coupled device
EPROM	erasable programmable read only memory
ESS	European Spallation source
FWHM	full width at half maximum
ILL	Institute Laue Langevin
IP	image plate
MCP	micro channel plate
MSGC	micro strip gas chamber
MWPC	multi-wire proportional chamber
PM	photomultiplier
PSL	photostimulated luminescence
SNS	spallation neutron source

BIBLIOGRAPHY

1. P. B. Convert and J. B. Forsyth, eds., *Positional Sensitive Detection of Thermal Neutrons*, Academic Press, London, 1983.
2. R. K. Crawford, *Proc. SPIE* **1,737**, 210–223 (1992).
3. G. W. Fraser, *Proc. SPIE* **2,339**, 287–301 (1995).
4. G. Lander, ed., *J. Neutron Res.* **4**, 1–274 (1996).
5. J. Hastings, G. C. Smith, and B. Yu, eds., Workshop on Neutron Detectors for Spallation Sources. BNL, 1998, *http://neutronworkshop.bnl.gov/*.
6. C. O. Fischer, J. Stade, and W. Bock, eds., *Proc. 5th World Conf. Neutron Radiogr.*, Berlin, 1996, Pub: DGZfp.
7. E. Lehmann, H. Pleinert, and S. Körner, *Proc. 3rd Int. Topical Meet. Neutron Radiogr.*, Lucerne, Switzerland, 1999.
8. A. J. Peurrung, *Nucl. Instrum. Methods* **A443**, 400–415 (2000).
9. G. F. Knoll, *Radiation Detection and Measurement*, 3, John Wiley & Sons, Inc., New York, NY, 1999.
10. L. V. Groshev et al., *Bull. Acad. Sci. USSR Phys. Ser.* **26**, 1,127–1,146 (1962).
11. D. I. Garber and R. R. Kinsey, Report No. BNL 325, Brookhaven National Laboratory, 1976.
12. V. Radeka, N. A. Schaknowski, G. C. Smith, and B. Yu, in B. P. Schoenborn and R. B. Knott, eds., *Neutrons in Biology*, Plenum Press, NY, 1996, pp. 57–67.
13. G. Charpak, R. Bouclier, T. Bressani, J. Favier, and C. Zupancic, *Nucl. Instrum. Methods* **62**, 262–268 (1968).
14. V. Radeka and R. A. Boie, *Nucl. Instrum. Methods* **178**, 543–554 (1980).
15. R. A. Boie, J. Fischer, Y. Inagaki, F. C. Merritt, H. Okuno, and V. Radeka, *Nucl. Instrum. Methods* **200**, 533–545 (1982).
16. S. A. McElhaney and R. I. Vandermolen, Oak Ridge National Laboratory Report ORNL/TM-11557, 1990.
17. M. Imai, T. Kato, and D. Schneider, *J. Chem. Phys.* **106**, 9,362 (1997).
18. B. P. Schoenborn and R. B. Knott, eds., *Neutrons in Biology*, Plenum Press, NY, 1996.
19. V. Radeka, *IEEE Trans. Nucl. Sci.* **NS-21**, 51–64 (1974).
20. A. Oed, *Nucl. Instrum. Methods* **A263**, 351–359 (1988).
21. N. Vellettaz, J. E. Assaf, and A. Oed, *Nucl. Instrum. Methods* **A392**, 73–79 (1997).
22. V. Dangendorf, A. Demian, H. Friedrich, V. Wagner, A. Akkerman, A. Breskin, R. Chechik, and A. Gibrekhterman, *Nucl. Instrum. Methods* **A350**, 503–510 (1994).
23. B. Gebauer, C. Schulz and T. Wilpert, *Nucl. Instrum. Methods* **A392**, 68–72 (1997).
24. B. Gebauer, C. Schulz, T. Wilpert, and S. F. Biagi, *Nucl. Instrum. Methods* **A409**, 56–62 (1998).
25. M. J. Knitel, Ph.D. Thesis, Technical University, Delft, 1998.
26. H. O Anger, *Rev. Sci. Instrum.* **29**, 27–33 (1958).
27. M. G. Strauss, R. Brenner, F. J. Lynch, and C. B. Morgan, *IEEE Trans. Nucl. Sci.* **NS-28**, 800–806 (1981).
28. R. Kurz and J. Schelten, *Nucl. Instrum. Methods* **A273**, 273–282 (1988).
29. W. Schäfer, E. Jansen, G. Will, A. Szepesvary, R. Reinartz, and K. D. Müller, *Physica* **B213 & 214**, 972–974 (1995).
30. P. L. Davidson and H. Wroe, in P. B. Convert and J. B. Forsyth, eds., *Positional Sensitive Detection of Thermal Neutrons*, Academic Press, London, 1983, pp. 166–174.
31. N. J. Rhodes, A. G. Wardle, A. J. Boram, and M. W. Johnson, *Nucl. Instrum. Methods* **A392**, 315–318 (1997).
32. B. Czirr, G. M. MacGillivray, R. R. MacGillivray, and P. J. Seddon, *Nucl. Instrum. Methods* **A424**, 15–19 (2001).
33. J. Schelten, M. Balzhauser, F. Hongesberg, R. Engels, and R. Reinhartz, *Physica* **B234–236**, 1,084–1,086 (1997).

34. C. Petrillo, F. Sacchetti, G. Maehlum, and M. Mancinelli, *Nucl. Instrum. Methods* **A424**, 523–532 (1999).
35. E. O. Wollan, C. G. Shull, and M. C. Marney, *Phys. Rev.* **73**, 527–528 (1948).
36. H. Holwein, in P. B. Convert and J. B. Forsyth, eds., *Position Sensitive Detection of Thermal Neutrons*, Academic Press, London, 1983, pp. 379–390.
37. M. Sonoda, M. Takano, J. Miyahara, and H. Kato, *Radiology* **148**, 833 (1983).
38. Y. Amemiya, T. Matsushita, A. Nakagawa, Y. Satow, J. Miyahara, and J. Chikawa, *Nucl. Instrum. Meth.* **A266**, 645–653 (1988).
39. F. Cipriani, J. -C. Castagna, M. S. Lehmann, and C. Wilkinson, *Physica* **B213 & 214**, 975–977 (1995).
40. M. J. Knitel, V. R. Bom, P. Dorenbos, C. W. E. van Eijk, I. Berezovskaya, and V. Dotsenko, *Nucl. Instrum. Methods* **A449**, 578–594 (2000).
41. M. Thoms, *Nucl. Instrum. Methods* **A424**, 34–39 (1999).
42. J. L. Wiza, *Nucl. Instrum. Methods* **162**, 587–601 (1979).
43. G. W. Fraser and J. F. Pearson, *Nucl. Instrum. Methods* **A293**, 569–574 (1990).
44. A. S. Tremsin, J. F. Pearson, J. E. Lees, and G. W. Fraser, *Nucl. Instrum. Methods* **A368**, 719–730 (1996).
45. R. A. Schrack, *Nucl. Instrum. Methods* **222**, 499–601 (1984).
46. C. R. Rausch, Thesis, Technical University, Munich, 1996.
47. J. Furukawa, T. M. Nakanishi, and M. Matsubayashi, *Nucl. Instrum. Methods* **A424**, 116–121 (1999).
48. L. L. Allee, P. M. Davis, and H. C. Aderhold, *Proc. 5th World Conf. Neutron Radiogr.*, Berlin, 1996, pp. 728–733.
49. R. Wacha, V. R. Crispim, C. Lage, and J. D. R. Lopes, *Radiation Meas.* **32**, 159–162 (2000).
50. C. F. Barton, *Trans. Am. Nucl. Soc.* **28**, 212–213 (1977).
51. J. Stade, J. J. Rant, and M. Kaling, *Proc. 5th World Conf. Neutron Radiogr.*, Berlin, 1996, pp. 298–306.
52. S. Tazaki, K. Neriishi, K. Takahashi, M. Etoh, Y. Karasawa, S. Kumazawa, and N. Niimura, *Nucl. Instrum. Methods* **A424**, 20–25 (1999).
53. G. Bayon, *Nucl. Instrum. Methods* **A424**, 92–97 (1999).
54. S. Fujine, K. Yoneda, M. Kamata, and M. Etoh, *Nucl. Instrum. Methods* **A424**, 200–208 (1999).
55. E. W. McFarland, R. C. Lanza, and G. W. Poulos, *IEEE Trans. Nucl. Sci.* **NS-38**, 612–621 (1991).
56. B. Schillinger, *J. Neutron Res.* **4**, 57–63 (1996).
57. B. Schillinger, W. Blümluber, A. Fent, and M. Wegner, *Nucl. Instrum. Methods* **A424**, 58–65 (1999).
58. M. R. Gibbons, W. J. Richards, and K. Shields, *Nucl. Instrum. Methods* **A424**, 53–57 (1999).
59. A. C. Kak and M. Slaney, *Principles of Computerized Tomographic Imaging*, IEEE, NY, 1988.
60. M. Masahito, T. Hibiki, K. Mishima, K. Yoshii, and K. Okamoto, *Nucl. Instr. Methods* **A463**, 324–330 (2001).

O

OPTICAL IMAGE FORMATION

PANTAZIS MOUROULIS
California Institute of Technology
Pasadena, CA

INTRODUCTION

For the greatest part of human history, visible light has been the only image forming medium. And until recently, visual telescopes and microscopes comprised the vast majority of image forming instruments. The image formation principles examined in this article derive from such venerable ancestry but are broader in their application. Whereas visible wavelengths span a range of less than 2 : 1 in the electromagnetic (EM) spectrum, the contents of this article can find application in devices operating across a wavelength range of $\sim 10^6$, from nanometers to millimeters. However, the middle part of that range, from about 100 nm to about 10 μm, contains practically all of the devices that are classified as optical instruments, and those are our main concern here.

An image is an n-dimensional map of a property of an object. An optical image, as we understand it here, is a spatial map of one or more of the following three properties of an object: emissivity (E), reflectance (R), or transmittance (T). In other words, an object may emit EM radiation or it may reflect/transmit it in a spatially dependent way. Let A stand for any of E, R, and T, or a combination. Then, an optical image is a map or representation of $A(x, y, z)$. The function A also depends on wavelength, so one may plot a map of $A(\lambda)$. However, although the wavelength dependence is implicit in $A(x, y, z)$, its explicit mapping is not part of an optical image. Also, we restrict the concept of optical image to be independent of any recording device. Thus, an optical image is the image projected on the film by the camera lens, but not the image recorded by the film because the latter depends on the characteristics of the film.

In most cases, it is sufficient or necessary to simplify this definition of an optical image by restricting it to two dimensions. This is not to say that optical instruments may not image a three-dimensional object. It is really only a matter of convenience arising from the fact that most image recording devices are inherently two-dimensional, so an image formed on a plane is of primary practical interest. Accordingly, most of the theory of optical instruments deals with planar images, and therefore, either with a projection of the three-dimensional object on a plane or a slice of such an object.

The object property $A(x, y, z)$ is mapped by the optical system to a function $A'(x', y', z')$, which is the image. A 1 : 1 correspondence is ideally assumed to exist between the functions A and A', so that the amount of EM energy arriving at point (x', y', z') depends only on the amount of energy leaving the corresponding object point (x, y, z). Clearly this does not hold in practice because any optical system will image a point into a small area (or volume). This implies that the energy at any point O′ depends on the energy emitted from points in the neighborhood of O. Thus, the image is, in general, an imperfect map of the object, which means that some of the information contained in the object will be lost during imaging. Nevertheless, if the imperfections are not large, we may satisfy ourselves that a one-to-one correspondence may be established between small object and image patches.

There are two fundamental ways of deriving the function A'. Let us imagine for simplicity that our object is an illuminated plane such as a photograph that has some reflectance variation. We may scan the object, probing one small area at a time, or we may form simultaneously an image of the entire object in what is called whole-field imaging that may be accomplished by using a simple lens. Scanning may be accomplished in two fundamental ways: (1) by moving a small aperture in close proximity to the object and (2) by moving a spot of light across the object and collecting the reflected (or scattered) radiation. These three different ways of imaging share common principles and therefore can be described by a common mathematical formulation given later. However, the specific techniques and instruments involved in the various methods are different and are covered by separate articles in this Encyclopedia. Within this article, we concentrate explicitly on whole-field imaging, although all of the principles and most of the techniques are immediately transferable to scanning using a spot of light.

The simplest device for whole-field imaging is the pinhole camera. It comprises a square or rectangular box that has a small opening on one side. The image is formed on the side opposite the pinhole, as indicated in Fig. 1. The small size of the pinhole (relative to its distance from the object and image planes) restricts light from any object point to be incident on only a small image patch, thus establishing the approximate one-to-one correspondence mentioned previously. In other words, the light cone accepted from an object point is small, and so is the overlap between the bases of such light cones that arise from different object points. As the size of the

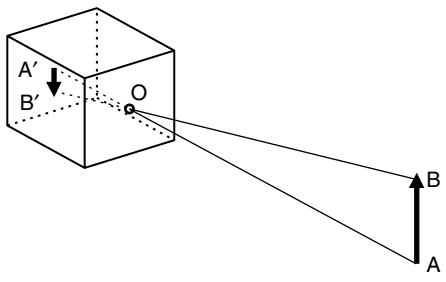


Figure 1. The pinhole camera.

pinhole increases, the bases of the light cones overlap significantly. Therefore, light from different object spots is incident on the same image spot, thus destroying the one-to-one correspondence and therefore the imaging ability of the pinhole camera.

Therefore, it would appear desirable for the pinhole to be as small as possible, but in fact this is not so. In the first place, as the pinhole gets smaller, the amount of light passing through it will diminish, resulting in a very faint image. But an even more fundamental limitation is caused by the wave nature of light. When a wave is made to pass through an opening of a size comparable to the wavelength, the wave spreads in directions that differ from the original direction of incidence due to *diffraction*. If the opening is sufficiently small, the wave spreads in all forward directions, across essentially a hemisphere. Thus, light from any point in the object would end up illuminating the entire image plane simultaneously, again destroying the correspondence between object and image.

The projection of the pinhole on the image plane, taken as approximately the same size as the pinhole itself, provides the *resolution limit* of the pinhole camera. Image details smaller than the pinhole size cannot be resolved. Any close examination of pinhole camera images reveals considerable blur. Thus, the pinhole camera has two fundamental problems: inferior resolution and a dim image.

Both of these problems are resolved by introducing a lens in place of the pinhole. As Fig. 2 shows, the lens collects light from an object point and concentrates (focuses) that light onto an image point. The amount of light collected by the lens depends on its diameter and can clearly be a lot more than the light collected by a pinhole. Interestingly, for an ideal lens, the ability of the lens to focus the light down to a very small area (focal spot) is inversely proportional to its size: the larger the lens, the smaller the focal spot size (assuming that the distance to the image is fixed). So we appear to gain both image brightness and image sharpness by increasing the lens size.

Unfortunately, for any nonideal lens, the focal spot size depends on the specific construction of the lens, as well as on the object and image size or the angle at which rays strike the lens. When these realistic considerations are taken into account, it is found that increasing the lens size beyond a certain point normally leads to an increase in spot size or requires a complicated lens construction.

The subject of this article is a more detailed appreciation of the contents and implications of the last two paragraphs. The necessary background for much of this material is that of a first-year college physics course. Specifically, some familiarity is assumed with the elementary concepts of a thin lens and its focal length/points and the basic concepts of wave optics. In addition to that, a proper understanding of the image formation concepts in a later section assumes some familiarity with the basic mathematics of Fourier transformations. Although some basic material of Fourier theory is reviewed in the Appendix, it is primarily useful as a refresher for those who already know something about the topic. However, a reader unfamiliar with such mathematics should still be able to appreciate that material because it is largely self-contained, although significantly compressed through omission of most mathematical proofs. A deliberate attempt has also been made to provide simple physical arguments for most equations, and many illustrations have been included. Clearly, for a detailed understanding, the reader must resort to the references given.

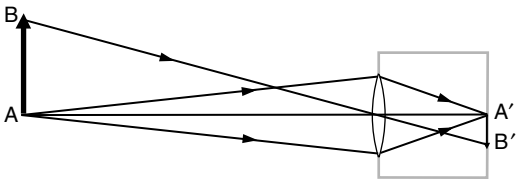


Figure 2. Schematic of a camera where lens is in place of the pinhole.

PARAXIAL OPTICS

A basic understanding of image formation and the associated instruments may be obtained through the concept of *rays*. A ray is a geometrical path that corresponds roughly to the direction of propagation of EM energy. This correspondence is generally good away from a focus and away from diffracting edges but breaks down dramatically in those two cases, where, as we will see, we have to employ the methods of wave optics. Inside isotropic materials, rays may be defined more strictly as lines perpendicular to the geometrical wave fronts, where the latter are defined as surfaces of constant phase measured relative to an originating point source. Space limitations prohibit a full discussion here; a more complete introductory discussion may be found in (1,2).

Along with the pinhole camera, a plane mirror may also be considered the simplest imaging device. As Fig. 3 shows, a plane mirror forms an image of a point at a distance (OA′) equal to −(AO). The image formed by a plane mirror is *virtual*. It is formed by a diverging bundle of rays that appear to have come from A′ even though they never get there. Combinations of plane mirrors allow us to perform interesting and useful image transformations such as image rotation or translation. Such combinations are often encountered as *reflecting prisms*, which are an essential element of many imaging systems, such as

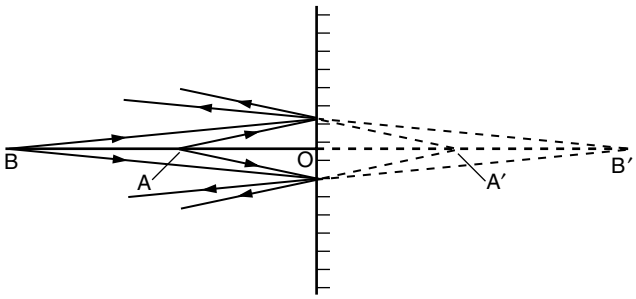


Figure 3. Imaging through a plane mirror.

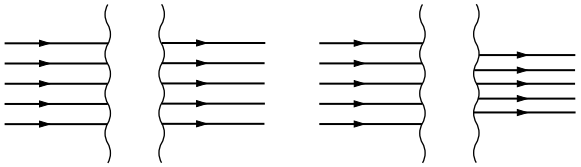


Figure 4. Schematic of afocal or zero-power systems. In both cases, the divergence angle of the incident bundle of rays is not affected.

cameras or binoculars. Details of the forms and properties of reflecting prisms may be found in (2–4).

A real image is formed by the lens shown in Fig. 2. This image is formed by a converging bundle of rays. One important difference between the plane mirror and the lens is that the plane mirror cannot alter the divergence (or convergence) of the rays, whereas the lens (or a curved mirror) can. For example, the lens of Fig. 2 transforms a diverging bundle, emanating from point A, into a converging bundle focusing on point A′. When an optical system changes the divergence or convergence angle of the rays, we say that it has *power*. More strictly, we may define a zero-power (or *afocal*) system from Fig. 4: If an incident parallel bundle of rays remains parallel after passing through (or interacting with) the system, then we say that the system has zero power, or is afocal. Systems that have nonzero power may also be called *focal*.

It should be understood that there is no fundamental difference between an afocal and a focal system. A plane mirror is the simplest afocal system. As Fig. 3 shows, although a plane mirror does not alter the divergence of the rays, nevertheless, it forms an image when a nearby object is presented. As the object recedes, so does the image. At the limit, when the object is at infinity, rays emitted from an object point will be parallel, and will remain parallel after impinging on the plane mirror. This means that they will intersect at infinity, which is also where the image will be located. In other words, parallel rays are focused at infinity. A lot of confusion can be avoided if one thinks of an object (or image) at infinity as the limiting case of a very distant object, and similarly, of an afocal system as the limiting case of a focal system.

We will restrict the exposition of image forming optical systems to systems that have an axis of rotational symmetry and contain only homogeneous materials, such as a compound lens system in which all of individual lens elements have the same axis. Practically all of the critical concepts of image formation can be developed within this framework, and once those are understood, the reader can extend them to other situations whose symmetry is different.

A first and important approximation to the behavior of rotationally symmetrical systems is obtained by considering only rays that remain infinitesimally close to the axis. Such rays are called *paraxial*. In practical terms, the paraxial domain is defined by neglecting terms of second or higher order in the power series expansion of the sine or cosine. Thus, if a is the angle that a ray forms with the axis,

$$\sin a = a + (\text{terms of third or higher power in } a) \qquad (1)$$

and

$$\cos a = 1 + (\text{terms of second or higher power in } a). \qquad (2)$$

Thus the paraxial approximation is $\sin a = a (= \tan a)$, and $\cos a = 1$. An important consequence of the paraxial approximation is that all smooth curved surfaces of revolution about the axis become indistinguishable from a sphere. When the origin of coordinates is at the apex, the equation describing such surfaces may be written as

$$z = \tfrac{1}{2}c(x^2 + y^2) + (\text{higher order terms}), \qquad (3)$$

where z is the coordinate along the axis and $c = 1/R$, where R is the local radius of curvature. The bracketed term in Eq. (3) can be neglected to within the same level of approximation as the corresponding terms in Eqs. (1) and (2). As a result of this restriction, paraxial image formation is perfect. This means that all wave fronts are spherical, and therefore, all rays meet at a single point, the center of the spherical wave front. Thus the image of point A in Fig. 2 is also a point. Similarly, the image B′ of the extra-axial point B is also a perfect point, but it is assumed that the lengths (AB) and (A′B′) are infinitesimally small. Although this assumption appears at first sight extremely restrictive, the value of paraxial optics is not restricted to an infinitesimal region around the axis. Appropriate extension to a finite region can be successfully made, as shown later. The departure of the wave front from sphericity is described as *aberration*.

Paraxial optics theory is used in performing the following four functions: (1) finding the location of the image, (2) finding the size of the image, (3) obtaining a first approximation to the brightness of the image, and (4) determining a 'first-cut' optical system that satisfies the previous three requirements. The fourth problem is generally considerably harder than the first three. That is the field of *optical design*, treated in (2–7), among several others. Within this article, we concentrate on the problem of finding the properties of a given system, without concerning ourselves how the system was designed to achieve those properties.

IMAGE LOCATION AND SIZE

Given an optical system and an object, how do we determine the image location and size? In this section, we show that both of these questions may be answered by tracing *a single paraxial ray*.

An optical system may comprise any number of lenses or mirrors, keeping in mind that we have restricted our attention to systems of rotational symmetry. It is evident, however, that plane mirrors, used to fold an otherwise long system, do not destroy this rotational symmetry and are ignored from our present viewpoint. Thus, we represent an arbitrary optical system as a collection of n refracting and/or reflecting surfaces (Fig. 5).

Paraxial Ray Tracing

Paraxial ray tracing involves two operations: refraction through a surface and transfer to the next surface. In

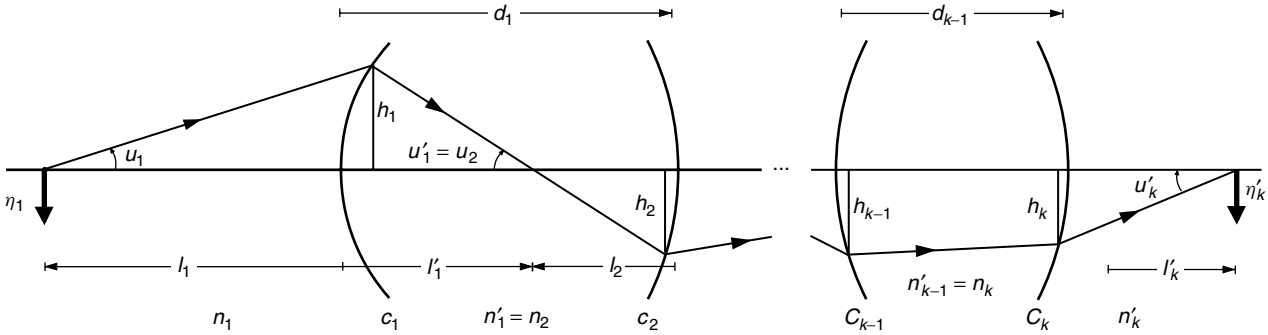


Figure 5. Schematic of an optical system comprising k optical surfaces that have the same axis of rotational symmetry. In the paraxial approximation, the height h and the angle u are both considered infinitesimal.

Table 1. Notation for Paraxial Ray Tracing

	Symbol	Meaning
Constructional parameters	n	Refractive index
	c	Surface curvature ($1/R$)
	d	Axial surface separation
Ray parameters	h	Paraxial ray/surface intersection height variable
	u	Paraxial ray angle variable
	l	Object/image distance from surface
	η	Paraxial object/image height variable

the first operation, we determine the new ray angle, and in the second its point of intersection with the next surface. Table 1 gives the meaning of the symbols in Fig. 5. A subscript is used to associate a variable with its corresponding surface; primed quantities refer to the *image space* of a surface, that is, after refraction; unprimed quantities refer to the *object space* of a surface, that is, before refraction. In this sense, "refraction" also includes reflection, if the surface in question is a mirror. The same set of equations can handle both cases by using an appropriate convention, as shown.

The following two equations describe the paraxial refraction (Eq. 4) and transfer (Eq. 5) processes:

$$n'_i u'_i - n_i u_i = -h_i c_i (n'_i - n_i), \tag{4}$$

$$h_{i+1} = h_i + d_i u'_i, \tag{5}$$

where $n'_i \equiv n_{i+1}$ and $u'_i \equiv u_{i+1}$.

The paraxial angle and height variables are related to the object distance through

$$u = -\frac{h}{l}, \tag{6}$$

and a similar equation (with primes) may be written for the image. Thus if paraxial ray tracing succeeds in establishing the height and angle variables at the last surface (after refraction), then, Eq. (6) may be used to determine the image location. It is also possible to remove the paraxial angle variable u from Eqs. (4), (5) in favor of the object/image distance l. It should be clear that the quantity l is normally finite, whereas both u and h are infinitesimal (although obviously they may not be represented pictorially as such).

In applying Eqs. (4)–(6), attention must be paid to the sign convention if one is to obtain correct numerical answers. The sign convention used is that of a Cartesian coordinate system where the local origin is at the apex of the surface. Thus, distances to the left and the bottom of the apex are negative, and acute angles measured from the axis to the ray are positive if counterclockwise. Light is assumed to come from the left, unless a mirror is involved.

Mirrors introduce an additional consideration to the sign convention. Reflection from a mirror may be treated as refraction if the following two conditions are observed: (1) after reflection, the refractive index of the "new" medium changes sign (so that the refractive index of air becomes -1 after the first reflection, and 1 after the second), and (2) the sign of the axial separation d also reverses. The second condition is in fact obvious because in going from surface i to surface $i+1$, we would have to move from right to left after an odd number of reflections.

Determination of Image Location

Sufficient background has already been provided to determine the image location for an arbitrary system and object. Clearly, the system must be fully specified in terms of refractive indexes, curvatures, and surface separations. The process is as follows:

1. Derive the initial values (at the first surface) for the paraxial height and angle variables u_1, h_1. If the object distance (l_1) from the first surface is the only parameter known, then one specifies an arbitrary value for u_1 and finds h_1 through Eq. (6). An arbitrary value for u may be used because all paraxial rays from a point will be imaged at the same one point (all wave fronts are spherical). This may also be seen by the fact that the variable u drops out of Eqs. (4), (5) if u is substituted in terms of l [shown by Eq. (6)].
2. Apply Eqs. (4), (5) to determine u'_k (after the last refraction) and h_k (at the last surface).
3. Again apply Eq. (6) to determine l'_k.

Clearly, this procedure may be reversed if we desire to determine the object location from a known image location.

Determination of Image Size (Magnification)

It would appear at first sight that to find the image height, we would need to trace a ray from the top of the object and determine its intersection at the image plane. Although this is possible, it is strictly unnecessary. Figure 5 contains all of the information necessary for determining the image size, and it is important to understand why. First, let us define the magnification M as the ratio of the image and object heights:

$$M = \frac{\eta_k'}{\eta_1}. \tag{7}$$

This type of magnification is called *transverse* magnification, to distinguish it from other possible types such as longitudinal, angular, and so on, described in the next section. When the adjective is omitted, transverse is normally implied.

Now, consider imaging by a single surface, as in Fig. 6. A ray is shown from the top of the object through the apex of the surface that intersects the image plane at the top of the image. Application of the paraxial form of Snell's law for this ray ($n \sin i = n' \sin i'$) and simple geometry gives us the following relationship:

$$M = \frac{\eta'}{\eta} = \frac{nl'}{n'l} = \frac{nu}{n'u'}, \tag{8}$$

where the last equality follows from $u = -h/l$. Successive application of this formula for the k surfaces of Fig. 5 gives

$$\begin{aligned} M = \frac{\eta_k'}{\eta_1} &= \frac{\eta_1'}{\eta_1}\frac{\eta_2'}{\eta_2}\cdots\frac{\eta_k'}{\eta_k} = \frac{n_1}{n_k'}\frac{l_1'}{l_1}\frac{l_2'}{l_2}\cdots\frac{l_k'}{l_k} \\ &= \frac{n_1u_1}{n_1'u_1'}\frac{n_2u_2}{n_2'u_2'}\cdots\frac{n_ku_k}{n_k'u_k'} \end{aligned} \tag{9}$$

because $\eta_j' = \eta_{j+1}$ and $n_j' = n_{j+1}$. Also, because $u_j' = u_{j+1}$, the last equality simplifies to

$$M = \frac{n_1u_1}{n_k'u_k'}, \tag{10}$$

showing the tremendous simplification that is achieved by using the paraxial angle variable u. The corresponding magnification formula utilizing the length l in Eq. (9) does not simplify further and can also cause additional computational problems if an intermediate distance is either infinite or zero. This is one of the key reasons for using the paraxial angle variable in the ray-tracing Eqs. (4), (5). Notice now that all of the quantities in Eq. (10) are associated with the single ray from the axial object point shown in Fig. 5. Thus, we have demonstrated that both the image location and the magnification can be determined by tracing a single paraxial ray from the axial object point.

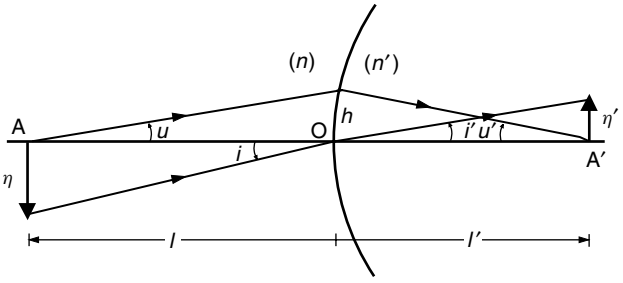


Figure 6. Imagery of an extended object by a single surface.

Other Types of Magnification

Longitudinal Magnification, Image of a Volume. If the object is three-dimensional or if it is anything other than a plane perpendicular to the axis, the previous concept of linear or transverse magnification needs to be extended. We can do so by introducing the concept of longitudinal magnification. Consider two points A_1 and A_2 along the axis of an optical system, separated by a distance δl. Let their images be A_1' and A_2', also along the axis, separated by a distance $\delta l'$. M_1 is the linear magnification between points A_1 and A_1', and similarly for M_2. It can be shown that the longitudinal magnification is the ratio of the two axial segments (2):

$$M_{\text{long}} \equiv \frac{\delta l'}{\delta l} = M_1M_2. \tag{11}$$

Now consider a cubic volume. The face of the cube centered at K will be imaged at magnification M_1, and the face centered at L will be imaged at a magnification M_2. This shows that the images of the two faces will be of different size and implies that the image of a cubic volume is a truncated pyramid, as shown in Fig. 7. This holds only if the focal points of the lens are outside the segment (KL), so that the image of the cube is wholly real or wholly virtual.

Afocal systems have the property that the linear magnification is constant, independent of the object location. For such systems, the image of a cubic volume becomes a parallelepiped.

Angular Magnification. When object and image are at infinity, the concept of transverse or linear magnification is not useful and is replaced by the concept of angular magnification. An optical system that forms an image at infinity of an object at infinity must necessarily be afocal because a parallel bundle of rays remains parallel. This is illustrated in Fig. 8, where the two lenses are separated by the sum of their focal lengths. A parallel bundle of rays from the edge of the infinite object is shown that forms an angle u with the axis. The bundle emerges parallel at an angle u' and is directed to the edge of the image. The angular magnification is

$$M_{\text{ang}} = \frac{u'}{u}. \tag{12}$$

As mentioned in the previous section, the linear magnification is constant for an afocal system. Let the object not be at infinity but at an arbitrary point A along the axis. If the system is afocal, any ray parallel to the axis in the object space will emerge parallel in the image space.

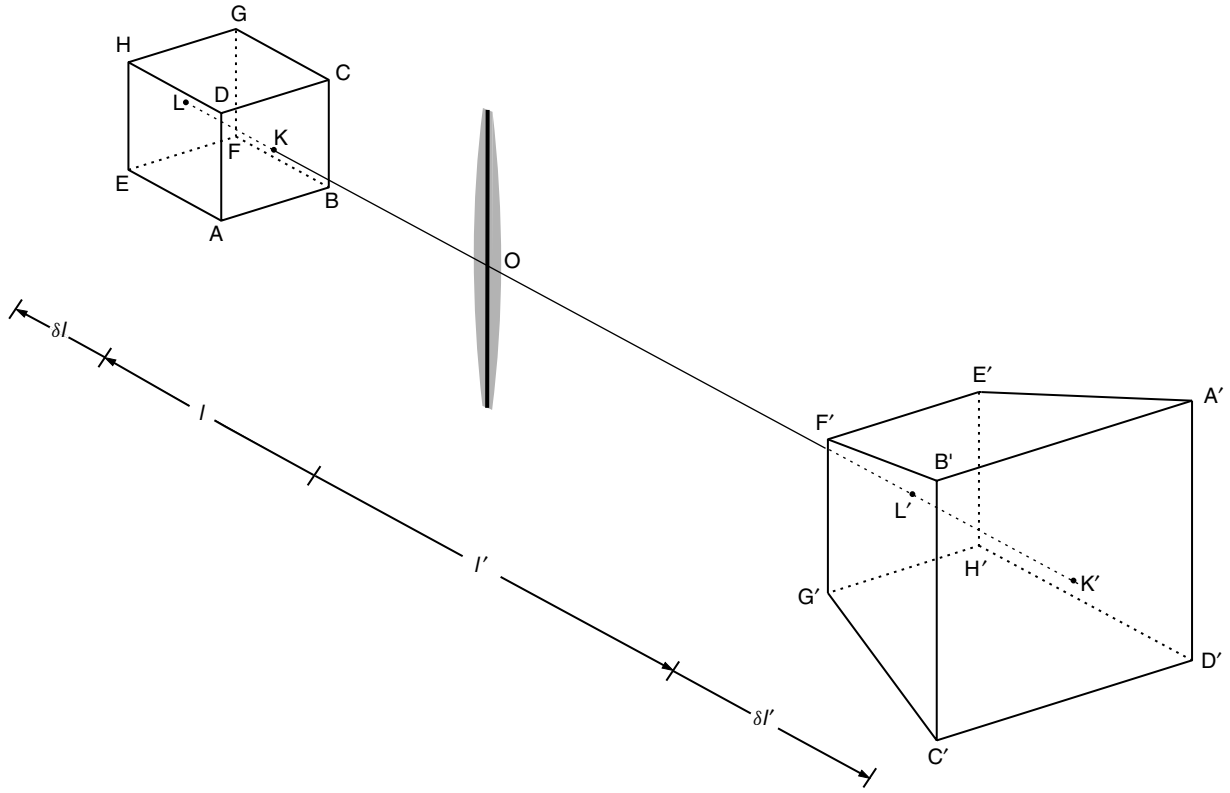


Figure 7. Image of a cube through an ideal focal system.

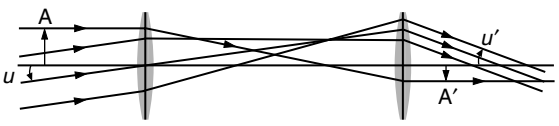


Figure 8. Imaging by an afocal system. The two lenses shown are assumed to be "thin." A thin lens is an idealization of a real lens of infinitesimally small thickness. Therefore, all refraction is assumed to take place on a single plane, even though two distinct surfaces of different curvatures may be associated with the lens.

It follows that such a ray will pass through the top of the image, independent of the image location. Hence, the linear image size is constant. It may be shown that the linear magnification of an afocal system is the inverse of the angular:

$$M_{\text{lin}} = \frac{1}{M_{\text{ang}}} \tag{13}$$

For a focal system, either the object or the image may be at infinity but not both together. In either case, the concept of linear magnification also breaks down. As Fig. 9 shows, an object point B at the focal plane of a lens gives rise to a parallel bundle of rays emerging from the lens (image at infinity). The magnification may then be thought of as the linear object size divided by the angular image size. From the figure, we see that this ratio is the focal length F, which therefore represents the system magnification in such a case. By reversing the direction of the rays, it is seen that a corresponding relationship holds when the object is at infinity.

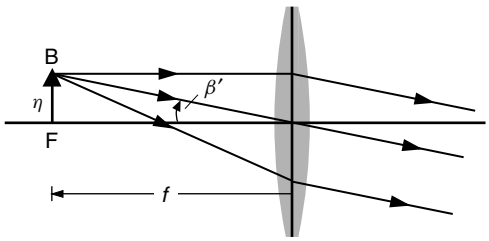


Figure 9. Imaging by an ideal focal lens where the object is at the focal plane. See Fig. 10 for an explanation of all symbols.

Visual Magnification. When a magnifying glass is used for visual observations, the important parameter is the gain in apparent image size between the aided and unaided eye, not the linear or the angular magnification. We set a (semiarbitrary) *near point* for clear vision at 25 cm away from the eye. When the object is at that distance, the maximum amount of detail may be resolved with naked eye. Let the object subtend an angle β in that case. When the object is viewed through a magnifier of focal length F, the magnified image subtends an angle β'. Visual magnification is the ratio of the two angles, and it can be shown that it varies between the limits (2)

$$\frac{25\text{ cm}}{F} \leq M_{\text{vis}} \equiv \frac{\beta'}{\beta} \leq \frac{25\text{ cm}}{F} + 1, \tag{14}$$

depending on the object location.

LENS (MIRROR) ASSEMBLIES: GAUSSIAN OPTICS

It is assumed that the reader is familiar with the following basic thin-lens conjugate equation:

$$\frac{1}{l'} - \frac{1}{l} = \frac{1}{F} \tag{15}$$

In this equation, l and l' are the object and image distances from the lens, and F is the *equivalent* or *effective* focal length (EFL). The lens is assumed to be in air. The signs are consistent with the Cartesian sign convention, established previously, which is almost universally followed in all recent advanced texts and almost never followed in early elementary texts. The quantity on the right-hand side is the *power* K of the lens given by

$$K \equiv \frac{1}{F} = (n-1)(c_1 - c_2). \tag{16}$$

The power of the lens is its fundamental characteristic and suffices to describe all of its imaging properties that we have discussed so far. The thin-lens equation is depicted in Fig. 10. In the thin-lens approximation, the axial distance between the two surfaces is assumed to be zero, so all refraction takes place along a single plane, from which the distances l and l' are measured. This is the reason for drawing the thin lens as a single line that has shaded surfaces. When the thickness of the lens cannot be ignored, it may be shown that the power is given by

$$\frac{1}{F} = K = (n-1)(c_1 - c_2) + dc_1c_2\frac{(n-1)^2}{n}, \tag{17}$$

where d is the axial lens thickness. Using this definition, a conjugate equation similar to Eq. (15) may be written for the thick lens, but only if the distances l and l' are measured from the proper locations. The way this may be accomplished and the way of finding the proper locations for measuring l and l' is explained later in the more general context of an arbitrary optical system.

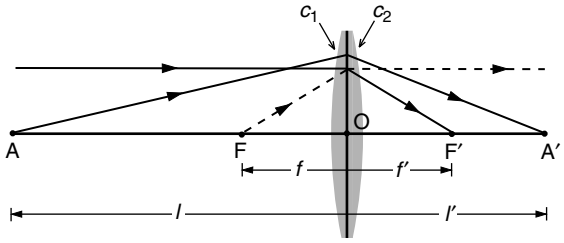


Figure 10. Depiction of the thin-lens conjugate equation. The object, point A, is imaged at A′. The distances l and l' are measured from the plane at which refraction is assumed to take place. Two more rays are also shown: one from an axial object at infinity that crosses the axis at the second focal point F′, and one that emerges parallel to the axis, having originated at the first focal point F. The distances (OF) = f and (OF′) = f' are the first and second focal lengths of the lens and are opposite in sign ($f = -f'$), irrespective of whether the lens is positive or negative. They are not to be confused with the equivalent focal length (EFL), F, whose sign is characteristic of the lens. In the simple situation depicted in this figure, the EFL happens to coincide numerically with f', but this is not generally the case. For example, if the lens is immersed in a medium of refractive index n', then $f' = n'F$. The general case is treated later in this section.

The paraxial ray-tracing techniques previously described suffice to determine the image location and size, if that is the sole concern. However, they are useful only if a detailed prescription of the system is available. But it is often necessary to understand the imaging action of a lens assembly whose internal construction is unknown. One such simple example is a compound camera lens comprising several unknown lens elements. The methods of Gaussian optics allow us to predict the behavior of such a lens in an abstract way. In effect, any focal lens/mirror assembly can be reduced to a thin-lens equivalent. Clearly, the reduction cannot be perfect if we are to retain some essential information about the compound lens. Whereas a thin lens is characterized by a single parameter (its power or EFL), a compound or thick lens requires three parameters; but even so, an arbitrary system is enormously simplified when reduced to a set of only three parameters.

The three parameters in question are the power or equivalent focal length of the entire system and the location of two reference planes from which object and image distances are to be measured. These are interrelated as follows: Consider an arbitrary system, for example, the compound camera lens of unknown internal construction mentioned earlier. The first and last surfaces of the lens are the only ones physically accessible, and we represent those in Fig. 11. Now, consider a ray from the axial object at infinity that crosses the axis at the second focal point F′ (if the system is not afocal). Inside the system, the ray is refracted in an unknown fashion, but we may determine the point Q′ through the intersection of the initial and final ray segments. The plane P′Q′ is then the plane perpendicular to the axis that passes through Q′. By proceeding in the opposite direction, we establish the plane PQ. These two planes are called *principal planes*, and they have the interesting property of unity magnification [so that the image of the segment (PQ) is the segment (P′Q′)]. The principal planes are the reference planes we need to apply the thin-lens equation to an arbitrary optical system.

This is shown in Fig. 12. Let the image of (AB) be (A′B′), and assume that we have established the location of the principal planes for the system (P, P′), following

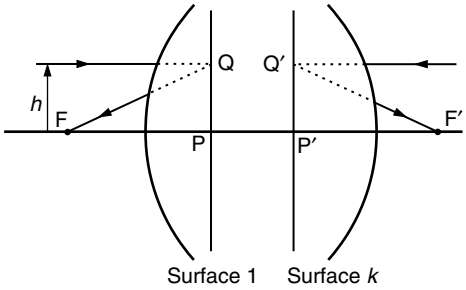


Figure 11. Establishing the principal planes of a focal lens system.

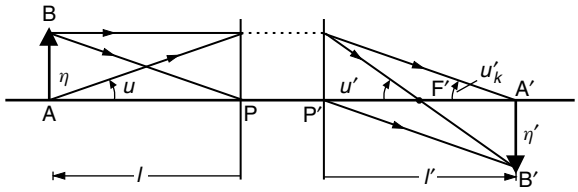


Figure 12. Imaging by a general focal system represented by its principal planes.

the previous procedure. Then, we may write an equation identical to Eq. (15) for the object and image distances, if they are measured from P and P′, respectively. In the more general case, where the object and image space refractive indexes are different, Eq. 15 becomes

$$\frac{n'}{l'} - \frac{n}{l} = K. \tag{18}$$

Now, the power of the system is defined with the aid of Fig. 12. A ray parallel to the axis at a height η emerges at an angle u'_k. If n' is the image-space refractive index, then

$$K = \frac{1}{F} = -\frac{n'u'_k}{\eta}, \tag{19}$$

where again F is the EFL of the system. It should be clearly understood that the power is independent of the choice of the height η. It is also independent of the direction of propagation, in other words, whether light comes from the left, as in Fig. 12, or from the right. This holds whether or not the system has any left to right symmetry.

In summary, both the thin-lens-equivalent conjugate equation (Eq. 18) and the magnification formula (Eq. 8) hold for an arbitrary system, provided that the distances l and l' are measured from the principal planes. It remains to specify the principal plane locations. These are normally specified by reference to the first and last surfaces of the system, as in Fig. 13. We trace a ray parallel to the axis at a height h_1 in the object space. It intercepts the last surface at a height h_k and forms an angle u'_k with the axis. If δ' is the distance between the last surface and the image-space principal plane,

$$\delta' = \frac{h_1 - h_k}{u'_k}(u_1 = 0). \tag{20}$$

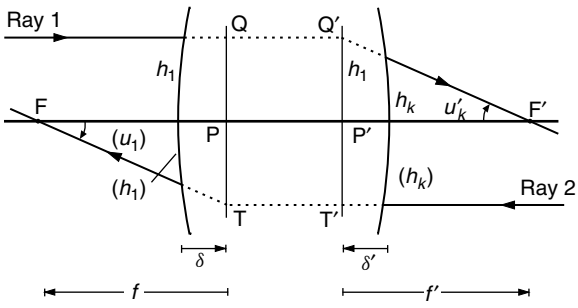


Figure 13. Location of the principal planes.

Similarly, for a ray parallel to the axis from the right, or equivalently, for a ray coming from the left that emerges parallel to the axis, we obtain (Fig. 13)

$$\delta = -\frac{h_1 - h_k}{u_1}(u'_k = 0), \tag{21}$$

where the meanings of h_1 and h_k in the two cases are different (associated with a different ray). These two formulas complete the basic equations of Gaussian optics and also show its equivalence to the previous ray-tracing methods. All of the quantities in Eqs. (19)–(21) can be computed by the ray-tracing procedures of the previous section. However, the importance of the power and the principal planes is that they allow us to make general inferences about the system and *they can be determined experimentally even without knowledge of the internal construction of the system*. The way to achieve this is beyond the scope of this article [see (2,8)].

Afocal Systems

Principal planes cannot be established for afocal systems, as the construction of Fig. 11 demonstrates (the ray would remain parallel to the axis after exiting the system). Also, Eq. (19) shows that the power of all afocal systems is zero ($u'_k = 0$), so we cannot distinguish between afocal systems by their power. Afocal systems are a special case, which nevertheless can also be treated generally in a simple way (2). Because the linear magnification for an afocal system is constant (say M_0), we can use it instead of the power. The procedure is as follows. First, we establish any set of convenient conjugate planes by tracing a single ray through the system. Then, we use those conjugates as reference planes, from which to measure object and image distances. Thus, if l and l' are the conjugate distances measured from our set of reference planes,

$$l' = \frac{n'}{n}M_0^2 l. \tag{22}$$

DETERMINATION OF IMAGE IRRADIANCE

The Marginal and Pupil Rays and the Optical Invariant

Up to this point, we have shown that the image location and size can be determined by tracing a single ray, given the object and the system construction. But the significance of paraxial and Gaussian optics can be extended even further to determine a first (but often very good) approximation of the amount and distribution of EM energy that reaches the image from the object through the lens. This can be accomplished by tracing only one more paraxial ray, thus providing a striking demonstration of the economy and power of these methods.

The magnification formula, Eqs. (8) or (10) may be rearranged as

$$nu\eta = n'u'\eta' \equiv H. \tag{23}$$

So the product $nu\eta$ is an invariant. This holds for the entire system as well as for an individual surface. The invariant H is usually called the Lagrange invariant, but

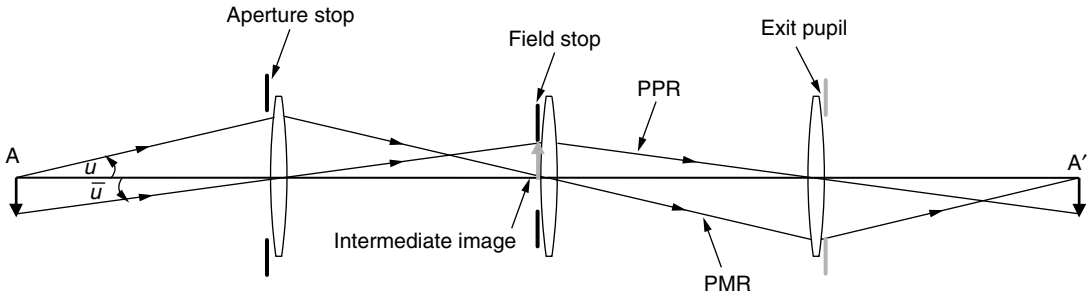


Figure 14. Imaging by a realistic optical system. The size of the light cones accepted into the system from any object point is limited by the aperture stop, or more generally, by the *entrance pupil*, which in this case coincides with the aperture stop. The maximum object size that can be imaged though the system is determined by the field stop.

it may be also called the Helmholtz invariant, or the optical invariant in (2). As we will see, it determines the amount of light accepted by the system (throughput), as well as the maximum number of resolvable spots in the system (information capacity).

Up to this point, we have set no limit on the actual length of the object (η)—we have been concerned only with the magnification. The angle u will also be limited to a maximum value by the size of the lens opening, which has not been a concern until now. Although the invariance of Eq. (23) holds for any values of u and η, it acquires its fundamental significance if we interpret both u and η as the maximum values permitted by the system. This is depicted in Fig. 14. In the figure, the object at A is imaged at A′ at unity magnification, but this is not essential. This figure contains all of the important elements one needs to understand the transfer of EM energy through the lens system to a first approximation. We trace two rays, one from the axial object point A and one from the edge of the object. The first ray also passes through the edge of the "aperture stop," which is a diaphragm that determines the size of the cone of light admitted into the system by any object point. This ray is called the marginal ray. The ray from the edge of the object is chosen to pass through the middle of the aperture stop; it is called the principal, chief, or pupil ray.

The image of the aperture stop through the system is the "exit pupil," as shown. The size of the exit pupil determines the angle of the light cones that converge onto image points. The corresponding size of the object-space cones is determined in this diagram by the aperture stop, but in the general case, the aperture stop may not be located at the front of the system, but somewhere inside it. In that case, the image of the aperture stop through the preceding elements is the "entrance pupil." For the system shown, the aperture stop and the entrance pupil coincide.

The size of the image is limited by the "field stop." This is a diaphragm placed at an intermediate image location. If no intermediate image is formed by the system, the field stop may be provided by an aperture across the image itself, for example, the dimensions of the photodetector that records the image. Similarly, there may be no clearly defined diaphragm that plays the role of the aperture stop. The aperture and/or field stops may be determined by the size of those lens elements that come closest to limiting the marginal and pupil ray angles. For example, it is evident from the figure that even if we had no physical diaphragm at the aperture stop location, the angle of the marginal ray would still be limited by the first or the third (but not the second) lens diameter. Similarly, in the absence of a field stop, the size of the image would be determined by the size of the second lens.

The two rays shown in Fig. 14 determine the size and the location of the object and image, as well as the size and location of the entrance and exit pupils. These two rays are traced according to the paraxial equations shown previously, and for this reason, we designate them as the paraxial marginal ray (PMR) and the paraxial pupil ray (PPR), correspondingly. Despite the use of paraxial equations, it is assumed that the field (object) size and the aperture size correspond to the actual system values, and are thus finite, and possibly quite large. Because the paraxial equations are not valid far from the axis, it would appear that their use would lead to unacceptably large errors. However, that is not the case. The two conditions shown as equations 24 and 25 allow us to extend the validity of the paraxial equations and in fact arrive at predictions that are quite accurate for systems that do not have large aberrations. The two conditions relate the paraxial angle variable associated with the marginal and pupil rays to the actual (finite) field and aperture angles. Thus, let u be the paraxial angle variable associated with the PMR, and let a be the corresponding finite aperture angle (or angular aperture size—see Fig. 14). Similarly, let $\overline{u}$ be the paraxial angle variable associated with the PPR and $\overline{a}$ the corresponding finite field angle (or angular field size). Then, we set

$$u = \sin a \tag{24}$$

and

$$\overline{u} = \tan \overline{a}. \tag{25}$$

Thus, for example, if a system is said to have a semifield of view of 45°, we set the PPR angle variable $\overline{u}$ as $\tan(45^\circ) = 1$. Or, if the system is said to have an aperture angle of 30°, we would set the PMR angle variable u equal to $\sin 30^\circ = 0.5$. These two equations represent the link between paraxial and finite optics and allow us to extend

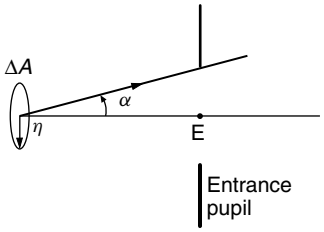


Figure 15. Flux collected by the entrance pupil of a lens from a small Lambertian source.

the usefulness of paraxial optics to the finite domain. Full justification of this extension is given in (2).

Flux Collected by a Lens. Image Irradiance

To proceed further, we must assume familiarity with some radiometric and photometric concepts that are discussed elsewhere in this Encyclopedia [see also (2,3,8)]. The basic concept borrowed from physical optics is that of the radiant power (or flux), measured in watts, and its photometric equivalent, the luminous flux, measured in lumen. Other radiometric (photometric) quantities are then constructed by the use of areas and solid angles. Image irradiance (illuminance) is simply the flux per unit area incident on the image surface. For the source, the most important quantity is the radiance (luminance), given as flux per unit solid angle per unit area projected normally in the direction of observation. In Fig. 15, let the source have area ΔA and a radius equal to η. At the entrance pupil of the lens, a fraction $\Delta\Phi$ of the total flux Φ emitted by the source is collected. This fraction is given by

$$\Delta\Phi = \left(\frac{\pi}{n}\right)^2 LH^2, \tag{26}$$

where L the source radiance (luminance) and $H = nu\eta = n\sin a\eta$, the Lagrange invariant of the lens. Thus, we see that *the flux collected by the lens is proportional to the square of the Lagrange invariant*.

If the lens has a transmission factor τ, then flux equal to $\tau\Delta\Phi$ will be distributed across the image surface $\pi\eta'^2$. The image irradiance is then

$$E = \frac{\tau\Delta\Phi}{\pi\eta'^2}. \tag{27}$$

The advantage of using Eq. (24) is that for systems in which the aberrations are reasonably well corrected, the relationship $u' = \sin a'$ also holds in the image space. Therefore, $H = n'\sin a'\eta'$. Substituting in Eq. (27) and using Eq. (26) gives

$$E = \frac{\pi\tau L}{n^2} n'^2 \sin^2 a'. \tag{28}$$

The quantity $n'\sin a'$ is called the *numerical aperture* (NA) of the lens. It is defined in the image space if the object is more distant than the image. More frequently, it is defined in the object space ($n\sin a$) for a nearby object, as for a microscope objective. The quantity $1/(2n'\sin a')$ is called the *relative aperture* or simply the *F number (F#)*. This quantity is more often used for a distant object than the NA, for example, for photographic objectives. In terms of the $F\#$ then, Eq. (28) becomes

$$E = \frac{\pi\tau L}{n^2}\frac{1}{4(F\#)^2}. \tag{29}$$

As mentioned, the last two equations are more appropriate when the object is distant. When both object and image are nearby, it may be useful to express the image irradiance in terms of the linear magnification M. By combining Eqs. (26) and (27) and using $H = n\sin a\eta$, we obtain

$$E = \frac{\tau\pi L\sin^2 a}{M^2}. \tag{30}$$

The problem of determining the image irradiance has now been solved.

Notice that pupil ray quantities do not appear explicitly anywhere in the last few equations. For example, both the angle α and the magnification can be thought of as associated with the marginal ray. Thus, it may appear that the image irradiance can be determined by using only one ray, rather than two, as we claimed at the beginning of this section. However, pupil ray quantities are implied. For example, the angle α is specified by the ratio of the pupil size (determined by the marginal ray) and the location of the pupil or the distance between the image and the pupil (determined by the pupil ray). It is best to think that the PMR determines the location of the object/image and the size of the pupils and the PPR determines the size of the object/image and the location of the pupils. This symmetry between the PMR and the PPR is exploited in aberration theory.

IMAGE QUALITY LIMITATIONS

Limitations due to the Dispersion of Optical Materials: Chromatic Aberrations

The paraxial approximation and the image formation principles discussed so far contain one unstated assumption, namely, that the refractive index n is constant. In reality, that is correct only for a single wavelength. Optical materials exhibit *dispersion*, that is, variation of the refractive index with wavelength. Therefore, for any system that contains refractive elements, all of the system properties will generally vary with wavelength. In the Gaussian sense, we are concerned with the power and the principal plane location. The chromatic variation of those system properties leads to two directly observable chromatic effects or aberrations: the chromatic variation of focus (also called longitudinal or axial chromatic aberration) and the chromatic variation of magnification (also called transverse or lateral chromatic aberration).

To characterize these effects, we need first to characterize the dispersion of optical materials. This is done through the so-called Abbe V number (or occasionally, constringence), defined as follows. Let n_1, n_2, n_3 be the values of a material's refractive index at wavelengths λ_1, λ_2, λ_3, respectively, where $\lambda_1 < \lambda_2 < \lambda_3$. These wavelengths, it is

assumed, span the range of interest, and λ_2 is approximately in the middle of that range. For the range of visible wavelengths, the most common wavelengths taken are 486 nm (F hydrogen line), 588 nm (d helium line), and 656 nm (C hydrogen line). Then, the V number is defined as

$$V = \frac{n_2 - 1}{n_1 - n_3}. \tag{31}$$

The Abbe V number characterizes the dispersiveness of glass in the following sense: its inverse, $1/V$, is equal to the chromatic variation of the angle by which a thin prism made from the given glass deviates light.

Now, we can give the chromatic aberrations of a thin lens in air, as follows. Let l'_3 and l'_1 be the image distances from the lens for wavelengths 3 and 1, respectively; then, the chromatic variation of focus is

$$\frac{1}{l'_1} - \frac{1}{l'_3} = \frac{K}{V}, \tag{32}$$

where K is implicitly K_2, that is, the power at the middle wavelength, and V is the Abbe value of the material from which the thin lens is made.

The chromatic variation of image height is given by

$$\eta'_1 - \eta'_3 = -\overline{h}l'\frac{K}{V}, \tag{33}$$

where K and l' refer to wavelength 2 and $\overline{h}$ is the pupil ray height at the thin lens. This formula makes it clear that there is no chromatic variation of magnification (or image height) if the aperture stop is located at the (thin) lens (in which case, $\overline{h} = 0$). Justification for Eqs. (32) and (33) can be found in (2).

In a compound system that may contain many different optical materials of different refractive indexes and dispersive properties, calculating the chromatic errors is more complicated. We may note, however, that the difference in V value between different materials may be exploited to create achromatic systems, that is, systems that exhibit small chromatic errors. Further details may be found in (2,3,8–10).

Limitations due to the Wave Nature of Light: Information Capacity

In the geometrical approximation examined so far, we have seen that it is possible for all rays to cross at a single point, leading to perfect point image formation at least in the paraxial region and for a single wavelength. However, the geometrical approximation is valid only when the wavelength of light can be considered very small. The wave nature of light dictates that it cannot be focused to a region much smaller than its wavelength. Thus, the image of a point will have a certain minimum size, even if the system is otherwise perfect. The irradiance distribution in the image of a point is referred to as the point-spread function (PSF). For a system that suffers no other imperfections, the PSF is determined by the shape of the exit pupil (e.g., circular, elliptical, irregular) and the irradiance distribution across it (see later sections). A most commonly encountered case is that of a circular pupil shape that is uniformly illuminated. In this case, the rotational symmetry means that the PSF is a function only of the radius ρ:

$$G(\rho) = (\text{const})\left[\frac{J_1(2\pi\rho)}{\pi\rho}\right]^2 \tag{34}$$

where J_1 is a Bessel function of the first kind. The form of $G(\rho)$ is illustrated in Fig. 16. The irradiance distribution that corresponds to Fig. 16 is called the Airy pattern (after Sir George Airy, Royal Astronomer, who had many good opportunities to observe images of point objects). Diffraction theory (11–13) shows that the radius of the first zero, ρ_0, is given by

$$\rho_0 = \frac{0.61\lambda}{\text{NA}}, \tag{35}$$

where the NA is understood to refer to the image space. The images of two points separated by ρ_0 are said to be just resolved, and ρ_0 is traditionally thought of as the resolution limit of an optical system (also called the Rayleigh resolution limit — but see later sections for some additional discussion).

The maximum number of resolvable spots in the image determines the information capacity of the optical system. If η' is the image height, the number of resolvable spots across the image area is proportional to $(\eta'/\rho_o)^2$. Using Eq. (35), we find that the information capacity I of a

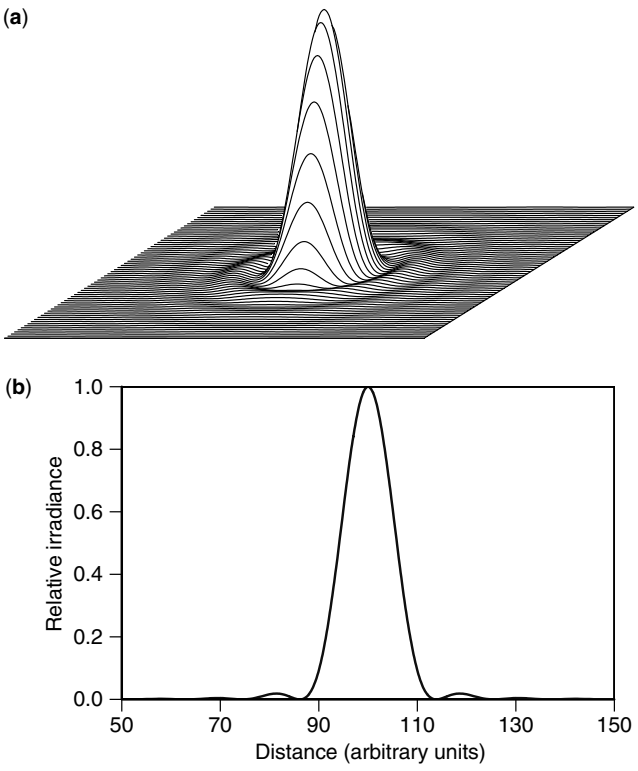


Figure 16. (**a**) Three-dimensional representation of the image of a point formed by a diffraction-limited system of circular aperture. (**b**) The same function in cross section, displaying the relative heights of the secondary maxima.

system is

$$I \propto (\mathrm{NA})^2 \eta'^2 = (n' \sin \alpha')^2 \eta'^2 = (n' u' \eta')^2 = H^2, \qquad (36)$$

using Eqs. (23) and (24). Thus the information capacity of an optical system is proportional to the square of the Lagrange invariant.

Clearly, this calculation holds for a single wavelength and thus implicitly excludes chromatic aberration. Monochromatic aberrations are also common (see next section) and have the effect of further blurring the image (or increasing the resolution limit), thus limiting the information capacity. In that case, it can be seen that H^2 represents the upper limit in information capacity that is achievable by any (incoherently illuminated) system performing the given imaging task.

Limitations due to Nonparaxial Geometry: Monochromatic Aberrations

Here, we begin to examine the properties of whole-field imaging systems outside the paraxial region, where images are no longer perfect. There are two complementary ways of examining those properties. The first is based on the geometric properties of the wave fronts and rays, and the second takes the wave nature of light into account. The geometric way is only an approximation, useful when the wavelength can be considered negligibly small. It owes its utility to the fact that geometric calculations are considerably faster than wave (or diffraction)-based ones. Even with the speed of today's computers, that is a considerable advantage. In principle, diffraction-based calculations are more accurate. However, as the image quality declines, diffraction-based calculations become extremely time-consuming and require very large arrays for an adequate representation of the aberration. Thus, geometric techniques are also useful at the limit of large aberrations.

The linear paraxial ray-tracing equations (4), (5) lead to perfect image formation (all paraxial rays that originate from a single object point meet at the same image point). Those equations are linear in u, due to the paraxial approximation $\sin u = u$. When the difference between the sine and the angle can no longer be neglected, the resulting nonlinear ray-tracing equations (involving $\sin u$) lead to the result that rays from the same object point cross the image plane at different places. There are a few special-case exceptions to this (e.g., a point object at the center of curvature of the surface), but in general, monochromatic aberrations are the inevitable result of image formation outside the paraxial region. Imaging of an axial point is also subject to aberrations if the rays collected by the lens form a finite angle with the axis.

Generally, the aim of optical design is to control the amount of aberration by using the smallest possible number of (preferably spherical) refracting or reflecting surfaces. A system in which the aberrations are negligible is called *diffraction-limited*, because the resolution of such a system will be determined only by the diffraction spread in the image of a single point, as per the previous section. A system may be diffraction-limited on-axis, but aberration-limited off-axis. It is also possible for a system to be diffraction-limited off-axis.

The basic tool for deciding whether the system has the desired image quality is finite ray-tracing. A finite ray is any ray that does not stay infinitesimally close to the axis. The only two exceptions to this definition are the two Gaussian rays, the PMR and the PPR, which are traced according to the laws of paraxial optics, even though they form finite angles with the axis. However, the ray paths of the PMR and the PPR are fictitious; they are close to the paths of real rays only in the object and image spaces, assuming that the aberrations are not large and that the correct convention is chosen [Eqs. (24), (25)]. Also, there is no point in tracing more Gaussian rays. Gaussian optics assumes perfect image formation, so no additional information would be obtained. By contrast, because finite rays all cross the image plane at different locations, in general, many finite rays are needed to determine the ray intersection pattern at the image plane.

Finite ray-tracing equations are considerably more involved than paraxial ones and are not given here. They may be found in several references, for example (2,9,14). They also form the basis of the many optical design programs available for personal computers. It is worth noting that those programs have evolved to the point that even a nonexpert can use them successfully to evaluate a given system, although design itself requires the appropriate expertise.

Aberration may be characterized either through the rays or through the wave front. Ray aberration is, in general, the distance between a finite ray intersection with a desired image plane and the corresponding intersection of a Gaussian ray. Wave-front aberration is the optical path difference between the wave front and a chosen reference sphere, measured along the ray. Optical path difference is defined as the geometric distance times the refractive index of the medium. Qualitatively, the wave-front aberration measures the departure of the wave front from sphericity.

Because rays are normal to wave fronts, ray and wave-front aberrations are related: The former is the derivative of the latter with respect to the exit pupil coordinates. Specifically, Fig. 17 shows a schematic of an optical system in which a ray proceeds from axial object point O to the image space, passing through the exit pupil located at E′. An exaggeratedly nonspherical wave front is also shown. Let the exit pupil have a height h'_e, and let the radius of the reference sphere be R', equal to the distance

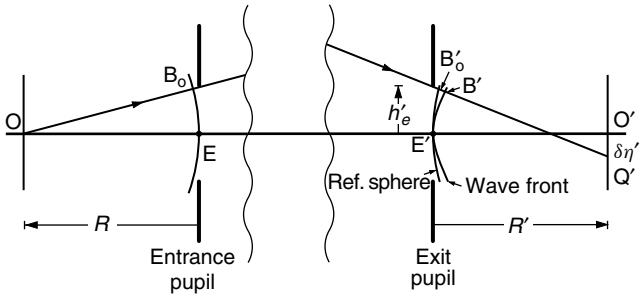


Figure 17. Wave front and ray aberrations.

between exit pupil and Gaussian image (E′O′). Let the refractive index of the image-space medium be n'. A real ray is perpendicular to the wave front at B′ and cuts the reference sphere at B'_0. The path $(B'_0O') = R'$ represents the path of a Gaussian ray that crosses the axis at the paraxial image point O′. The wave-front aberration W in this figure is given by

$$W = n'(B'_0B'). \tag{37}$$

Finally, let the vertical axis be y' at the pupil and η' at the image. The ray aberration is $\delta\eta' = (O'Q')$. All of these quantities are related through

$$\delta\eta' = -\frac{R'}{n'h'_e}\frac{\partial W}{\partial y'}. \tag{38}$$

A similar equation can be written for the x direction. In this expression, R', h'_e, W, and $\delta\eta'$ have units of length, and y' is dimensionless, normalized with respect to h'_e. Proof for this expression can be found in (2,9,15).

By choosing to represent the wave-front aberration at the exit pupil, we can conveniently fold the effects of diffraction into the description because diffraction effects will be dominated by the limiting aperture of the system. In the image space, the limiting aperture is the exit pupil. The wave-front aberration is the preferred way of describing aberrations and provides a lucid and understandable description, although, traditionally, optical designers have favored the more esoteric ray aberration. It is worth reviewing the traditional aberration types here because they are often encountered in imaging systems even outside of optics, for example, in imaging by an electron beam.

The wave theory of aberrations was put in a consistent theoretical framework by H. H. Hopkins, in his out-of-print 1950 classic by the same title (15). Welford (9) gave an alternative but similar treatment, and Mouroulis and Macdonald (2) provided a treatment of the introductory aspects of the problem following Hopkins's notation. The book by Mahajan (16) also contains derivations of the wave aberration of several common optical systems, as well as a complete theoretical treatment.

For a single point object, the wave-front aberration can be described as the difference in shape between the actual (generally nonspherical) wave front in the image space and a corresponding reference sphere that represents an ideal wave front. The wave-front aberration can then be represented as a polynomial with respect to the pupil coordinates. However, every point in the object (e.g., an on-axis point and an off-axis point) will generate a different wave-front shape. In other words, the wave-front aberration is, in general, field-dependent. For this reason, we represent the wave-front aberration as a polynomial that contains both aperture and field terms. However, the inclusion of field terms in the wave-front aberration means that it is possible to have field-dependent aberrations where the corresponding wave-front shape is spherical but the location of the image departs from the paraxial prediction. This is explained further later.

Considerations of symmetry dictate that some combinations of terms can be excluded from the polynomial. The types of terms depend on the particular form of the symmetry. For example, a cylindrical lens has a set of aberrational terms different from those of a spherical lens. We restrict our attention to the common case of rotationally symmetrical systems. In that case, the wave-front aberration polynomial can be written as

$$W = W_{040}r^4 + W_{131}\tau r^3\cos\theta + W_{222}\tau^2r^2\cos^2\theta + W_{220}\tau^2r^2 + W_{311}\tau^3r\cos\theta + (\text{higher order terms}) \tag{39}$$

where r is the pupil radial coordinate, τ the field radial coordinate, and θ the azimuth. The coefficients W_{ijk} represent the various aberrational terms and have units of length, and the radial coordinates are normalized with respect to their maximum values. In this notation, the subscripts are chosen to recall the exponents of the variables. Equations (38) and (39) are related through $y = r\cos\theta$. The magnitude of W is typically less than one wavelength for systems of high image quality and not more than a few wavelengths for most well-corrected systems. There are, however, cases where high image quality is not a concern, for example, lenses used with array detectors that possess coarse pixel size, in which case the aberration can be many wavelengths.

The five terms in Eq. (39) are the so-called primary aberrations, also known as third-order or Seidel aberrations. Specifically, $W_{040}r^4$ is called *spherical aberration*. This is a radially symmetrical term that has no field (τ) dependence. The departure of the corresponding wave front from the spherical shape is shown in Fig. 18. Such a departure from sphericity gives a variety of possible irradiance distributions that depend on the amount of aberration and the exact location of the focus. These images are best appreciated when compared with the image of a nonaberrated point. For this reason, we show them in the next section.

$W_{131}\tau r^3\cos\theta$ is called *coma*. The departure of the comatic wave front from sphericity is shown in Fig. 19. It can be seen that coma does not impart an overall curvature to the wave front and thus does not change the position of best focus along the axis. It does, however, produce an overall tilt to the wave front, that results in a shift of the best image point from the Gaussian prediction.

$W_{222}\tau^2r^2\cos^2\theta$ is called *astigmatism*. This aberration imparts a cylindrical error to the wave front, as shown in

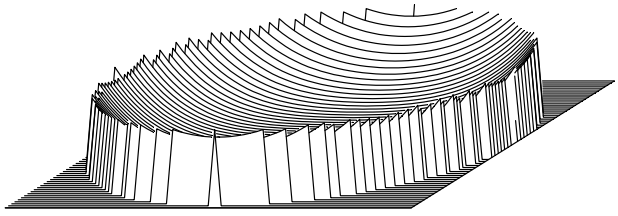


Figure 18. Wave-front shape for a system suffering from spherical aberration. The corresponding reference sphere is flat. The departure from the reference has a fourth power dependence on the radius. Notice that spherical aberration adds an overall curvature to the wave front.

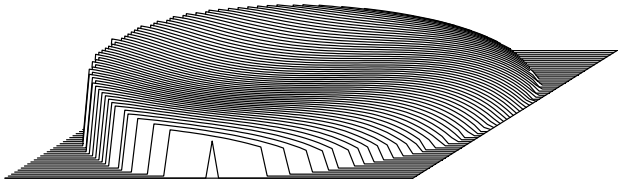


Figure 19. Wave-front shape for a comatic wave front (flat reference).

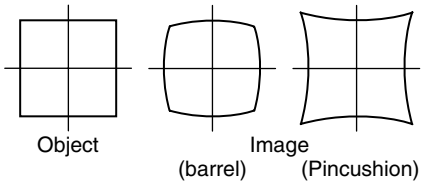


Figure 22. Image of a square object (left) through a system suffering from distortion.

Fig. 20. Again, the corresponding point images are shown in the next section. For the last two figures, we note that the field dependence is not particularly relevant; it merely determines the amount of aberration but does not change the basic wave-front shape. However, for the two remaining aberrations following, the wave-front shape is actually spherical, and only the field dependence is of importance.

$W_{220}\tau^2 r^2$ is called *field curvature*. Although the wave front is spherical, the plane of best focus shifts as field angle increases. The result is that the best image is not obtained on a plane but on a curved surface. If it were possible to curve the receiving (detector) shape to conform to the image surface, a stigmatic (point-like) image would be obtained throughout. However, usually the detector surface is flat, and in that sense field curvature is as significant an aberration as the first three. Figure 21 shows the effect of field curvature.

$W_{311}\tau^3 r\cos\theta$ is called *distortion*. Again, this represents a spherical wave-front shape. Distortion represents a change in magnification across the field. In negative or pincushion distortion, the magnification increases with field, and in positive (or barrel) distortion, the magnification decreases with field. The corresponding image of a square is shown in Fig. 22.

Image imperfections are not exhausted by these five types of aberrations. Higher order aberrations exist, as do other types in systems of different symmetry. However, these primary aberrations are important because they describe the most fundamental and commonly encountered types of image degradation. They are also very easy to compute: for rotationally symmetrical systems, it can be shown (2,9) that the computation of these primary aberrations involves only paraxial ray tracing of the two Gaussian rays, the PMR and PPR previously discussed.

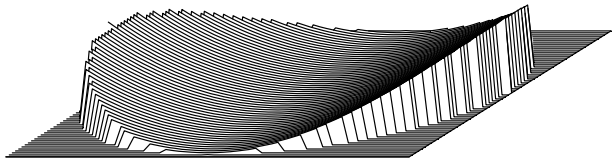


Figure 20. Wave-front shape for an astigmatic wave front (flat reference).

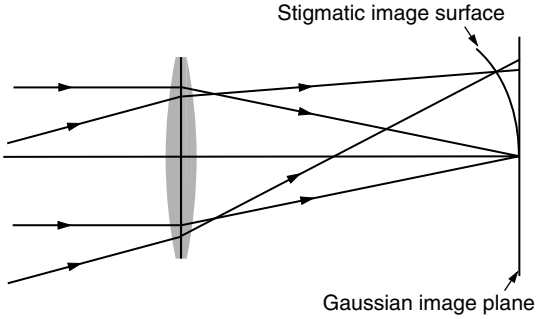


Figure 21. Field curvature: The best (stigmatic) image is obtained on a curved surface.

IMAGE FORMATION IN THE SPATIAL DOMAIN (CONVOLUTION): THE POINT-SPREAD FUNCTION

The Convolution Integral

Intuitively, we can consider the image of an extended object as the summation of the images of all points that comprise the object. Let $G(u, v)$ be a function that represents the image of a point, called the point-spread function, (PSF). An example of such a function is shown in Fig. 16 for a system that has no aberration. More examples will be shown later. Let $B(u, v)$ be the irradiance distribution in the object. The summation of the point images [where peak irradiances are weighted by $B(u, v)$] is expressed mathematically as a convolution integral (see also Appendix):

$$B'(u', v') = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} B(u, v)G(u' - u, v' - v)\,du\,dv. \qquad (40)$$

The intuitive appeal of this way of describing image formation hides two critical assumptions in Eq. (40). The act of summing (or integrating) implies that the total irradiance is obtained by summing the irradiances of individual points. This is the *linearity* assumption, and it is satisfied if the system is illuminated *incoherently*. This corresponds to the everyday experience of adding the irradiances produced on a receiving screen by two different lamps. However, with coherent illumination, interference effects cause the total irradiance to differ from the sum of the individual irradiances. For our purposes here, we may define incoherent illumination as follows: The phase difference of the EM field emitted from two different source (object) points varies randomly and rapidly relative to the integration time of the photodetector. If this condition holds, interference effects from the two source points are averaged out over time and thus are not observed.

The second assumption is that the function $G(u' - u, v' - v)$ does not change form as it shifts across the image, in other words, that it is the same function for every u', v'. This is the assumption of *stationarity*. To

connect this discussion with the previous section, notice that if the system suffers from any aberration other than spherical (in other words, any field-dependent aberration), it is strictly speaking nonstationary. However, we can apply Eq. (40) even in such a case, provided we restrict the region of integration over an area where the variation of the PSF is negligibly small and provided such area is large compared with the extent of the PSF. Thus, we can consider that an extended image comprises a set of patches across which the PSF is constant.

For the vast majority of whole-field imaging optical systems, the assumption of incoherence is justified. However, there are cases where the system is coherently illuminated. Even in that case, image formation can still be described as a convolution, provided that we do not sum the irradiances but the complex amplitudes (amplitude and phase) of the field. In other words, for coherent illumination, the PSF is the complex amplitude (rather than irradiance) distribution in the image of a point. If $F(u, v)$ is the coherent PSF and $A(u, v)$ is the complex amplitude distribution in the object, then coherent image formation can be expressed as

$$A'(u', v') = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} A(u, v)F(u' - u, v' - v)\, du\, dv, \qquad (41)$$

where A' describes the complex amplitude distribution in the image.

In coherent illumination, the stationarity requirement is more involved, and we postpone a more complete discussion until later. In summary, Eqs. (40) and (41) show that *the problem of image formation is reduced to the problem of determining the image of a single point*. The next section shows how the image of a point may be determined.

Since the image of a point has a certain minimum spread determined by diffraction, even in the absence of aberrations, it follows that the image of an extended object is bound to be an inaccurate, or blurred, representation of the object. This means that some details present in the object will be lost. Thus convolution can be thought of as a blurring operation.

The Relationship between the PSF and the Pupil Function

If we know the complex amplitude distribution over the exit pupil, we can determine the form of the PSF. Referring back to Fig. 17, consider the exit pupil of an optical system at E′, and let a reference sphere be centered at point O′. This point may be the Gaussian image point or any nearby convenient point, for example, one that minimizes the difference between the reference sphere and the wave front. The wave front shown is generally nonspherical, although the difference is greatly exaggerated for clarity. The reference sphere represents a perfect wave front. Let $f(x, y)$ be the complex amplitude distribution at the reference sphere and $s(x, y)$ be the real amplitude distribution (representing the variation of transparency or transmittance). The wave-front aberration W is the difference in the optical path length between the reference sphere and the wave front, so the phase difference between the two is $2\pi W/\lambda \equiv kW$, where k is the propagation constant. Therefore,

$$f(x, y) = s(x, y) \exp[ikW(x, y)]. \qquad (42)$$

In practice, the procedure for determining the PSF is as follows. First, the wave-front aberration is determined by finite ray tracing up to the exit pupil. This establishes $W(x, y)$ for a grid of sampling points. Here, a 2-D polynomial may be interpolated if desired. Information about the variation in system transmittance (technically called *apodization*) must also be supplied, although often it is assumed that the system has uniform transmittance. Thus $f(x, y)$ is established. At this point, the effects of diffraction are folded in by considering the propagation between the exit pupil and the image plane as a diffraction process. This process is described in some detail in the books by Goodman (17) and Gaskill (18). Here we give the result, which is that apart from a constant factor, the complex amplitude distribution in the image of a point is the Fourier transform of the pupil function (see Appendix for an introduction to the Fourier transform):

$$F(u, v) = \text{const} \int_A \int f(x, y) \exp[i2\pi(ux + vy)]\, dx\, dy \qquad (43)$$

The integration extends over the aperture or exit pupil area A because $f(x, y)$ is zero outside the pupil area. In this expression, the pupil coordinates x and y, it is assumed, are normalized with respect to the radius of the exit pupil, and the normalized image coordinates u and v are related to the real coordinates ξ and η through

$$u = \frac{(n \sin a)\xi}{\lambda}$$

and

$$v = \frac{(n \sin a)\eta}{\lambda}, \qquad (44)$$

where $(n \sin a)$ is the numerical aperture in the image space. The validity of Eq. (43) assumes that the conditions for scalar diffraction apply, namely, that the angles involved are not large. In practice, vector diffraction effects are small, even for a numerical aperture of 0.5 or so. It also assumes that the field angle is not too large, otherwise the coordinate normalization scheme must be redefined (19).

The complex amplitude distribution F is the PSF for a coherently illuminated system. For an incoherent system, the PSF is

$$G(u, v) = |F(u, v)|^2. \qquad (45)$$

Coherence has two aspects, spatial and temporal. Spatial coherence is defined as the degree of correlation between the EM fields emitted by two different source points. Temporal coherence describes the degree of monochromaticity of the light (or mathematically, the autocorrelation of the waveform emitted by the source). It is worth noting that even the incoherent PSF given before applies for a single wavelength, implying temporally coherent but spatially incoherent light. Quasi-monochromatic light

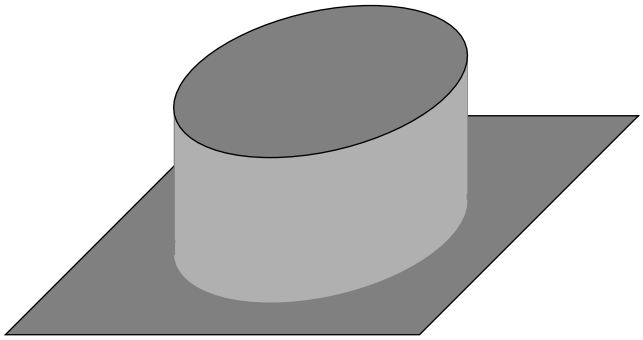


Figure 23. Shape of the pupil function for a system that has no aberration and a circular pupil.

(narrow wavelength band) provides a good approximation to the incoherent PSF.

The Diffraction-Limited PSF

For a system that has a circular pupil and no aberrations, the pupil function is simply a 'top-hat' function, unity inside the pupil and zero outside, as shown in Fig. 23. The incoherent PSF resulting from that function has already been shown in Fig. 16. The pattern has regular zeros and rapidly diminishing maxima. The peak irradiance of the first ring is only 1.9% of the maximum, and 83.8% of the total energy is contained (encircled) in the area up to the first zero.

Incoherent Two-Point Resolution Limit. The radius of the first zero (Airy disk) is given by Eq. (35) for a focusing system. This can also be given in angular terms as $1.22\lambda/d$, where d is the exit pupil diameter, which is valid for afocal systems as well. When two nearby points are separated by a distance equal to the radius of the Airy disk, they are said to be just resolved, and the radius of the Airy disk is considered the *incoherent, two-point resolution limit* or *Rayleigh resolution limit* (see also later for other kinds or resolution limits). The irradiance recorded by the photodetector is shown in Fig. 24 in cross section. Resolution of two points implies that the central dip is detectable. However, a low-noise photodetector can reliably detect a smaller dip than that shown in Fig. 24. For this reason, the resolution limit of Eq. (35) should be seen as a convention that does not reflect any deeper physics. For example, a more rational definition of the resolution limit could have been the distance that makes the central dip just vanish. This is referred to as the *Sparrow resolution limit* (20).

The Coherent PSF

The coherent PSF that corresponds to the pupil function of Fig. 23 is shown in Fig. 25. The coherent PSF is normally a complex function, and thus can be represented only in two separate graphs (amplitude and phase, or real and imaginary part). However, in this case, the phase of the function alternates between zero and π, so it can be represented as a real function. A photodetector, such as the human eye, will record only irradiance, and thus will not perceive directly any difference between the functions of

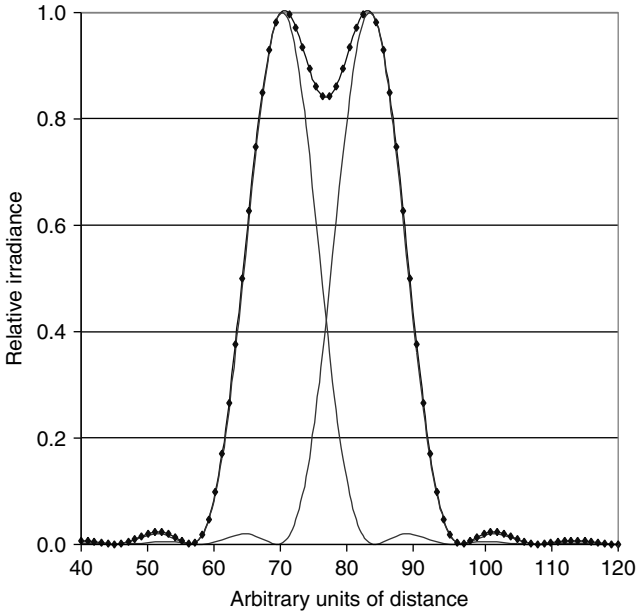


Figure 24. Irradiance distribution in the image of two incoherent points at the Rayleigh limit of resolution (diamonds). The individual irradiances of the two points are also shown in thinner lines.

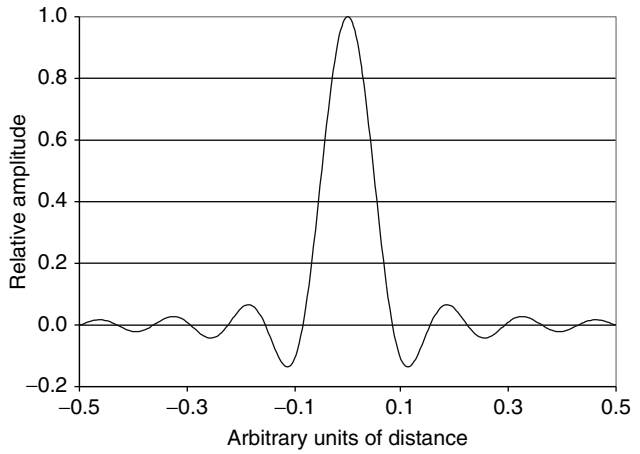


Figure 25. Coherent PSF of a system that has a circular pupil and no aberrations.

Figs. 16 and 25 (the former is simply the squared modulus of the latter). However, two coherent point images in close proximity give rise to a pattern much different from that of Fig. 24, due to the fact that the light from the two point images will interfere constructively or destructively, depending on the relative phase. The pattern will also depend on whether the two object points are cophasally illuminated. Figure 26 shows an example of the amplitude and irradiance distribution in the combined image of two coherent, nearby points, where the distance between the points is the same as that of Fig. 24. It can be seen that when the two object points are cophasally illuminated, they are not resolved, but when they are in antiphase, they are fully resolved. This example shows the complications of coherent imaging.

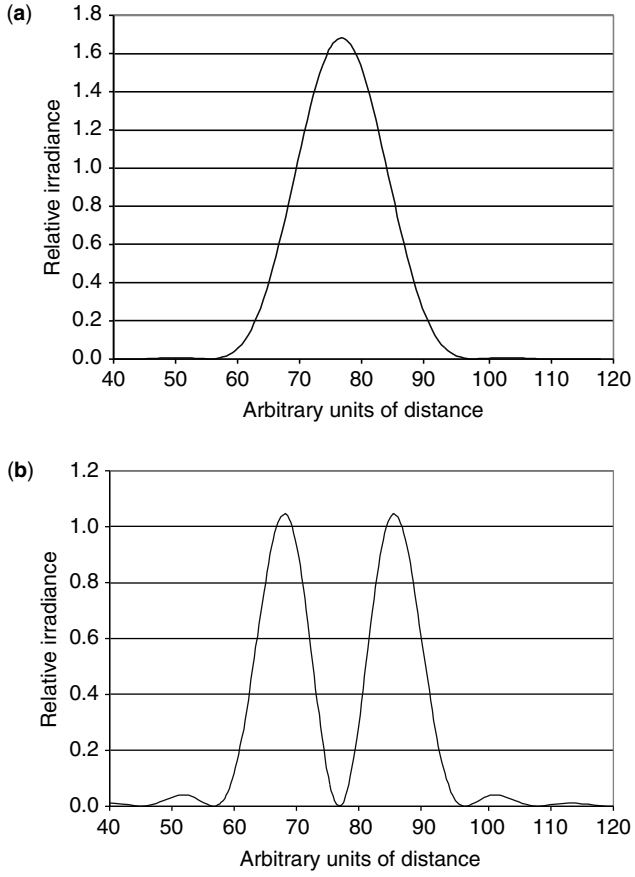


Figure 26. (**a**) Irradiance cross section plot in the image of two coherent and cophasal points at the Rayleigh distance. The two points are not resolved. (**b**) The same plot when the points are π out of phase. See color insert.

The Effect of Defocus. It is interesting to see how defocus modifies the pattern of Fig. 16. In this sense, defocus can be considered an aberration of the form $W_{020}r^2$ following the aberration notation in "Limitations due to Nonparaxial Geometry." This has the effect of simply changing the radius of the reference sphere (in other words, the perturbed wave front differs from the ideal by a spherical term that can be eliminated through refocusing). Figure 27 shows the effect of defocus using W_{020} as a parameter.

Apodization

Because there is no phase error (aberration), the diffraction-limited PSF is determined by the shape of the pupil and the real amplitude term of the pupil function. A circular pupil is commonly assumed, but it should be appreciated that even in the absence of aberrations, the pupil shape becomes elliptical or irregular at significant off-axis angles. The effect of pupil shape is evident at rather large field angles of some tens of degrees or when the system suffers from *vignetting*, where rays from the extreme parts of the aperture are (sometimes deliberately) cut off.

An example of a nonuniform real amplitude term (apodization) is given by astronomical telescopes, in which

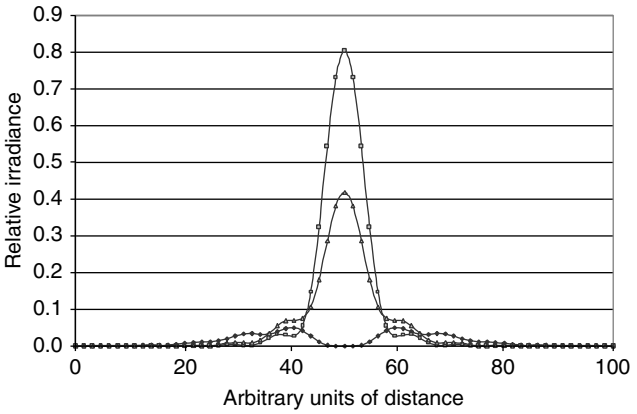


Figure 27. Effect of defocus on the incoherent PSF. The Airy disk is omitted for clarity, but it would have a peak of unity on the same scale. Squares: $W_{020} = 0.25\lambda$, triangles: $W_{020} = 0.5\lambda$, diamonds: $W_{020} = 1\lambda$. In the latter case, the central value of the pattern is zero. The total energy (integrated over x and y) is constant in all three patterns.

the secondary mirror often obscures the central part of the aperture. Figure 28 gives the form of the pupil function for a central obscuration of 12% of area. The corresponding PSF is in Fig. 29. It can be seen that the effect of the obscuration is to increase the energy in the rings at the expense of the central maximum. Various cases of nonuniform pupil shapes have been described by Mahajan (21).

The PSF in the Presence of Aberrations

Here, a brief, visual representation of the PSF in the presence of aberrations is given. Each aberrational term is a phase term, and the corresponding PSF is the squared modulus of the Fourier transform of the corresponding pupil function, $s(x, y) \exp[ikW(x, y)]$, where $s(x, y)$ is the same as in Fig. 23, and $W(x, y)$ is the corresponding aberrational term, as given in "Limitations due to Nonparaxial Geometry." The original work on this topic was done by Nijboer (22–24). Excellent pictorial representations of PSFs can be found in (25), and in the more easily accessible (10).

Spherical Aberration. The PSF in the presence of spherical aberration varies in a complicated way as

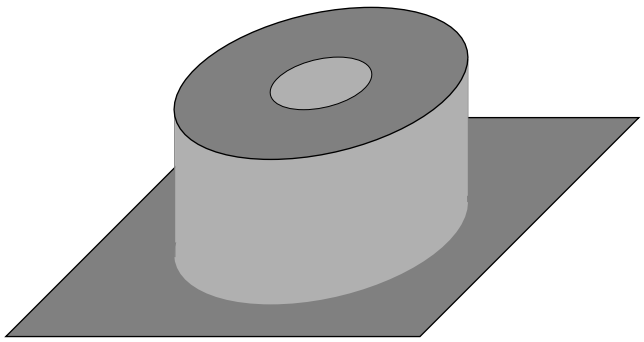


Figure 28. Form of the pupil function for a system that has no aberration and a centrally obscured pupil.

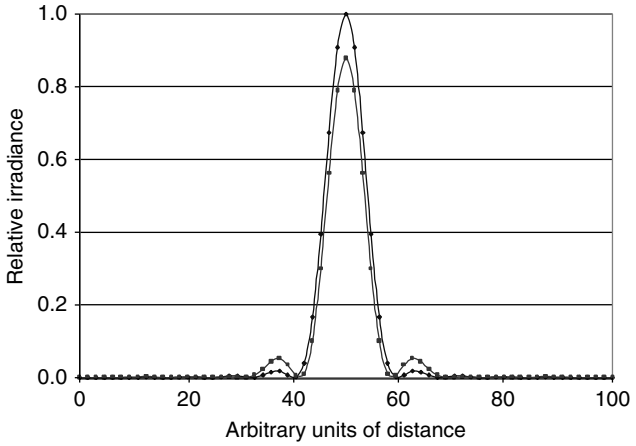


Figure 29. Comparison of the incoherent PSFs of an apodized (squares) and an unapodized system (diamonds). The latter PSF is the same as in Fig. 16. The two PSFs are normalized to the same amount of energy.

a function of defocus. Thus, in fact, the aberration function is taken as $W_{040}r^4 + W_{020}r^2$, where the second term represents defocus (a change in the radius of the reference sphere). This is a rotationally symmetrical aberration, so the PSFs are also rotationally symmetrical. Figure 30 shows the PSF section for one wave of aberration ($W_{040} = 1\lambda$) and at three different foci as noted, and Fig. 31 shows the PSF section for two waves of aberration.

Coma. Coma is an asymmetrical aberration that does not interact with defocus in the sense that the best image in the presence of coma is at the paraxial focal plane. However, the location of the PSF centroid is shifted

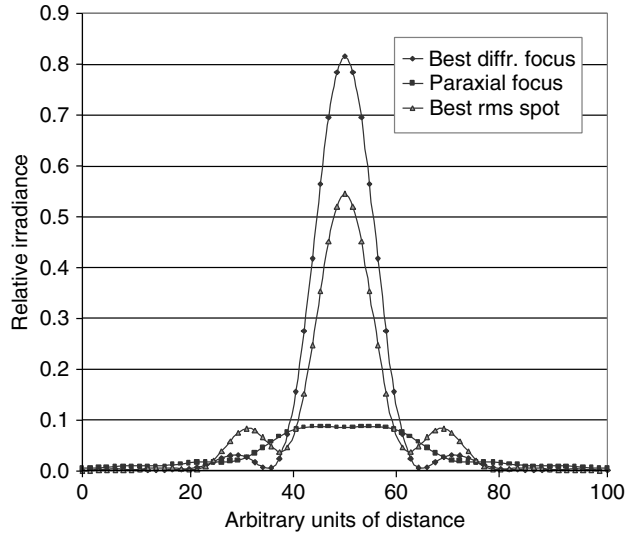


Figure 30. Plot of the incoherent PSF in the presence of one wave of spherical aberration and at various positions along the axis. Diamonds: best diffraction focus, triangles: minimum geometrical spot size; squares: paraxial focus. The pattern that corresponds to the minimum geometrical spot size is called traditionally the "circle of least confusion," although the name is not justifiable from the diffraction viewpoint.

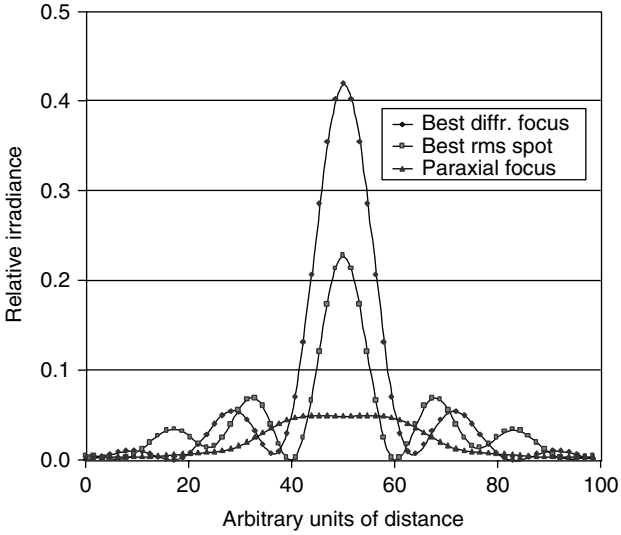


Figure 31. Plot of the incoherent PSF for two waves of spherical aberration and various positions along the axis: Diamonds: best diffraction focus, squares: best geometrical spot size; triangles: paraxial focus. Notice the difference in scale between this and the previous figure. The Airy disk is unity at the peak in both cases.

laterally with respect to the paraxial prediction. Figure 32 shows the PSF for one and two waves of coma.

Astigmatism. Astigmatism is a cylindrical deformation of the wave front that makes the PSF look like a short line. As we move through focus, we encounter two such focal lines, the sagittal and the tangential, perpendicular to each other. Halfway between them, the circle of least confusion is formed, which (although it does not resemble a circle) represents the minimum spread of the PSF and thus is taken as the best focus. Figure 33 shows the astigmatic line and the circle of least confusion for one astigmatic wave.

Image Quality Criteria Based on the PSF

Traditionally, the notion of image quality applies to incoherently illuminated systems, which represent the

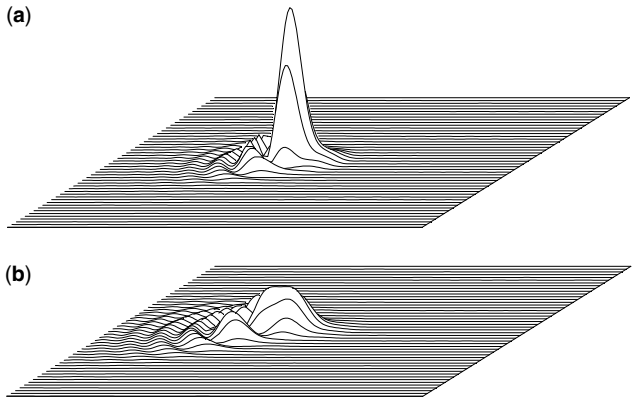


Figure 32. (**a**) Irradiance distribution in the PSF in the presence of one wave of coma. The peak value of the pattern is about 0.53 (relative to an Airy disk of unity peak). (**b**) Two waves of coma.

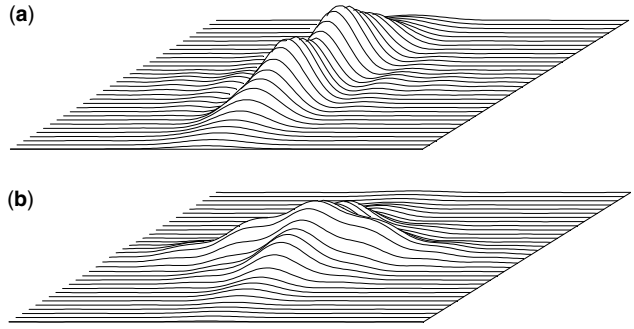


Figure 33. (**a**) Irradiance distribution in the PSF of a system suffering from one wave of astigmatism, at one of the two astigmatic foci (sagittal or tangential), where the image resembles a line. (**b**) A symmetrical image is obtained between the sagittal and tangential foci.

vast majority of image forming systems. It would be possible to devise modifications for coherent systems, but no particular need for doing so has arisen.

The Strehl Ratio. The Strehl ratio I is the ratio composed of the peak value of the PSF divided by the peak value of the PSF of an unaberrated system that has the same aperture (26). This definition is valid when the PSF has a well-defined, dominant peak, a condition that does not necessarily hold in the presence of a substantial amount of aberration (e.g., Fig. 33). Thus the Strehl ratio is primarily useful for well-corrected systems. The Strehl ratio can be read from Figs. 27, 29, and 30 as the maximum PSF value. An alternative definition of the Strehl ratio is the PSF value at the centroid. This has the advantage of retaining validity for any aberrational condition, although the scale of a PSF plot is best appreciated from its peak value. It can be shown that the two definitions lead to the same result under the appropriate conditions, which are small aberrations balanced by the appropriate amount of longitudinal or transverse focal shift.

A system that has a Strehl ratio $I \geq 0.8$ is considered practically diffraction-limited. If the system is free of aberration but is somewhat out of focus, the value of $I = 0.8$ corresponds to the *Rayleigh quarter-wave criterion*, according to which a wave-front error of less than a quarter of a wavelength is negligible. In other words, a wave-front focus error of $W_{020} = \lambda/4$ results in a PSF where $I = 0.8$ (see also Fig. 27). For other aberrations however, the permissible amount differs from a quarter wave. These values are given in Table 2 for the three primary aberrations [justification may be found in (27].

Related to the Strehl ratio is another often used quantity, the wave-front variance, which is also (inappropriately but commonly) called rms (root-mean-square) aberration. For any one field position, the wave-front variance is given by

$$E = \overline{(W^2)} - (\overline{W})^2, \tag{46}$$

where the average refers to integration across the pupil area. Maréchal examined in detail the relationship between the Strehl ratio and the wave-front variance (28). For values of $I > \sim 0.5$, there is a one-to-one relationship between the wave-front variance and the Strehl ratio (29). However, for lower values, where the PSF may not have a well-defined single maximum, its peak value is no longer relevant. In such a case, the energy spread may be characterized by the encircled energy, as in the next section. The wave-front variance remains a useful way of specifying the departure of the wave front from sphericity, especially because it can be readily measured by interferometric techniques. However, it is not necessarily a good way of specifying image quality for all applications. For example, if the final detector is the human eye, it is known that the wave-front variance is poorly correlated to subjective image quality (30).

Table 2. Strehl Tolerances for Primary Aberrations

Aberration	Symbol	Value for $I = 0.8$ (in waves)	Comment
Defocus	W_{020}	0.25	
Spherical	W_{040}	0.95	At best focus ($W_{020} = -W_{040}$)
Coma	W_{131}	0.62	
Astigmatism	W_{222}	0.36	At best focus ($W_{020} = -W_{222}/2$)

Encircled Energy. There is no single criterion akin to the Strehl ratio for encircled energy. Although such criteria could be devised (e.g., the radius encircling a fixed amount of energy or the amount of energy inside a fixed radius), no particular need for doing so has arisen. Encircled energy becomes important in the presence of a photodetector of limited area or a photodetector array such as a CCD. Then, it is important to know how much of the PSF energy is contained in the pixel. Because pixels are typically square, we may also speak of "ensquared" energy. Figure 34 compares the encircled energy between the PSFs of Fig. 29 (diffraction-limited system with and without 12% central obscuration). Various cases have been examined by Mahajan (31).

Line- and Edge-Spread Functions

Line-Spread Function. The line-spread function (LSF) is the irradiance (incoherent) or the complex amplitude (coherent) distribution in the image of an ideal line object. Image formation can be described in terms of the LSF for the orientation perpendicular to the line. In other words, we can write the irradiance distribution in any direction in the image as a convolution of the irradiance distribution in the object (in the same direction) and the LSF. The LSF has two advantages; it reduces all two-dimensional integrals to one-dimensional ones, and it is generally easier to measure the irradiance distribution in the image of a line than in the image of a point. The LSF is obtained from the PSF by simply integrating in the

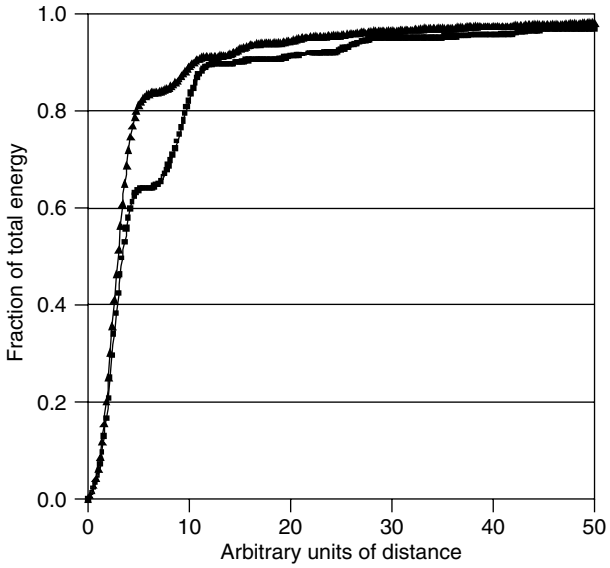


Figure 34. Encircled energy comparison between an obscured (squares) and unobscured (triangles) system, corresponding to the PSFs of Fig. 29.

direction of the line:

$$\text{LSF} \equiv G(u) = \int_{-\infty}^{+\infty} G(u, v)\, dv \tag{47}$$

In the incoherent case and for a system that has no aberrations and a uniformly illuminated circular aperture, the LSF is given by integrating Eq. (34), expressed in rectangular coordinates. The form of the function is given in Fig. 35. The behavior of the LSF in the presence of aberrations can also be readily predicted numerically. Aberrational tolerance theory based on the LSF has been developed (32).

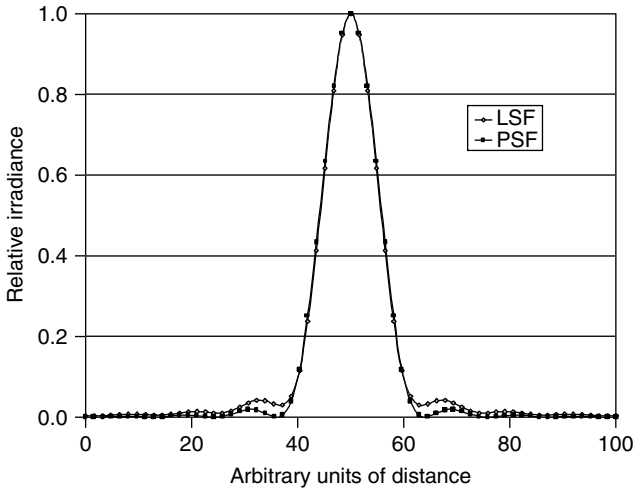


Figure 35. Incoherent LSF (diamonds) for a diffraction-limited system that has a circular pupil, compared with the corresponding PSF (squares).

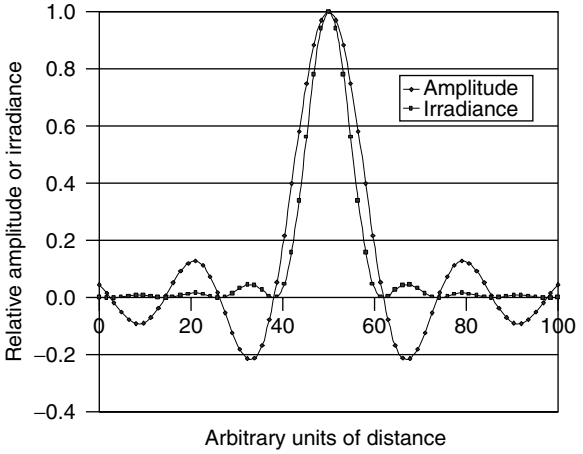


Figure 36. Amplitude (diamonds) and irradiance (squares) distribution in the image of the coherent LSF of a system that has no aberrations and a circular pupil.

If the line object is coherently and cophasally illuminated, the complex amplitude distribution in the image is given simply by

$$A(u) = (\text{const})\frac{\sin(2\pi u)}{\pi u} \tag{48}$$

and the corresponding irradiance distribution by the square of this function. The two functions are shown in Fig. 36. The irradiance distribution in the coherent LSF has zeros, unlike the incoherent case. The location of the first zero is given by $u = 0.5$, or, for nonnormalized coordinates (Eq. 44),

$$\xi = \frac{0.5\lambda}{n \sin a} \tag{49}$$

Edge-Spread Function. The edge-spread function (ESF) is the irradiance (incoherent) or complex amplitude (coherent) distribution in the image of an edge. In the incoherent case, there is no closed-form expression even for a diffraction-limited system. The ESF is obtained by convolving the LSF with a step function that represents the edge. Alternatively, we may express the LSF [$G(u)$] as the derivative of the ESF [$E(u)$]:

$$G(u) = \frac{dE(u)}{du}. \tag{50}$$

This formula is more useful in practice because the most common use of the ESF is to determine the LSF and through it, the modulation transfer function (MTF) see next section). The incoherent ESF is a monotonic function for any combination of aberrations. It is shown in Fig. 37.

For coherent case in the absence of aberrations, the complex amplitude distribution in the image of an edge is given by

$$A(u) = 0.5 + \frac{1}{\pi} Si(2\pi u) \tag{51}$$

(assuming cophasal illumination), where $Si(z) = \int_0^z \sin x/x\, dx$, the so-called Fresnel sine integral. The

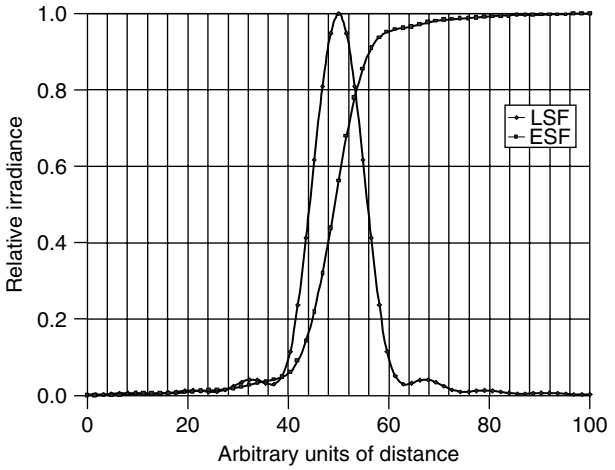


Figure 37. Incoherent ESF (squares) for a system that has no aberration and a circular pupil, compared with the corresponding LSF (diamonds).

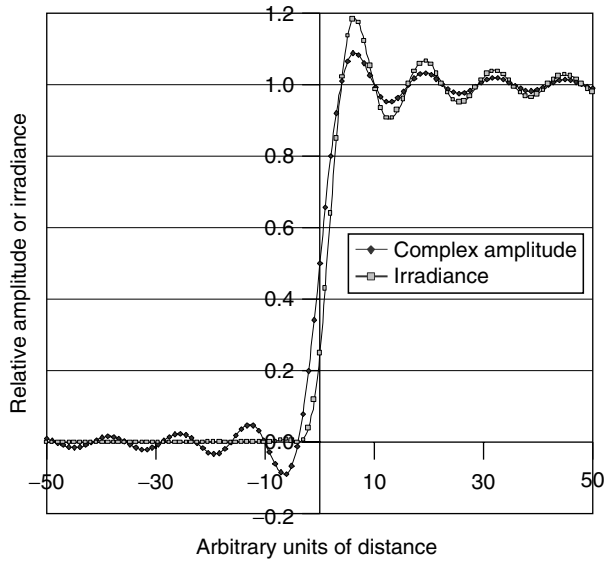


Figure 38. Amplitude (diamonds) and irradiance (squares) distribution in the coherent image of an edge or a system that has no aberrations and a circular pupil.

corresponding irradiance pattern, $B(u) = |A(u)|^2$, is shown in Fig. 38. It can be seen that the coherent image of an edge shows ringing, unlike the incoherent ESF. The maximum value of the ESF is 1.18 (if the geometric edge is unity). This maximum value occurs at $u = 0.5$, or $\xi = 0.5\lambda/(n \sin \alpha)$.

IMAGE FORMATION IN THE FREQUENCY DOMAIN (FILTERING). THE OPTICAL TRANSFER FUNCTION

The necessary mathematical background for this section is summarized in the Appendix. A reader unfamiliar with the basics of Fourier transforms will need to consult the Appendix before proceeding further.

The Image as a Fourier Transform

Any periodic function $f(x)$ that satisfies certain simple conditions (see Appendix) can be written as a summation of sinusoidal functions whose frequencies are equal to integral multiples of the frequency of the original function. Thus an image (or irradiance distribution) can be analyzed in terms of its harmonic (Fourier) components, or the components can be added (synthesized) to reconstruct the image.

Fourier synthesis is illustrated by an example in Fig. 39, where we show a periodic function (sawtooth wave) and the function that results from summing the first four harmonic components (whose periods are p, $p/2$, $p/3$, $p/4$). A remarkably good representation of the function is obtained, even with only four harmonic components.

The (ir)radiance distribution that corresponds to the sawtooth function of Fig. 39 is shown in Fig. 40. A sinusoidal irradiance distribution, whose period is the same as the sawtooth, is shown in Fig. 41. It is the addition of sinusoidal patterns (or gratings) similar to this that can be used to synthesize the distribution of Fig. 40, but there is a subtlety that must be appreciated.

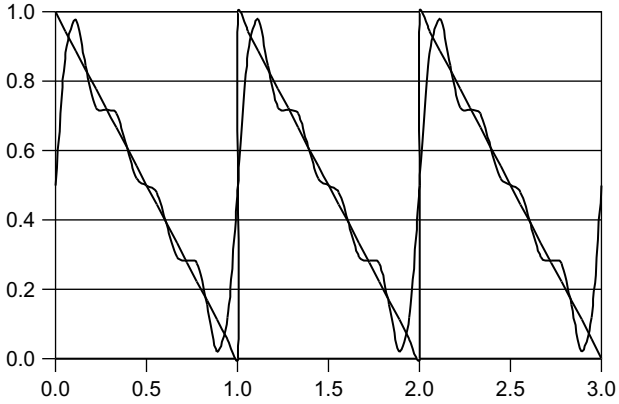


Figure 39. A sawtooth function whose slope is -1 and period 1, and its harmonic approximation, limited to the first four Fourier coefficients: $f(x) \cong 1/2 + 1/\pi \sum_{n=1}^{4} 1/n \sin(2\pi n)$.

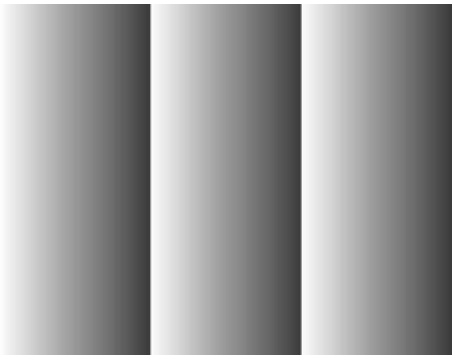


Figure 40. A sawtooth irradiance distribution. This function may appear to differ visually from a sawtooth near the discontinuity due to the property of the human eye that exaggerates luminance discontinuities, as well as imperfections in the printing process.

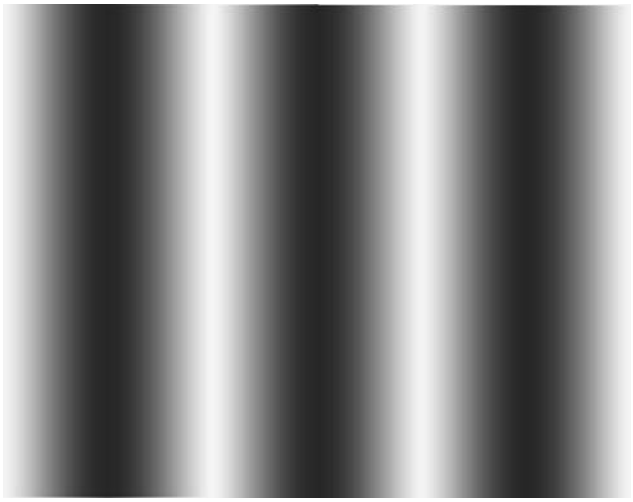


Figure 41. An approximately sinusoidal irradiance distribution.

A function that represents a physically plausible (ir)radiance distribution cannot contain negative values. The full Fourier series expansion of the sawtooth function of Fig. 39 is given by [see Eqs. (A1)–(A4)]

$$f(x) = \frac{1}{2} + \frac{1}{\pi}\sum_{n=1}^{\infty}\frac{1}{n}\sin\left(2\pi\frac{n}{p}x\right), \tag{52}$$

where the coefficients of the cosine terms are all zero. It will be seen that the sinusoidal components contain negative values, unlike the irradiance distribution of Fig. 41. Thus, the harmonic components of an image are not necessarily themselves true images (or physically possible irradiance distributions). In the example of Eq. (52), the addition of the constant term 1/2 makes $f(x)$ overall positive, but this is not the same as adding a series of wholly positive sinusoids. Such a series would be represented as $\sum_{n=1}^{\infty}[1/n + 1/n\sin(2\pi(n/p)x]$ because the minimum constant that can be added to a sinusoid to make it positive is its own amplitude. This last series would give the same result as Eq. (52) if the series $\sum_{n=1}^{\infty}(1/n)$ converged to $\pi/2$. However, this series does not converge; hence, the addition of fully positive sinusoids does not reproduce $f(x)$. The fact that there is no physically negative (ir)radiance[1] ultimately does not affect the utility of Fourier analysis in describing an image, nor does it diminish the significance of sinusoidal (ir)radiance distributions as test objects that have special properties.

Now, we wish to consider the image in two dimensions and extend the description to nonperiodic functions. Before proceeding further, the variables x and y used in the Appendix will be replaced by u and v to keep a consistent notation with the section "Image Formation in the Spatial Domain," where the image coordinates were (u, v) and the pupil coordinates were (x, y). If $f(u, v)$ is an irradiance distribution, it may be written as

$$f(u,v) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} F(\sigma,\tau)\exp[i2\pi(\sigma u + \tau v)]\,d\sigma\,d\tau, \tag{53}$$

where

$$F(\sigma,\tau) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f(u,v)\exp[-i2\pi(\sigma u + \tau v)]\,du\,dv. \tag{54}$$

The last two equations show that we can consider an image as the summation (or integral) of an infinite number of sinusoidal components of various periods (frequencies) and orientations.

Fourier analysis provides us with an alternative way of thinking about an image: instead of being concerned about its visible features, we are concerned about its spatial frequency content. This shift of viewpoint proves to be of great utility. For example, one may perform operations that directly affect the spatial frequency spectrum of an image; some examples are given in the following sections.

The Optical Transfer Function

Definitions

The fundamental significance of the sinusoidal distribution in image formation is that it remains sinusoidal when transmitted by an optical system, except for amplitude and phase. This is a cornerstone fact of image formation, and it is an immediate consequence of the linearity property of optical systems. Linearity means that no higher harmonics are generated by the system.

When this fact is coupled with the Fourier method of describing an image, the general problem of image formation is reduced to the problem of imaging a sinusoid (or better, a set of sinusoids of various frequencies and orientations), just as in the section "Image Formation in the Spatial Domain," the problem of image formation was reduced to determining the image of a point. And although substituting the image of a point by that of many sinusoids might appear a more laborious way of describing image formation, it ultimately leads to a simpler way of describing the action of an optical system — or, if not simpler, at least complementary and of equal significance.

The concept of the transfer function is perhaps more familiar in audio engineering. Audio amplifiers are designed to have a flat frequency response over the 20 Hz to 20 kHz band. This means that they must transmit all pure tones (sinusoids) across that band in undiminished (or uniformly amplified) amplitude, independent of frequency. Similarly, they are not supposed to introduce any phase shift, and harmonic distortion (extraneous frequencies not present in the original signal) must be kept very low. An incoherently illuminated optical

[1] There are certain special cases, for example, the phenomenon of *lateral inhibition* in vision or of *adjacency* in photography, where the response of the recording medium can be described through a point-spread function that has negative irradiance. The physical interpretation of such negative irradiance is that the excitation of a point reduces the response of neighboring points. This exception has no effect on this discussion, and in any case, it is associated with recording media (retina, film) rather than purely optical systems.

system can be thought of as an amplifier with a gain ≤ 1 that is inherently incapable of introducing harmonic distortion but may introduce a nonlinear, frequency-dependent phase shift.

Let

$$B(u, v) = a + b\cos[2\pi(\sigma u + \tau v)] \tag{55}$$

be a cosinusoidal radiance distribution (grating) or object. We assume incoherent illumination for now. The period d and the orientation θ of this cosinusoidal grating are given by

$$\begin{aligned} d &= (\sigma^2 + \tau^2)^{-1/2}, \\ \theta &= \arctan\left(\frac{\tau}{\sigma}\right) \end{aligned} \tag{56}$$

where the first of these two relationships arises simply from Pythagoras' theorem. We also define the *contrast* or *modulation* of the grating as the ratio b/a.[2] Now, apply the convolution Eq. (40) to determine the image of this grating. If $G(u', v')$ is the system PSF, the irradiance distribution in the image will be (from Eq. 40)

$$\begin{aligned} B'(u', v') = \iint_{[}^{a} &+b\cos(2\pi(\sigma u + \tau v))] \\ &\times G(u' - u, v' - v)\, du\, dv, \end{aligned} \tag{57}$$

where it is understood that the integration extends across the image area (or the area across which the PSF remains stationary). Performing the integration and after appropriate substitutions of variables that are not detailed here, we obtain the following result:

$$\begin{aligned} B'(u', v') = \text{const.}\{&a + bM(\sigma, \tau)\cos[2\pi(\sigma u' + \tau v') \\ &+ \phi(\sigma, \tau)]\} \end{aligned} \tag{58}$$

This equation shows that the irradiance distribution is also cosinusoidal and has the same frequency and orientation as the object, but has different contrast (bM/a) and a different phase (ϕ). Notice that to derive Eq. (58), we only made the assumptions of linearity and stationarity that are inherent in the convolution Eq. (57). No additional assumptions were needed; hence, the fact that a sinusoid is imaged as a sinusoid is inherent in those two assumptions.

The reader may object at this point that the image of the sinusoid cannot be another sinusoid of the same period unless the system has unity magnification. This is true of course, but Eqs. (57) and (58) are strictly correct as written; the normalized coordinates we are using [defined through Eq. (44)] actually account for the magnification because they contain the numerical aperture [see also Eqs. (10) and (24)]. Thus, when we say that the image is a sinusoid of the same frequency, we mean the same frequency in the normalized coordinate space. In real coordinate space, the grating period scales according to the magnification. When numerical values for the frequency are quoted in a system that has nonunity magnification, they usually refer to the image space.

The functions M and ϕ in Eq. (58) are the modulus and phase of a single function:

$$D(\sigma, \tau) = M(\sigma, \tau)\exp[i\phi(\sigma, \tau], \tag{59}$$

where $D(\sigma, \tau)$ is given (from the derivation of Eq. 58) as

$$D(\sigma, \tau) = \frac{\iint_{G}^{(} u, v)\exp[-i2\pi(\sigma u + \tau v)]\, du\, dv}{\iint_{G}^{(} u, v)\, du\, dv}, \tag{60}$$

where the limits of integration are the same as in Eq. (57). This last equation may explain how $G(u, v)$ seemed to have disappeared in going from Eq. (57) to (58). It was in fact absorbed in $D(\sigma, \tau)$.

The function $M(\sigma, \tau)$ is the ratio of the image contrast or modulation (bM/a) to the object contrast (b/a). It is called the *modulation transfer function* (MTF). The function $\phi(\sigma, \tau)$ is the phase shift between image and object. It is called the *phase transfer function* (PTF). The function $D(\sigma, \tau)$, which has the MTF as its modulus and the PTF as its phase, is called the *optical transfer function* (OTF). Under the assumptions of linearity and stationarity, the OTF fully characterizes image formation by an optical system. Eq. (60) shows *that the OTF is given by the normalized (inverse) Fourier transform of the PSF* $G(u, v)$.

The two ways of describing image formation (from the PSF and from the OTF) are intimately related. This may first be appreciated from Eq. (60) that related the PSF and OTF. The convolution theorem (Eqs. (A25), (A26) in the Appendix) also states that if a function can be written as the convolution of two others, its Fourier transform is given by the product of the Fourier transforms of the two functions. From the convolution Eq. (40) that relates the (ir)radiance distribution in the object and the image through the PSF, we obtain the following relationship:

$$b'(\sigma, \tau) = b(\sigma, \tau)g(\sigma, \tau), \tag{61}$$

where $g(\sigma, \tau)$ is the nonnormalized version of the OTF. This equation states that the Fourier (spatial frequency) spectrum of the image $[b'(\sigma, \tau)]$ is given by the Fourier spectrum of the object, multiplied by the OTF. Thus the OTF acts as a spatial frequency filter on the frequency spectrum of the object.

Calculation of the OTF. The steps for calculating the OTF have already been outlined. First, the pupil function $f(x, y)$ must be established, which involves computing (by ray tracing) the wave-front aberration $W(x, y)$ as well as the pupil transmittance [Eq. (42)]. Next, the Fourier transform of the pupil function gives the complex amplitude distribution $F(u, v)$ in the image of a point [Eq. (43)]. The squared modulus of F is the incoherent PSF, $G(u, v)$ [Eq. (45)]. Finally, the

[2] If both a and b are positive and $b \geq a$, then the radiance distribution B is physically possible, that is, it contains no negative values. However, the derivation that follows is independent of that assumption.

(normalized) inverse Fourier transform of the PSF gives the OTF [Eq. (60)].

This method of computing the OTF is occasionally employed because of the availability of fast Fourier transform (FFT) numerical routines that speed up the calculations considerably. However, there is another, more direct method of computation, which is preferable in many ways (33). This alternative is based on a mathematical theorem that allows us to relate the OTF directly to the pupil function. The theorem may be stated as follows: If a function $A(x, y)$ is the Fourier transform of another function $a(u, v)$, then the autocorrelation of $a(u, v)$ is given by the inverse Fourier transform of the squared modulus of $A(x, y)$ (see also Appendix). In this case, the OTF is the inverse Fourier transform of the squared modulus of the complex amplitude distribution in the image, which in turn is the Fourier transform of the pupil function. Hence, we may write the OTF directly as the autocorrelation of the pupil function:

$$D(\sigma, \tau) = \frac{1}{A} \iint_{f} x, y) f^*(x - \sigma, y - \tau)\, dx\, dy \tag{62}$$

where A is the pupil area and * denotes complex conjugate. The limits of integration can be considered infinite, but in practice the integral is nonzero only over the area of overlap of $f(x, y)$ and the shifted function $f^*(x - \sigma, y - \tau)$. In writing this equation, it would appear that the physical units are wrong if we are subtracting σ (spatial frequency or inverse length) from x (length). It is important to understand that all variables are normalized to be unitless [see Eq. (44) and associated discussion]. Without the appropriate normalization, all of these equations expressing the PSF, the OTF and the Fourier transformations become a lot more cumbersome.

Some simple examples of the OTF are given below, for zero aberrations and unity pupil transmittance, that is, $f(x, y) = 1$ inside the pupil and zero outside. In that case, the OTF depends solely on the shape of the pupil. Along the axis of a rotationally symmetrical optical system, the pupil shape is typically circular, but off-axis it may be elliptical or more complicated. Of course, it is also possible to have a noncircular aperture stop that would then define the pupil shape. Because there is no phase term, the OTF is the same as the MTF, and the integral, Eq. (62), simplifies to

$$D(\sigma, \tau) = M(\sigma, \tau) = \frac{1}{A} \iint_{S} dx\, dy = \frac{S(\sigma, \tau)}{A} \tag{63}$$

where S is the overlap area between the pupil function and its shifted version. For an example that provides the simplest analytical expression, suppose that the system has a square pupil shape, as in Fig. 42. The pupil area A is 4, and the MTF (S/A) is given by $1 - \sigma/2$. Thus, the MTF is linear with σ and reaches zero at $\sigma = 2$. Clearly, the autocorrelation can proceed in directions other than the x or y axes, for example, along the diagonal. In the latter case, the cutoff frequency would be $2\sqrt{2}$. The MTF in the x direction is shown in Fig. 43.

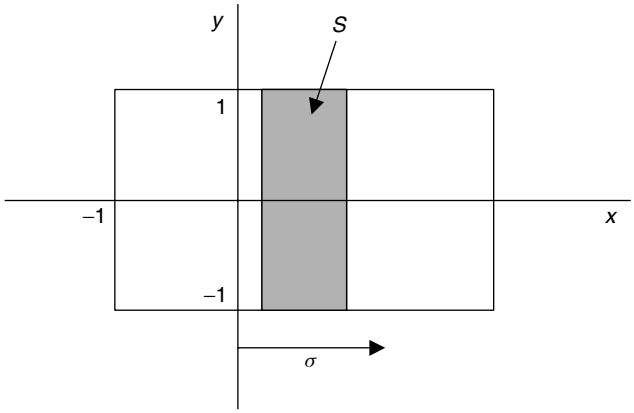


Figure 42. Illustration of the autocorrelation of a square pupil in the x direction. The area of overlap S is shown shaded.

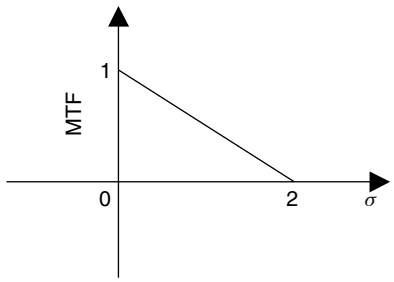


Figure 43. MTF (or OTF) shape for a system that has a square pupil and no aberration, corresponding to the autocorrelation operation of Fig. 42.

In the more common circular pupil of unit radius, the overlap area between two circles has a somewhat more complicated analytical expression. The MTF in this case is given by

$$D(\sigma) = M(\sigma) = \frac{4}{\pi} \left\{ \frac{1}{2} \arccos\left(\frac{\sigma}{2}\right) - \frac{\sigma}{4} \sin\left[\arccos\left(\frac{\sigma}{2}\right)\right] \right\}, \tag{64}$$

and the MTF is the same in any direction. It is shown in Fig. 44.

The cutoff frequency is $\sigma = 2$ because of the unit radius pupil normalization scheme. In other words, the pupil overlap goes to zero when the shifted function has moved by two radii. By inverting Eq. (44), we can obtain the denormalized spatial frequency N as

$$N = \frac{(n \sin a)\sigma}{\lambda}, \tag{65}$$

where all quantities refer to the image space.

The OTF in the Presence of Aberrations. In the presence of aberrations, the MTF and the OTF are no longer the same. The simplest phase term or aberration is defocus. The effect of defocus on the OTF was determined analytically by Hopkins (34), although nowadays similar computations are always performed numerically. Like the PSF, the OTF in the presence of defocus is a wholly

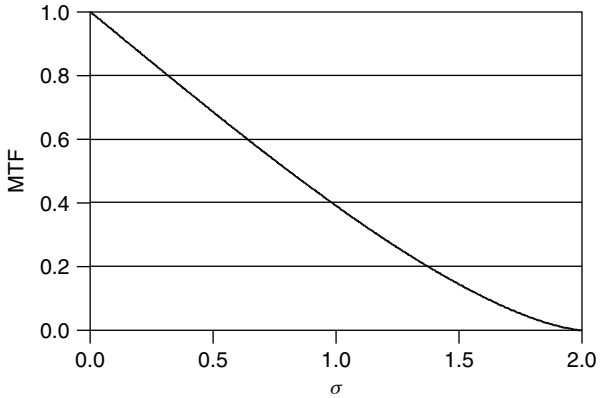


Figure 44. MTF (or OTF) for a system that has a circular pupil and no aberrations. The normalized spatial frequency σ relates to the nonnormalized frequency N through $\sigma = \lambda N/(n \sin a)$, where $n \sin a$ is the numerical aperture in the image space.

real function, but may contain negative values. This is illustrated in Fig. 45. The negative values imply a phase shift of 180° which produces a contrast reversal in practice; the bright grating bars become dark and vice versa. This is depicted in Fig. 46 using a target that has two different spatial frequencies.

When the MTF goes to zero, the corresponding spatial frequency is not transmitted. This is a form of *incoherent spatial filtering*; certain frequencies are filtered out of the image, and others are modified in different ways. Such incoherent spatial filtering can also be performed by pupil apodization, that is, by modifying the pupil transmittance. For an extreme example, suppose that the pupil is obstructed by a mask that allows light to pass through only two squares as shown in Fig. 47. By performing the autocorrelation of the pupil function in the absence of aberrations [Eq. (63)] in the $+x$ direction, we obtain the MTF shape in Fig. 48 that shows an entire band of missing frequencies. A less extreme example is given in Fig. 49, where the MTF of a centrally obscured system such as an astronomical telescope is compared with the MTF of an unobscured system of the same aperture. It can

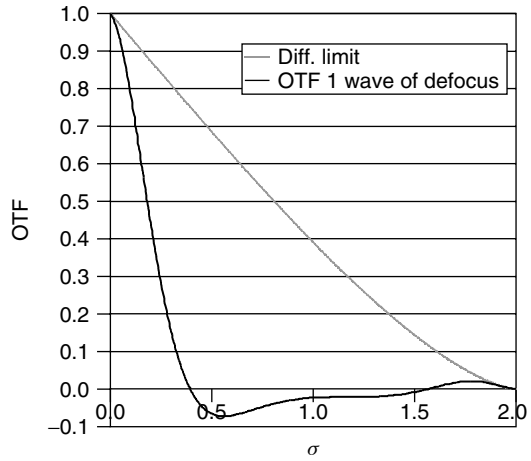


Figure 45. OTF for a system suffering from one wave of defocus, compared with the diffraction-limited case (top curve).

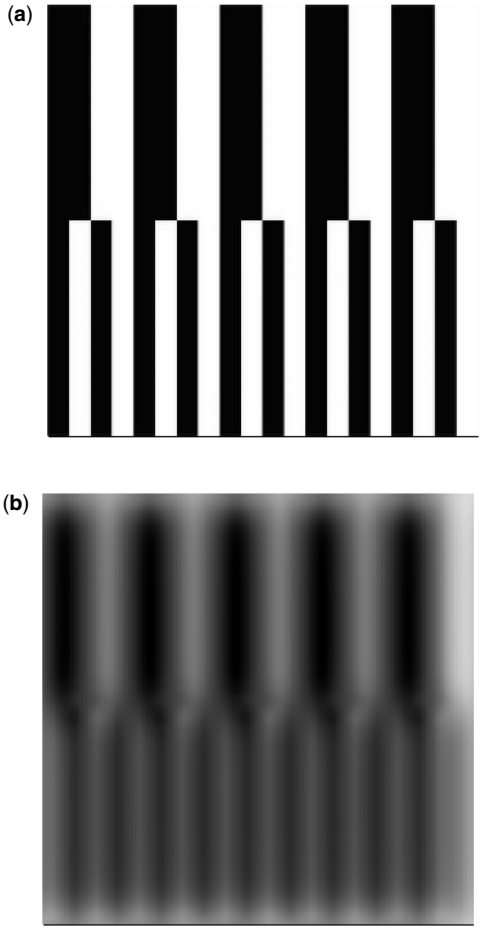


Figure 46. (**a**) A square wave object that has two different frequencies, such that the low frequency falls in the positive part of the MTF of Fig. 45, and the high frequency falls in the negative part. (**b**) The image of the square wave object through a system that has an OTF, as shown in Fig. 45. Notice that the bright and dark bars are reversed for the high-frequency grating (contrast reversal).

be seen that the effect of the obscuration is to suppress low frequencies and to enhance higher frequencies.

In the presence of aberrations other than defocus, the MTF can take a variety of shapes that are constrained by the following conditions (35):

1. The MTF has a maximum of unity at zero spatial frequency[3] (this says in effect that the best imaged object is no object at all; any contrast in the object will be attenuated).
2. At any given spatial frequency, the aberrated MTF is always lower than the unaberrated MTF of a system that has the same aperture.
3. The MTF slope at the origin is independent of aberration.

[3] Some nonoptical systems like photographic film can have an MTF that peaks at a frequency >0. This is intimately related to the so-called adjacency or inhibition effects, mentioned previously.

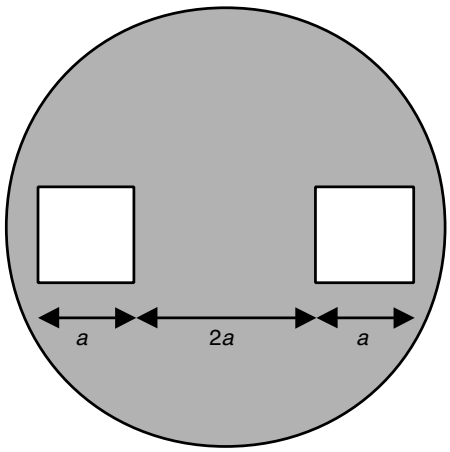


Figure 47. An apodized pupil that has two square transmitting areas.

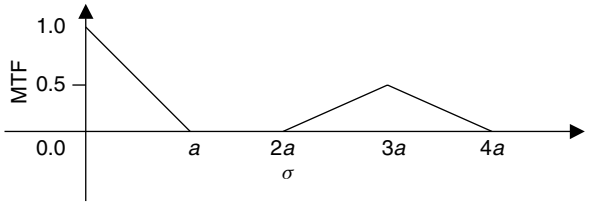


Figure 48. MTF (or OTF) in the x direction of a system that has a pupil shape such as shown in Fig. 47.

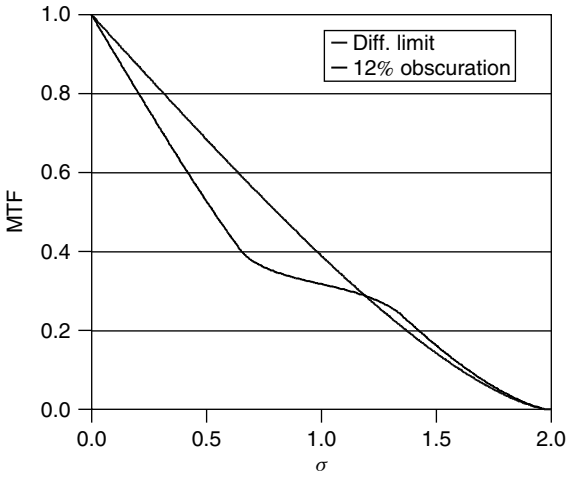


Figure 49. Comparison between obscured and unobscured diffraction-limited MTFs, corresponding to the PSFs of Fig. 29.

The MTF is a useful tool, but it is not always used as carefully as it requires. If the MTF is to be used successfully, one must keep in mind the following:

1. The MTF is a two-dimensional function. When shown as one-dimensional, one must always be aware that it represents only one object (grating) orientation unless the system is either diffraction-limited or suffers from spherical aberration but no other asymmetrical aberrations.
2. Several aberrations like spherical or astigmatism and their higher order counterparts interact with defocus in a complicated way. In the presence of such aberrations, the optimum focal plane is not universally defined. A different choice of focus will yield a different MTF. In general, the effect of aberrations on the MTF is to suppress certain frequency bands more than other bands, as a function of focus. Thus it is often impossible to define a "best" MTF for an optical system suffering from significant aberration, unless a specific criterion for choosing the best MTF is applied.
3. It is dangerous to use only the MTF and neglect the PTF because the latter may cause significant image distortion. If the PTF is zero or linear, then it has no effect on the image. In the linear case, all sinusoidal components are shifted by an amount proportional to their frequency, resulting in a bodily image shift but no distortion because all sinusoidal components still add in phase. The effect of the nonlinear part of the PSF on image quality has been examined by Hopkins (36). Asymmetrical aberrations like coma have the greatest effect on the nonlinear components of the PTF. Examples have been shown by Hopkins (36) and Mahajan (10). Finally, the PTF may also be neglected when the MTF is high, as, for example, at low frequencies. If an optical system is followed by a sampling detector that cannot reproduce the high frequencies transmitted by the optics, then only the low frequencies matter, and in such a case the PTF may be entirely negligible.

The OTF can be measured in a number of ways, and a considerable body of literature exists on this topic. In brief, we may state that although the OTF can be computed from a measurement of the PSF, such a measurement is not always feasible due to the small amount of light that passes through a pinhole (although the introduction of low-noise CCD detector arrays has facilitated such measurements). More often, the LSF is used, whose Fourier transform yields the OTF in the direction perpendicular to the line. The procedure may then be repeated for additional azimuths. When employing this process, a correction must generally be applied for the finite width of the object slit, and for the partial coherence in the slit illumination (37). An edge may also be used as an object, although the differentiation needed to derive the LSF (and hence the MTF) tends to increase the computational error. Interferometric techniques can provide a direct representation of the pupil function and hence, the OTF by autocorrelation. A detailed account of techniques and methods has been provided by Williams (38).

The frequency response techniques outlined here can be generalized to handle optical systems and also the detector of the optical radiation or even the transmitting medium, such as the atmosphere. So, for example, it is possible to model a complete Earth-orbiting remote sensing system by taking the product of the transfer functions of the atmosphere, the optics, and the detector (such transfer functions are discussed in other articles

in this encyclopedia). The multiplication of the transfer functions is valid if the effects of the various system components are decoupled. This means, for example, that the atmosphere cannot correct the effects of the lens aberrations, or that the lens aberrations cannot correct the effect of the detector. This condition is normally satisfied for physically different systems such as a detector and a lens, but it is not satisfied for two cascaded lenses because the aberrations of the first may cancel the aberrations of the second. Two lenses in cascade may be decoupled if the image of the first lens is formed on a diffusing screen that is then reimaged by the second lens. In such a case, the transfer functions of the two lenses may be multiplied because the diffusing screen destroys the coupling between the lenses and does not permit the second lens to correct the aberrations of the first.

Coherent Image Formation

Fundamentals. Coherent image formation has several subtleties as well as conceptual differences from the incoherent case. A detailed treatment of coherent image formation is outside the scope of this article because the entire subject holds, for now, relatively little interest outside the field of optics. Thus we will only hint at the subtleties and broadly describe the conceptual differences and similarities.

The first broad similarity between incoherent and coherent imaging is that the latter can also be described as a convolution with the coherent PSF, as shown by Eq. (41). The linearity condition is satisfied for a coherently illuminated system, and the stationarity condition merits some more discussion later. Equation (41) relates the image and object complex amplitude distributions through a convolution with the PSF. The convolution theorem of the Appendix then tells us that the Fourier transforms of those functions will be related through a mere multiplication. Specifically,

$$a'(x', y') = a(x, y)f(x, y), \tag{66}$$

where $f(x, y)$, the Fourier transform of the PSF is the pupil function, as we have established previously [Eq. (42)]. Therefore, Eq. (66) tells us that the Fourier (frequency) spectrum of the image is given by the Fourier spectrum of the object multiplied by the pupil function, which plays the role of the coherent transfer function (CTF). Thus, for example, in the absence of aberrations and apodization, the CTF is the simple top-hat function of Fig. 23, for a system that has a circular pupil shape.

In summary, then, coherent image formation can be described as follows: From the complex amplitude distribution in the object $A(u, v)$, we pass to the amplitude distribution across the entrance pupil $a(x, y)$ through a Fourier transformation. Multiplication of $a(x, y)$ by the pupil function $f(x, y)$ gives the amplitude distribution across the exit pupil (a'). Finally, an (inverse) Fourier transformation gives the complex amplitude distribution in the image (A').

Based on this description, we can predict the images of several typical objects. Consider first a low-frequency object, whose spatial frequency spectrum does not extend outside the pupil radius (taken as unity). This means that $a(x, y) = 0$ for any $x^2 + y^2 > 1$. If the pupil function contains no apodizing term or aberrations, all spatial frequencies inside the pupil will be transmitted with undiminished amplitude, and the image will be a perfect reconstruction of the object. This contrasts sharply with the incoherent case, where the only perfectly imaged object was no object at all (zero spatial frequency).

The images of objects that contain frequencies that fall outside the pupil will appear low-pass filtered. Take, for example, a sawtooth function similar to that of Fig. 39, where the abscissa is understood as amplitude rather than irradiance. If the extent of the pupil allows only the first four terms to pass, then the resulting image will have an *amplitude* distribution similar to that shown by synthesizing the first four harmonics. However, a photodetector such as the eye records only the irradiance and not the amplitude of the wave. Therefore, the final appearance of the image would correspond to the square of the function that arises from the superposition of the first four harmonics. Mills and Thompson (39) described in detail the effect of aberrations and pupil apodization in the images of coherently illuminated systems for standard test objects, including a point, slit, edge, and two adjacent points.

It is common to compare coherent and incoherent transfer functions, as shown in Fig. 50. However, this figure, a version of which appears in several texts, is conceptually misleading because the spatial frequency axis refers to two different physical quantities, amplitude and irradiance. A sinusoidal amplitude distribution (grating), when squared, will yield an irradiance distribution of twice the frequency [through the trigonometric identity $\cos^2 x = 1/2 + 1/2\cos(2x)$]. Thus the sinusoidal grating that corresponds to the cutoff point of the CTF would appear visually as a grating of the same frequency as the cutoff point of the incoherent OTF.

It is important to understand in more detail the nature of the "object" in coherent imaging. Up to this point, we have assumed that the object is a spatial

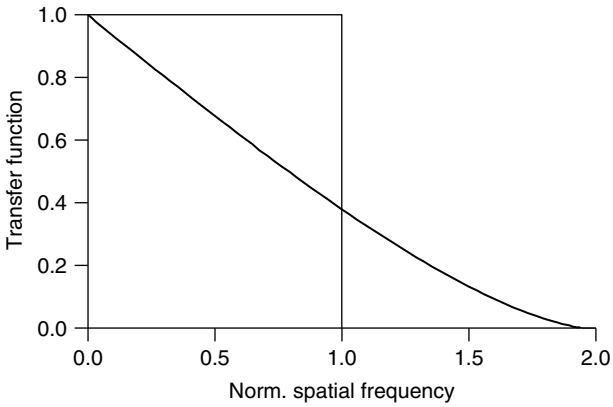


Figure 50. A common but somewhat misleading comparison between the diffraction-limited coherent and incoherent transfer functions for a system that has a circular pupil. In the coherent case, all frequencies up to unity (the pupil radius) pass through without attenuation. However, the spatial frequency axis refers to different types of objects in the two cases (amplitude or irradiance gratings).

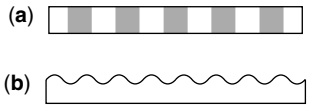


Figure 51. Schematic of an amplitude (**a**) and a phase object (**b**). The latter is like a perfectly transparent piece of glass that has a surface relief pattern as shown.

distribution of complex amplitude, without giving any further details. Such a complex amplitude distribution will have the general form $f(x,y) = U(x,y)\exp[i\phi(x,y)]$. In this expression, U is the real amplitude (transmittance or reflectivity) variation, and ϕ is the variation in optical thickness or phase (Fig. 51). Physically, the object may be taken as an illuminated transparency. The illuminating wave itself may have amplitude and phase variations and is thus represented by a function $f_i(x,y) = U_i(x,y)\exp[i\phi_i(x,y)]$. Under the assumption that the spatial detail in all four of these functions is considerably coarser than the wavelength, the complex amplitude distribution emanating from the object plane is given by the product of the two distributions

$$f_o(x,y) \equiv f_i(x,y)f(x,y) = U_i(x,y)U(x,y)\exp\{i[\phi_i(x,y) + \phi(x,y)]\} \equiv U_o(x,y)\exp[i\phi_o(x,y)] \tag{67}$$

Details in the object comparable to the wavelength are generally of interest for their diffractive, rather than imaging properties (in any case, high frequencies are likely to be cut off by the system aperture, as Fig. 50 indicates). Thus the term "object," as used so far, implies the complex amplitude distribution $f_o(x,y)$ that contains the characteristics of the object and of the illuminating wave.

Now, consider the question of stationarity of the amplitude PSF that was postponed previously. In addition to the requirement for the invariance of aberrations with field, as for incoherent imaging, we also need to ensure that there are no phase terms that vary with field and would not appear in the incoherent PSF. In the normal imaging situation for a planar object and image and finite pupil locations, there are two such terms. One is a linear obliquity term, and the other is a quadratic optical path variation term (40). Space prohibits further details, but it is noted that both of these terms can be eliminated if instead of a planar object and image surfaces, we consider spherical surfaces centered on the entrance and exit pupil, respectively. Then, the illuminating wave must also be a spherical wave focusing on the entrance pupil. Under these conditions, the illumination is cophasal, and the additional phase terms disappear, ensuring stationarity of the amplitude PSF and the validity of the convolution integral, Eq. (41). Of course, a planar object can also be handled if illuminated by a plane wave. This places the entrance pupil at infinity, and therefore the optical system must be appropriately designed.

An alternative way of handling planar objects is by matched illumination. In this case, the planar object is illuminated by a spherical wave focusing at the entrance pupil. Then, there is an exact Fourier transform relationship between the object distribution and the entrance pupil, provided the entrance pupil surface is taken as a sphere. (If the illuminating wave is not perfectly spherical, that is, it has aberrations, then the FT of the product of the object function and the aberration terms is obtained, as noted previously). If further the exit pupil and the image surfaces are also taken as spherical, then the stationarity requirement is satisfied exactly (40). However, many practical situations can be handled without concern about the difference between spherical and plane surfaces because the irradiance distribution is very similar in the two cases, and several operations can be performed on the irradiance alone.

Finally, we devote a little space to the intuitive understanding of Fourier analysis for a complex amplitude object. To recall the incoherent case, we saw that the fundamental building blocks of any incoherent image were sinusoidal irradiance distributions (albeit fictitious ones, that contain negative values). What are the corresponding building blocks of a complex amplitude distribution? The answer is that any complex amplitude distribution can be analyzed as a set of interfering plane waves at various directions of propagation. This is the so-called *angular spectrum*. This may be seen as follows: The Fourier transform relationship

$$a(x,y) = \iint_{A(u,v)}^{\exp} [i2\pi(ux+vy)]\,du\,dv \tag{68}$$

holds between the complex amplitude distribution in the entrance pupil $a(x,y)$ and the object $A(u,v)$. The inverse relationship,

$$A(u,v) = \iint_{a(x,y)\exp[-i2\pi(ux+vy)]}^{\cdot} dx\,dy \tag{69}$$

expresses the object distribution as a superposition of plane waves. The plane waves are represented by the exponential term and have a direction of propagation given by the direction cosines $\alpha = x\lambda$ and $\beta = y\lambda$ [see also (17)]. This is illustrated in Fig. 52, where we consider a simple, one-dimensional (coarse) diffraction grating, which is simply a one-dimensional periodic distribution of amplitude and/or phase. The grating is illuminated by a plane wave, as shown. It is a well-known result of wave optics [e.g., (12)] that after the grating, there will be plane waves emanating at various angles. Each such wave is called a grating order and corresponds to a harmonic term in the Fourier series of the grating profile [e.g., Eq. (A21) or (52) for a specific example]. This is the physical implementation of Eq. (68), which, however, refers to a general nonperiodic object. Now, if the direction of propagation is reversed, then the amplitude distribution at the plane of the grating is obtained as the superposition of the plane waves. This is the physical meaning of Eq. (69).

Coherent and incoherent illumination are two extreme cases. The general case is partially coherent illumination. Fortunately, most practical imaging problems and systems satisfy the coherence or incoherence conditions quite well.

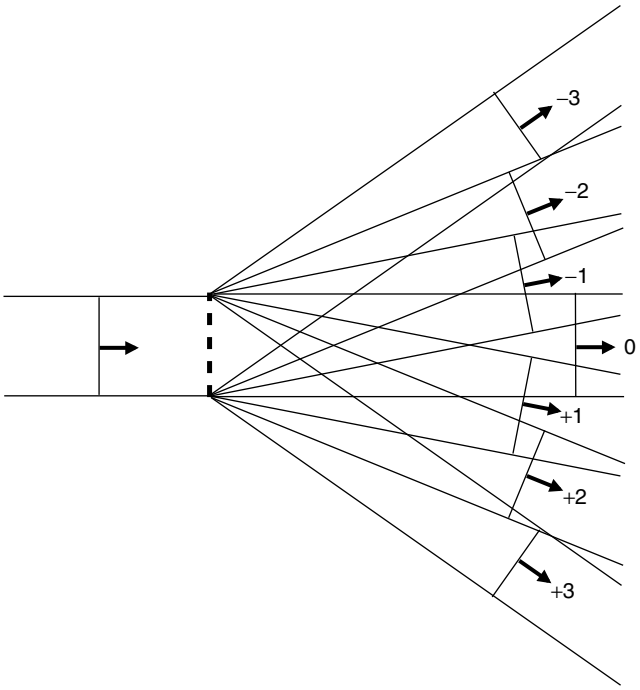


Figure 52. A plane wave incident on a periodic object (grating) and the resulting angular spectrum of diffracted plane waves, each corresponding to a grating order.

In general, partial coherence is significant in systems that incorporate illumination subassemblies, most commonly microscopes, or, more recently, photolithographic projection systems. Hopkins (40) showed that the problem of partially coherent imaging can be reduced to the superposition of the image intensities produced by waves emanating from each point of a properly defined "effective source" that simulates the partial coherence conditions across the illuminated object plane.

Spatial Filtering: An Example and Application of Coherent Imaging.

All of the previously discussed concepts of coherent imaging can be illustrated by the following example of a system, shown in Fig. 53. A laser beam is focused to a point and then expanded to illuminate the object. The system uses a plane wave to illuminate the planar object and therefore has the entrance pupil at infinity. If the image is to be obtained on a plane, then the exit pupil must also be at infinity. The plane noted as "filter plane" is also the image of the entrance and exit pupils (i.e., the aperture stop). The Fourier transform of the object amplitude distribution is formed at the entrance pupil by the angular spectrum of plane waves that emanate from the object. The lens immediately after the object will image the entrance pupil at the aperture stop, where a scaled version of the Fourier transform is obtained. The following lens performs the inverse Fourier transformation, and the image appears at the image plane.

If the lenses are well corrected for aberrations, then the Fourier transformations are exact and the stationarity conditions are satisfied across the entire image.

This system becomes more interesting if one places various filters at the filter plane. Then, amplitude leaving the filter plane will be the product of the incident and the filter transmission, as in Eq. (67). This allows us to perform operations in the Fourier domain and thus alter the appearance of the image in specified ways. To illustrate the process using an experiment of some historic importance, consider a square grid as an object (Fig. 54). The scaled Fourier transform of the object transmittance will appear at the filter plane and will have a form similar to that shown in Fig. 55. In a landmark experiment (although performed in a somewhat different setup), Porter (41) took a transparent strip as a filter that allowed only the vertical or horizontal orders to pass. In each case, the image turned into a vertical or horizontal one-dimensional grating (Fig. 56). Although the outcome of this experiment might sound obvious after the introduction of the Fourier transform and the imaging equations of the previous sections, it was performed at a time when none of these concepts were in place and led to a revolutionary understanding of the process of image formation, as described briefly in the next section.

Several other filtering operations can be performed by using the same setup. For example, a still picture of a TV frame displays the characteristic raster scan pattern in the form of fine horizontal lines. The spectrum of these lines is a set of dots in the perpendicular direction that are quite clearly distinguishable from the spectrum of the rest of the image. The dots may be described mathematically as a series of delta functions (if the grating is considered of infinite extent) that correspond to the Fourier transform of a periodic object [Eq. (A21)]. A well-constructed filter will eliminate the spectrum of the lines and will only minimally impact the rest of the image. Matched filters can also be constructed to aid in optical recognition of specific features within the image. All of these applications are described in several texts (11,13,17,20).

THE ABBE THEORY OF IMAGE FORMATION AND RESOLUTION LIMITS

Ernst Abbe (42) was the first to understand image formation as a diffraction and interference process. He was particularly interested in image formation by a

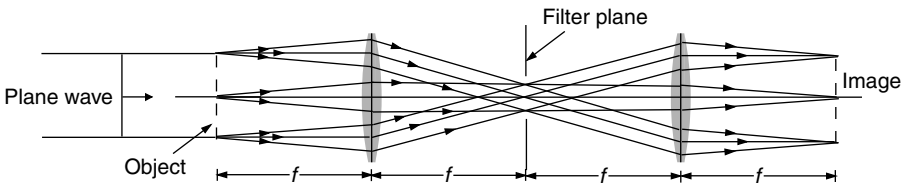


Figure 53. A system for illustrating coherent imaging concepts and for performing spatial filtering experiments. The two lenses have equal power and are separated by twice their focal length. Object and image are one focal length away from the first and second lens, respectively.

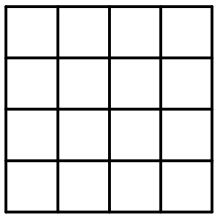


Figure 54. A square grid object.

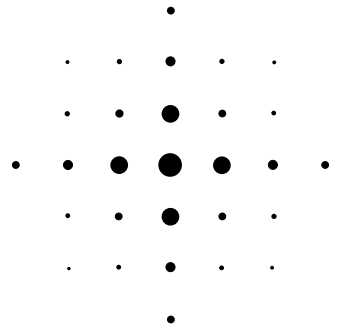


Figure 55. Approximate appearance of the spectra corresponding to the square grid object of Fig. 54. The size of the circles attempts to indicate the relative strength of the diffracted orders, except for the central (undiffracted) one, which would be considerably brighter than indicated.

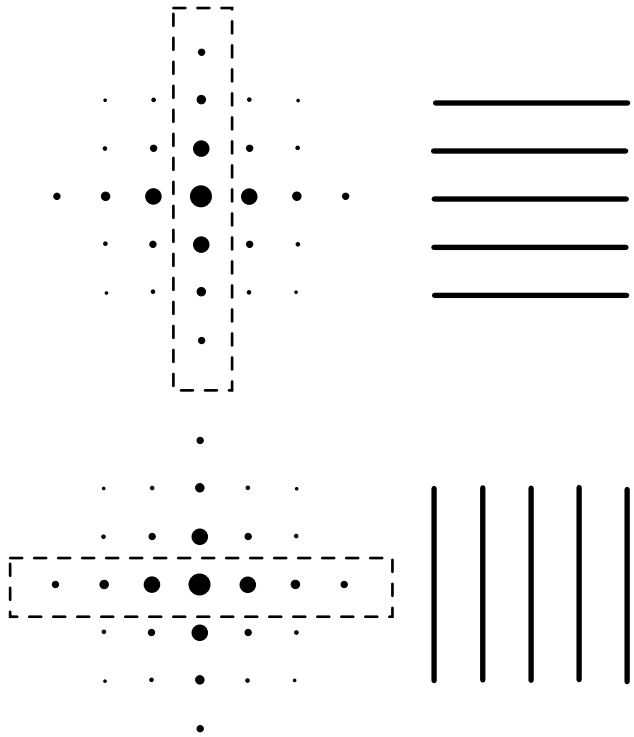


Figure 56. Spatial frequency filtering of the square grid object. A rectangular aperture (shown in dashed lines) allows only the vertical or horizontal orders, giving the image the appearance of a vertical or horizontal grating, respectively. Although not shown, other orientations of the aperture (e.g., 45°) would produce a similar effect.

microscope, so we use his theory as a way of summarizing the various types of resolution limits. Abbe's view of image formation can be described as follows: Consider a periodic object, such as a grating, illuminated by a monochromatic plane wave. An angular spectrum of plane waves is diffracted from the grating. These waves are brought to a focus at the exit pupil plane, where the spectrum (or Fourier transform) of the grating is formed (this is shown in Fig. 53, where the Fourier transform appears at the filter plane). The spectra form a series of point sources that give rise to another set of waves in the image plane. The interference of those waves produces the image. To understand this last step aided by Fig. 53, notice that the image would be formed at infinity without the help of the second lens, or, if the object was not exactly one focal length away, the image would be formed at a finite distance. The second lens is not essential to this description and in this case serves merely to form the image at a finite distance.

We may generalize the situation shown in Fig. 53 if we allow the refractive indexes of the object and image space to be different (n, n') and the magnification to be other than unity (so that the second lens may have a focal length different from the first). We borrow this result from physical optics, that the first-order diffracted beam from the grating is at an angle θ given by $d \sin\theta = \lambda/n$ where n is the refractive index of the object (grating) space. From this equation, the grating spacing d may be expressed as

$$d = \frac{\lambda}{(n \sin\theta)} \tag{70}$$

The spectra formed at the plane of the aperture stop are points [or more precisely, functions depending on the aperture shape such as given by Eq. (34)] separated by a distance proportional to the focal length $(F \tan\theta)$. Now, suppose that the radius of the aperture stop is larger than this distance, so that all orders pass through. Each of the orders generates a plane wave at the image plane. The interference of the three plane waves gives rise to a sinusoidal amplitude distribution or interference fringes, just as in Young's interference experiment (12,13). The period of this grating is given by $d' = \lambda/(n' \sin\theta')$. The ratio $d'/d = (n \sin\theta)/(n' \sin\theta')$ is equal to the geometric magnification [see Eqs. (8), (24)], as expected.

The limit of resolution of a given system is a function of the aperture size, as we have already seen. In this case, the resolution limit is obtained when the diffraction angle is such that the point spectra corresponding to the ± 1 orders are formed exactly at the rim of the aperture. If the diffraction angle is larger and the first-order spectra are cut off by the aperture, only the zero-order spectrum remains and gives rise to a single plane wave that results in uniform illumination of the image plane. Therefore, the image detail has been lost. Thus, if a is the maximum angle in the object space for which the first-order spectra just pass, the resolution limit of a coherent system can be expressed as $\lambda/(n \sin a)$.

It is possible to obtain higher resolution from a coherent system if oblique illumination is used. If, on the same system shown in Fig. 53, the illumination forms an angle

a with the object plane, then the undiffracted beam will be focused at the rim of the aperture. Only one more grating order is needed to pass through the aperture to create interference fringes (and thus visible detail) at the image plane. Thus, the angle between the zero and the first order can now be $2a$, corresponding to a grating of twice the frequency. Using oblique illumination, then, the resolution limit can be halved to $0.5\lambda/(n \sin a)$.

It is worth appreciating the difference between an amplitude and a phase object relative to this discussion. The sinusoidal grating employed in the example of this section represents a sinusoidal variation in real amplitude (and hence transparency) and has a transmittance of the form $a + b\cos(2\pi x/x_0)$. As the expression shows, this type of grating has a Fourier transform that contains only the first and zeroth orders. The appearance of the negative first order in the diffraction pattern justifies the preferential use of the complex exponential form [Eq. (A6)] in writing the Fourier series, because a sinusoidal grating in that form has both a positive and negative first order. However, the Fourier (or diffracted) spectrum of a sinusoidal phase grating contains higher harmonics, as can be seen by taking the Fourier transform of the transmittance of such a grating, which has the form $\exp[i\cos(2\pi x/x_0)]$.

A phase object is invisible to a detector such as the human eye that responds to irradiance. Mathematically, the squared modulus of the phase object is a constant; thus, it contains no spatial variation that would render it visible to the eye. If the aperture of the imaging system is such that even the highest spatial frequency present in the object is transmitted, then the corresponding image irradiance distribution is uniform, and the image is invisible. Departures from that condition, whether in the form of truncated high spatial frequencies or defocus/aberrations, can render the image visible by altering the phase relationship of the Fourier components and hence, the order of their interference at the image plane. This is also the principle of phase contrast microscopy, which places a phase mask at the Fourier plane to retard the phase of a chosen frequency band.

Table 3. Various Forms of Resolution Limit

Formula	Name	Condition	Resolved object
$d' = 0.61\lambda/(n' \sin a')$	Rayleigh	Image space	Two mutually incoherent points
$d' = 0.5\lambda/(n' \sin a')$	MTF limit	Image space	Incoherent sinusoidal grating
$d = \lambda/(n \sin a)$	Abbe	Object space	Sinusoidal grating illuminated coherently by normally incident wave
$d = 0.5\lambda/(n \sin a)$	—	Object space	Sinusoidal grating illuminated coherently and obliquely

The various types of resolution limits shown in Table 3 summarize the entire wave theory of image formation for both incoherent and coherent imaging.

APPENDIX

SOME CONCEPTS OF FOURIER ANALYSIS

The Fourier Series for Periodic Functions

According to the Fourier theorem, any periodic function that satisfies certain simple conditions (to be stated later) can be written as a summation of sinusoidal functions whose frequencies are equal to integral multiples of the frequency of the original function. Let $f(x)$ be a periodic function of period p, that is, $f(x) = f(x + np)$ where n an integer, then

$$f(x) = \frac{a_0}{2} + \sum_{n=1}^{\infty}\left[a_n \cos\left(2\pi \frac{n}{p}x\right) + b_n \sin\left(2\pi \frac{n}{p}x\right)\right], \quad \text{(A1)}$$

where

$$a_0 = \frac{2}{p}\int_{x_1}^{x_2} f(x)\,dx, \quad \text{(A2)}$$

$$a_n = \frac{2}{p}\int_{x_1}^{x_2} f(x)\cos(2\pi \frac{n}{p}x)\,dx, \quad \text{(A3)}$$

$$b_n = \frac{2}{p}\int_{x_1}^{x_2} f(x)\sin(2\pi \frac{n}{p}x)\,dx, \quad \text{(A4)}$$

and $x_2 - x_1 = p$. A form of the Fourier theorem is employed when, for example, one breaks down a certain musical tone into pure tones (sinusoidal vibrations). In the example of sound, all pure tones have positive frequencies. However, in certain cases in optics, there is plausible physical meaning assigned to negative frequencies. For this reason, the preferred way of expressing Fourier's theorem in optics (at least most of the time) is through the complex exponential notation, in which sines and cosines are substituted by complex exponentials through

$$\cos x = \frac{1}{2}[\exp(ix) + \exp(-ix)],$$
$$\sin x = \frac{1}{2i}[\exp(ix) - \exp(-ix)]. \quad \text{(A5)}$$

Now, Fourier's theorem for periodic functions may be restated as follows:

$$f(x) = \sum_{n=-\infty}^{+\infty} F_n \exp\left(i2\pi \frac{n}{p}x\right), \quad \text{(A6)}$$

$$F_n = \frac{1}{p}\int_{x_1}^{x_2} f(x)\exp\left(-i2\pi \frac{n}{p}x\right)dx \quad \text{(A7)}$$

and $x_2 - x_1 = p$. Sufficient conditions (called the *Dirichlet* conditions) for the validity of Eqs. (A1) and (A6) are

1. $f(x)$ has only a finite number of discontinuities in any finite interval of x within the range $x_1 < x < x_2$.
2. $f(x)$ has only a finite number of extrema in any finite interval of x in the range x_1 to x_2.
3. The integral of $|f(x)|$ over p converges (is finite).

The coefficients F_n are the *spectrum* of $f(x)$, also called the *Fourier spectrum* or the *Fourier coefficients*, or *harmonics*. When we use Eq. (58) or (53)–(55) to determine the strength of the Fourier coefficients, we speak of Fourier analysis. When we add the sinusoidal (or complex exponential) terms to approximate $f(x)$ (i.e., when using Eq. (52) or (57), we speak of Fourier synthesis. Fourier synthesis is illustrated by an example in Fig. 39.

Nonperiodic Functions: Fourier Transforms

The Fourier theorem can be extended to nonperiodic functions, thus allowing Fourier analysis to represent an arbitrary (ir)radiance distribution. In the general case, as p goes to infinity to represent a nonperiodic function, the variable n/p in Eq. (A6) becomes continuous, and the Fourier series becomes a Fourier integral, called the Fourier transform:

$$f(x) = \int_{-\infty}^{+\infty} F(\sigma) \exp(i2\pi\sigma x)\, d\sigma, \tag{A8}$$

where

$$F(\sigma) = \int_{-\infty}^{+\infty} f(x) \exp(-i2\pi\sigma x)\, dx. \tag{A9}$$

The reader should be aware that several different ways exist for expressing the Fourier transform: the negative sign in the exponent may appear in Eq. (A8) instead of Eq. (A9), a constant $1/\sqrt{(2\pi)}$ may appear in front of both integrals when 2π is removed from the exponents, or a constant $1/(2\pi)$ may appear in front of one integral and not the other. All of these ways of expressing the Fourier transform are equivalent and lead to the same physics. The equation with the minus sign is called the inverse Fourier transform, and the equation with the positive sign is called the Fourier transform. But this can be cause for confusion because it is acceptable to call either function $f(x)$ or $F(\sigma)$ the Fourier transform of the other. In fact, the terminology "inverse Fourier transform" need be employed only when Fourier transformations are cascaded, in which case it helps to ensure that the correct sign is used in successive operations. Otherwise, the sign of the Fourier transform has no physical relevance.

The function $F(\sigma)$ is the *Fourier or frequency spectrum* of $f(x)$. Notice that if x has dimensions of length, then σ has dimensions of inverse length and is called *spatial frequency*, although in the theory of image formation, it pays to normalize these variables and make them dimensionless, as we have already done in equation 44[4]. The variable σ is analogous to the discrete variable n/p in Eqs. (A6), (A7). The two functions $f(x)$ and $F(\sigma)$ are called a Fourier transform pair to emphasize the symmetry between them. In other words, the two functions are mathematically similar objects. Notice that there is no such symmetry between a periodic function and its Fourier spectrum, as defined by Eqs. (A6) and (A7): the function and its spectrum are mathematically different objects because the spectrum is a set of integrals, rather than a single function. These two facts pose a problem. If the Fourier transform of a periodic function were to exist, then, according to Eqs. (A8) and (A9), it would have to be a mathematical object similar to the original function (it might not be a periodic function, but it should at least be a function). That would mean that it could not be written in the form of Eq. (A7). Therefore, either the Fourier transform of a periodic function could not exist, or it would have to represent something different from the spectrum given by Eq. (A7).

The full resolution of this mathematical conundrum would fill several pages (43,44) and can be only hinted at here. It is noted that this dilemma is intimately connected with the conditions for the existence of the Fourier transform, because as $p \to \infty$, the Dirichlet conditions would exclude many simple functions, starting with the function $f(x) = \text{const}$, whose integral from $-\infty$ to $+\infty$ does not converge. However, it is worth noting that if $f(x)$ is to represent a physical quantity such as a function of distance (an image) or time (a signal), then it cannot extend from $-\infty$ to $+\infty$ but must necessarily be defined within finite limits. This would mean that the function is not truly periodic; hence, the Fourier transform would be applicable, but the Fourier series would not be. Thus, if we allow the mathematics to be tempered by physical reality, the conflict disappears. Nevertheless, when a quasi-periodic function (periodic within an interval) has very many periods, it approximates a fully periodic function so well that it is economical to treat it as such. This is another way of saying that the mathematical abstraction represented by a periodic function is useful because it simplifies many concepts and calculations. Thus, it is desirable to define the Fourier transform of periodic functions, and of course, it is desirable that it should have the same meaning as the discrete Fourier spectrum, Eq. (A7).

To define the Fourier transform of a periodic function, we actually need to define a new class of functions, called *generalized functions* (43,44), which contain ordinary functions as a subset. Within this new set of functions, the Fourier transform and its inverse always exist. However, only one generalized function need concern us here, the so-called *delta function*; other generalized functions of interest can be considered derived from it.The delta function is defined through the following relationships:

$$\begin{aligned} \delta(\sigma) &= \infty \text{ if } \sigma = 0, \\ &= 0 \text{ if } \sigma \neq 0, \end{aligned} \tag{A10}$$

[4] The variable σ can also represent the wave number (or inverse wavelength) when the Fourier transform is used in coherence theory. Fourier transforms are ubiquitous in optics!

and

$$\int_{-\infty}^{+\infty} \delta(\sigma)\, d\sigma = 1. \tag{A11}$$

The delta function can be given physical meaning as, for example, the idealized concept of an electrical point charge. The charge is considered concentrated at a single point (taken here as zero), but the integral of the function that represents it must equal the charge value (taken here as unity). The simplest visualization of the delta function can be obtained as a rectangle of height b and width $1/b$, centered at zero. The area (integral) of the rectangle remains unity for all values of b. As $b \to \infty$, we obtain a delta function.

Now, consider the function $f(x) = 1$, if $-a/2 < x < a/2$, and zero elsewhere. The Fourier transform of this function is

$$F(\sigma) = \int_{-\infty}^{+\infty} f(x) \exp(-i2\pi\sigma x)\, dx = \int_{-a/2}^{+a/2} \exp(-i2\pi\sigma x)\, dx, \tag{A12}$$

from which, after performing the integration, we obtain

$$F(\sigma) = \frac{\sin(\pi a\sigma)}{\pi\sigma}. \tag{A13}$$

Now, as $a \to \infty$, $f(x)$ becomes the unit constant function, $f(x) = 1$. It may also be shown that

$$\lim_{a\to\infty} \int_{-\infty}^{+\infty} \frac{\sin(\pi a\sigma)}{\pi\sigma} = 1 \tag{A14}$$

and

$$\lim_{a\to\infty} \frac{\sin(\pi a\sigma)}{\pi\sigma} = 0 \text{ for } \sigma \neq 0, = \infty \text{ for } \sigma = 0. \tag{A15}$$

Upon comparing Eqs. (A14) and (A15) with Eqs. (A10) and (A11), we see that

$$\lim_{a\to\infty} \frac{\sin(\pi a\sigma)}{\pi\sigma} = \delta(\sigma) \tag{A16}$$

[where the operations of taking the limit and integrating are defined as meaningful only if carried out in the order indicated in Eq. (A14)]. Therefore, the functions $f(x) = 1$ and $\delta(\sigma)$ are a Fourier transform pair:

$$f(x) = 1 = \int_{-\infty}^{+\infty} \delta(\sigma) \exp(i2\pi\sigma x)\, d\sigma, \tag{A17}$$

and

$$\delta(\sigma) = \int_{-\infty}^{+\infty} \exp(-i2\pi\sigma x)\, dx. \tag{A18}$$

The function F in Eq. (A13) is encountered frequently enough to be given a name. It is called a *sinc* function [see also Eq. (48)].

The Fourier Transform of a Periodic Function

Consider first the simplest periodic function, a sinusoid or simple harmonic function, written in exponential form:

$$f(x) = \exp(i2\pi\sigma_0 x) \tag{A19}$$

The Fourier transform of this function is [from Eq. (A9)]

$$F(\sigma) = \int_{-\infty}^{+\infty} \exp[i2\pi(\sigma_0 - \sigma)x]\, dx$$

and by virtue of Eq. (A18),

$$F(\sigma) = \delta(\sigma - \sigma_0). \tag{A20}$$

Thus the Fourier spectrum of a simple harmonic function of frequency σ_0 is a delta function centered at $\sigma = \sigma_0$. This is one of the reasons that the negative sign appears in the exponent of the Fourier spectrum [Eq. (A9], instead of Eq. (A8). Had this alternative way been followed, the spectrum of the harmonic function of frequency σ_0 would be centered at $\sigma = -\sigma_0$, which is less satisfactory. There is, however, no actual physical difference between the two representations.

Now, consider a general periodic function that can be expressed as a Fourier series, per Eq. (A6). Without the need for introducing new mathematics, it can be shown that the Fourier transform of that function is

$$F(\sigma) = \sum_{n=-\infty}^{+\infty} F_n \delta\left(\sigma - \frac{n}{p}\right), \tag{A21}$$

by virtue of Eqs. (A9) and (A13). Thus, the Fourier transform of a periodic function is a series of equally spaced delta functions, whose strengths are equal to F_n, located at frequencies n/p. This is the way that the Fourier spectrum of a periodic function may be described in terms of the Fourier transform. The function described by Eq. (A21) is shown schematically in Fig. 57.

The Two-Dimensional Fourier Transform

The definition of the Fourier transform can be readily extended to two dimensions. The two-dimensional Fourier

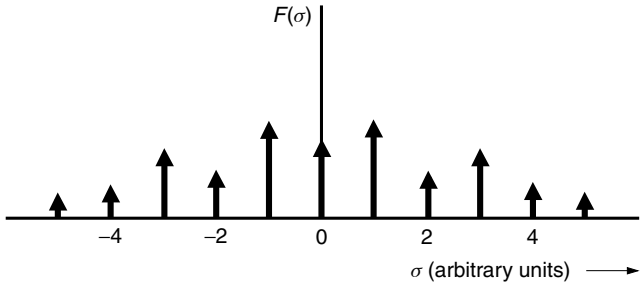


Figure 57. Schematic of the Fourier transform of a periodic function. Each arrow represents a delta function, and the corresponding harmonic strength is indicated by the height of the arrow.

transform may be written as

$$f(x,y) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} F(\sigma,\tau)\exp[i2\pi(\sigma x+\tau y)]\,d\sigma\,d\tau, \qquad \text{(A22)}$$

where

$$F(\sigma,\tau) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f(x,y)\exp[-i2\pi(\sigma x+\tau y)]\,dx\,dy \qquad \text{(A23)}$$

Similar extension can apply to functions of three or more variables, which, however, are not normally encountered in the theory of image formation.

Some Important Definitions and Theorems

We give these theorems in one-dimensional form for simplicity of notation, but the extension to two dimensions is straightforward. Proofs can be found in (43–45).

Parseval's Theorem. If $f(x)$ and $F(\sigma)$ are a Fourier transform pair, then,

$$\int_{-\infty}^{+\infty} |F(\sigma)|^2\,d\sigma = \int_{-\infty}^{+\infty} |f(x)|^2\,dx. \qquad \text{(A24)}$$

This theorem expresses, in general, the conservation of energy (or irradiance). Thus, it ensures, for example, that the integral of the PSF always converges for a system that has a finite pupil size.

Convolution. A function $A'(u')$ is said to be the convolution of two functions $A(u)$ and $F(u)$ if

$$A'(u') = \int_{-\infty}^{+\infty} A(u)F(u'-u)\,du. \qquad \text{(A25)}$$

If $a'(x)$, $a(x)$, and $f(x)$ are the Fourier transforms of A', A, and F, respectively, then,

$$a'(x) = a(x)f(x). \qquad \text{(A26)}$$

This theorem can be extended to an arbitrary number of functions (repeated convolution).

Autocorrelation. The autocorrelation function of $F(x)$ is defined as

$$C(\xi) = \int_{-\infty}^{+\infty} F^*(x)F(x+\xi)\,dx, \qquad \text{(A27)}$$

where * denotes complex conjugate. The (inverse) Fourier transform of $C(\xi)$ is given by

$$c(u) = |f(u)|^2, \qquad \text{(A28)}$$

where $f(u)$ is the inverse Fourier transform of $F(x)$.

ABBREVIATIONS AND ACRONYMS

CTF	coherent transfer function
EM	electromagnetic
ESF	edge spread function
FT	fourier transform
LSF	line spread function
MTF	modulation transfer function
OTF	optical transfer function
PMR	paraxial marginal ray
PPR	paraxial pupil ray
PSF	point spread function
PTF	phase transfer function

BIBLIOGRAPHY

1. D. S. Goodman, in M. Bass, ed., *Handbook of Optics*, vol. I, McGraw-Hill, 1995.
2. P. Mouroulis and J. Macdonald, *Geometrical Optics and Optical Design*, Oxford University Press, 1997.
3. W. J. Smith, *Modern Optical Engineering*, McGraw-Hill, 1990.
4. L. Levi, *Applied Optics: A Guide to Optical System Design*, Wiley, 1968.
5. R. Kingslake, *Optical System Design*, Academic Press, 1983.
6. D. C. O'Shea, *Elements of Modern Optical Design*, Wiley, 1985.
7. W. J. Smith, in M. Bass, ed., *Handbook of Optics*, vol. I, McGraw-Hill, 1995.
8. R. S. Longhurst, *Geometrical and Physical Optics*, Longman, UK, 1967.
9. W. T. Welford, *Aberrations of Optical Systems*, Adam Hilger, 1986.
10. V. N. Mahajan, *Aberration Theory Made Simple*, SPIE Press, Bellingham, WA, 1991.
11. R. Guenther, *Modern Optics*, Wiley, 1990.
12. M. Born and E Wolf, *Principles of Optics*, 7th ed., Cambridge University Press, 1999.
13. E. Hecht, *Optics*, Addison-Wesley, 1987.
14. G. H. Spencer and M. V. R. K. Murty, *J. Opt. Soc. Am.* **52**, 672–678 (1962).
15. H. H. Hopkins, *The Wave Theory of Aberrations*, Oxford University Press, 1950.
16. V. N. Mahajan, *Optical Imaging and Aberrations Part 1: Ray Geometrical Optics*, SPIE Press, 1999.
17. J. W. Goodman, *Introduction to Fourier Optics*, 2nd ed., McGraw-Hill, 1996.
18. J. D. Gaskill, *Linear Systems, Fourier Transforms, and Optics*, Wiley, 1978.
19. H. H. Hopkins, in R. Shannon and J. Wyant, eds., *Applied Optics and Optical Engineering I*, vol. IX, Academic Press, 1983.
20. G. O. Reynolds, J. B. DeVelis, G. B. Parrent Jr., and B. J. Thompson, *The New Physical Optics Notebook*, SPIE Press, Bellingham, WA, 1989.
21. V. N. Mahajan, *J. Opt. Soc. Am.* **72**, 1,258–1,266 (1982).
    V. N. Mahajan, *J. Opt. Soc. Am. A* **3**, 470–485 (1986).
22. B. R. A. Nijboer, *Physica* **10**, 679–692 (1943).
23. B. R. A. Nijboer, *Physica* **13**, 605–620 (1947).
24. K. Nienhuis and B. R. A. Nijboer, *Physica* **14**, 590–603 (1949).

25. M. Cagnet, M. Francon, and J. C. Thrierr, *Atlas of Optical Phenomena*, Springer-Verlag, 1962.
26. K Strehl, *Zeitschrift f. Instrumentenkunde* **22**, 213–217 (1902).
27. See sect. 8.3 in ref. 10, and also V. N. Mahajan, *J. Opt. Soc. Am.* **73**, 860–861 (1983).
P. Mouroulis, in P. Mouroulis, ed., *Visual Instrumentation: Optical Design and Engineering Principles*, McGraw-Hill, 1999.
28. A. Maréchal, *Revue d'Optique* **26**, 257–277 (1947).
29. W. B. King, *J. Opt. Soc. Am.* **58**, 655–661 (1968).
30. P. Mouroulis and H. Zhang, *J. Opt. Soc. Am. A* **9**, 34–42 (1992).
31. V. N. Mahajan, *Appl. Opt.* **17**, 964–968 (1978).
32. H. H. Hopkins and B. Zalar, *J. Mod. Opt.* **34**, 371–406 (1987).
33. J. Macdonald, *Optica Acta* **18**, 269–290 (1971).
34. H. H. Hopkins, *Proc. R. Soc. A* **231**, 91–103 (1955).
35. H. H. Hopkins, K. J. Habell, ed., *Proceedings of the Conference on Optical Instruments and Techniques*, Chapman and Hall, London, 1961, pp. 480–514.
36. H. H. Hopkins, *Optica Acta* **31**, 345–368 (1984).
37. A. P. Shaw, *Optica Acta* **33**, 1,389–1,396 (1986).
38. T. L. Williams, *The Optical Transfer Function of Imaging Systems*, Institute of Physics, 1999.
39. J. P. Mills and B. J. Thompson, *J. Opt. Soc. Am. A* **3**, 694–716 (1986).
40. H. H. Hopkins, *Photogr. Sci. Eng.* **21**, 114–123 (1977).
41. A. B. Porter, *Philos. Mag.* **11**(6), 154 (1906).
42. E. Abbe, *Archiv. Mikroskopische Anat.* **9**, 413–468 (1873).
43. M. J. Lighthill, *Introduction to Fourier Analysis and Generalized Functions*, Cambridge University Press, 1958.
44. P. Dennery and A. Krzywicki, *Mathematics for Physicists*, Harper and Row, 1967.
45. R. Bracewell, *The Fourier Transform and Its Applications*, 2nd ed., McGraw-Hill, 1965.

OPTICAL MICROSCOPY

MICHAEL W. DAVIDSON
The Florida State University
Tallahassee, FL

MORTIMER ABRAMOWITZ
Olympus America, Inc.
Melville, NY

INTRODUCTION

The past decade has witnessed enormous growth in the application of optical microscopy to micron and submicron level investigations in a wide variety of disciplines (1–5). Rapid development of new fluorescent labels has accelerated the expansion of fluorescence microscopy in laboratory applications and research (6–8). Advances in digital imaging and analysis have also enabled microscopists to acquire quantitative measurements quickly and efficiently on specimens that range from photosensitive caged compounds and synthetic ceramic superconductors to real-time fluorescence microscopy of living cells in their natural environment (2,9). Optical microscopy, helped by digital video, can also be used to image very thin optical sections, and confocal optical systems are now operating at most major research institutions (10–12).

Early microscopists were hampered by optical aberration, blurred images, and poor lens design, which restricted high-resolution observations until the nineteenth century. Aberrations were partially corrected by the mid-nineteenth century by the introduction of Lister and Amici achromatic objectives that reduced chromatic aberration and raised numerical apertures to around 0.65 for dry objectives and up to 1.25 for homogeneous immersion objectives (13). In 1886, Ernst Abbe's work with Carl Zeiss led to the production of apochromatic objectives based for the first time on sound optical principles and lens design (14). These advanced objectives provided images that had reduced spherical aberration and were free of color distortions (chromatic aberration) at high numerical apertures.

Several years later, in 1893, Professor August Köhler reported a method of illumination, which he developed to optimize photomicrography, that allowed microscopists to take full advantage of the resolving power of Abbe's objectives. The last decade of the nineteenth century saw innovations in optical microscopy, including metallographic microscopes, anastigmatic photolenses, binocular microscopes that had image-erecting prisms, and the first stereomicroscope (14).

Early in the twentieth century, microscope manufacturers began parfocalizing objectives, that allowed the image to remain in focus when the microscopist exchanged objectives on the rotating nosepiece. In 1924, Zeiss introduced a LeChatelier-style metallograph that had infinity-corrected optics, but this method of correction was not used widely for another 60 years.

Shortly before World War II, Zeiss created several prototype *phase contrast* microscopes based on optical principles advanced by Frits Zernike. Several years later, the same microscopes were modified to produce the first time-lapse cinematography of cell division photographed with phase-contrast optics (14). This contrast-enhancing technique did not become universally recognized until the 1950s and is still a method of choice for many cell biologists today.

Physicist Georges Nomarski introduced improvements in Wollaston prism design for another powerful contrast-generating microscopy theory in 1955 (15). This technique is commonly referred to as *Nomarski interference* or *differential interference contrast* (DIC) microscopy and, along with phase contrast, has allowed scientists to explore many new arenas in biology using living cells or unstained tissues. Robert Hoffman (16) introduced another method of increasing contrast in living material by taking advantage of phase gradients near cell membranes. This technique, now termed Hoffman Modulation Contrast, is available as optional equipment on most modern microscopes.

The majority of microscopes manufactured around the world had fixed mechanical tube lengths (ranging from 160–210 mm) until the late 1980s, when manufacturers largely changed over to infinity-corrected optics. Ray paths through both finite-tube-length and infinity-corrected microscopes are illustrated in Fig. 1. The upper portion

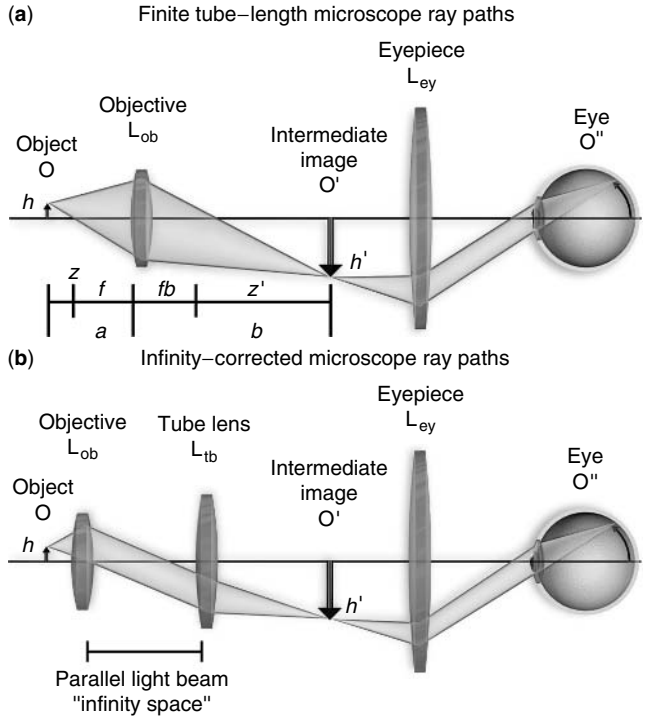


Figure 1. Optical trains of finite-tube and infinity-corrected microscope systems. (**a**) Ray traces of the optical train representing a theoretical finite-tube-length microscope. The object **O** is a distance **a** from the objective $\mathbf{L}_{ob}$ and projects an intermediate image **O'** at the finite tube length **b**, which is further magnified by the eyepiece $\mathbf{L}_{ey}$ and then projected onto the retina at **O"**. (**b**) Ray traces of the optical train that represents a theoretical infinity-corrected microscope system.

of the figure contains the essential optical elements and ray traces that define the optical train of a conventional finite-tube-length microscope (17). An object O of height h is being imaged on the retina of the eye at O". The objective lens L_{ob} projects a real and inverted image of O magnified to the size O' into the intermediate image plane of the microscope. This occurs at the eyepiece diaphragm, at the fixed distance $fb + z'$ behind the objective. In this diagram, fb represents the back focal length of the objective and z' is the optical tube length of the microscope. The aerial intermediate image at O' is further magnified by the microscope eyepiece L_{ey}) and produces an erect image of the object at O" on the retina, which appears inverted to the microscopist. The magnification factor of the object is calculated by considering the distance a between the object O and the objective L_{ob}, and the front focal length f of the objective lens. The object is placed a short distance z outside of the objective's front focal length f, such that $z + f = a$. The intermediate image of the object O' is located at distance b, which equals the back focal length of the objective fb plus z', the optical tube length of the microscope. Magnification of the object at the intermediate image plane equals h'. The image height at this position is derived by multiplying the microscope tube length b by the object height h, and dividing this by the distance of the object from the objective: $h' = (h \times b)/a$. From this argument, we can conclude that the lateral or transverse magnification of the objective is equal to a factor of b/a (also equal to f/z and z'/fb, the back focal length of the objective divided by the distance of the object from the objective. The image at the intermediate plane h' is further magnified by a factor of 25 centimeters (called the *near* distance to the eye) divided by the focal length of the eyepiece. Thus, the total magnification of the microscope is equal to the magnification by the objective multiplied by that of the eyepiece. The visual image (virtual) appears to the observer as if it were 10 inches away from the eye.

Most objectives are corrected to work within a narrow range of image distances, and many are designed to work only in specifically corrected optical systems that have matching eyepieces. The magnification inscribed on the objective barrel is defined for the tube length of the microscope for which the objective was designed.

The lower portion of Fig. 1 illustrates the optical train that uses ray traces of an infinity-corrected microscope system. The components of this system are labeled similarly to the finite-tube-length system for easy comparison. Here, the magnification of the objective is the ratio h'/h, which is determined by the tube lens L_{tb}. Note the *infinity space* that is defined by parallel light beams in every azimuth between the objective and the tube lens. This is the space used by microscope manufacturers to add accessories such as vertical illuminators, DIC prisms, polarizers, and retardation plates, whose designs are much simpler and have little distortion of the image (18). The magnification of the objective in the infinity-corrected system equals the focal length of the tube lens divided by the focal length of the objective.

FUNDAMENTALS OF IMAGE FORMATION

When light from the microscope lamp in the optical microscope, passes through the condenser and then through the specimen (assuming that the specimen is light absorbing), some of the light passes undisturbed in its path around and through the specimen. Such light is called direct light or undeviated light. The background light (often called the surround) that passes around the specimen is also undeviated light.

Some of the light that passes through the specimen is deviated when it encounters parts of the specimen. Such deviated light (as you will subsequently learn, called diffracted light) is rendered one-half wavelength or 180° out of phase with the direct light that has passed through undeviated. The one-half wavelength out of phase, caused by the specimen itself, enables this light to cause destructive interference with the direct light when both arrive at the intermediate image plane at the fixed diaphragm of the eyepiece. The eye lens of the eyepiece further magnifies this image which finally is projected onto the retina, the film plane of a camera, or the surface of a light-sensitive computer chip.

What has happened is that the direct or undeviated light is projected by the objective and spread evenly across the entire image plane at the diaphragm of the eyepiece. The light diffracted by the specimen is brought to focus at various localized places on the same image plane,

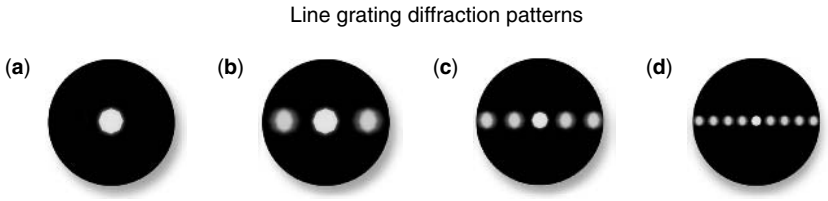


Figure 2. Diffraction spectra seen at the rear focal plane of the objective through a focusing telescope when imaging a closely spaced line grating. (**a**) Image of the condenser aperture diaphragm whose stage's empty. (**b**) Two diffraction spectra from a 10X objective when a finely ruled line grating is placed on the microscope stage. (**c**) Diffraction spectra of the line grating from a 40X objective. (**d**) Diffraction spectra of the line grating from a 60X objective.

where the diffracted light causes destructive interference, reduces intensity, and results in more or less dark areas. These patterns of light and dark are what we recognize as an image of the specimen. Because our eyes are sensitive to variations in brightness, the image becomes a more or less faithful reconstitution of the original specimen.

To help understand the basic principles, it is suggested that readers try the following exercise and use an object of known structure, such as a stage micrometer or similar grating of closely spaced dark lines as a *specimen*. To proceed, place the finely ruled grating on the microscope stage and bring it into focus using first a 10X and then the 40X objective (18). Remove the eyepiece and, insert a phase telescope in its place, so that the rear focal plane of the objective can be observed. If the condenser aperture diaphragm is closed most of the way, a bright white central spot of light will appear at the back of the objective, which is the image of the aperture diaphragm. To the right and left of the central spot, will be a series of spectra (also images of the aperture diaphragm), each colored blue on the part closest to the central spot and colored red on the part of the spectrum farthest from the central bright spot (as illustrated in Fig. 2). The intensity of these colored spectra decreases according to the distance from the central spot to the spectrum (17,18).

Those spectra nearer the periphery of the objective are dimmer than those closer to the central spot. The diffraction spectra illustrated in Fig. 2 use three different magnifications. In Fig. 2b, the diffraction pattern visible at the rear focal plane of the 10X objective contains two diffraction spectra. If the grating is removed from the stage, as illustrated in Fig. 2a, these spectra disappear, and only the central image of the aperture diaphragm remains. If the grating is reinserted, the spectra reappear once again. Note that the spaces between the colored spectra appear dark. Only a single pair of spectra can be observed if the grating is examined with the 10X objective. In this case, one diffraction spot appears to the left, and one appears to the right of the central aperture opening. If the line grating is examined with a 40X objective, as shown in Fig. 2c, several diffraction spectra appear to the left and right of the central aperture. When the magnification is increased to 60X (and assuming that it has a higher numerical aperture than the 40X objective), additional spectra (Fig. 2d) appear to the right and left than are visible by using the 40X objective.

Because the colored spectra disappear when the grating is removed, it can be assumed that the specimen itself affects the light passing through, thus producing the colored spectra. Further, if the aperture diaphragm is closed down, we will observe that objectives of higher numerical aperture *grasp* more of these colored spectra than objectives of lower numerical aperture. The crucial importance of these two statements for understanding image formation will become clear in the ensuing paragraphs.

The central spot of light (image of the condenser aperture diaphragm) represents the direct or undeviated light that passes undisturbed through or around the specimen as illustrated in Fig. 3b. It is called the zeroth order. The fainter images of the aperture diaphragm on each side of the zeroth order are called the first, second, third, fourth, etc. orders, respectively, as represented by the simulated diffraction pattern in Fig. 3a that would be observed at the rear focal plane of a 40X objective. All of

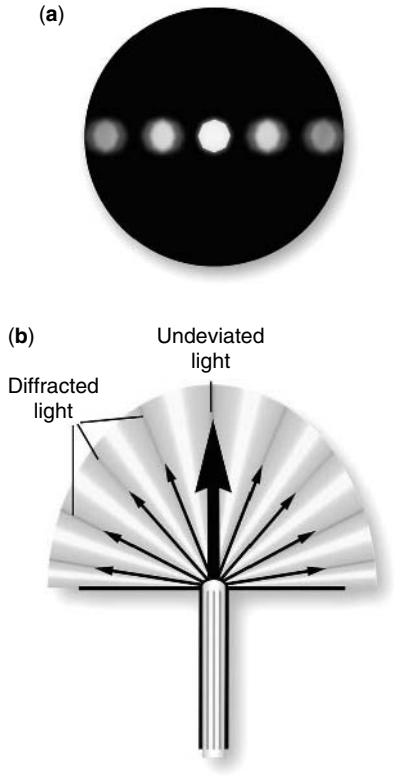


Figure 3. Diffraction spectra generated at the rear focal plane of the objective by undeviated and diffracted light. (**a**) Spectra visible through a focusing telescope at the rear focal plane of a 40X objective. (**b**) Schematic diagram of light both diffracted and undeviated by a line grating on the microscope stage.

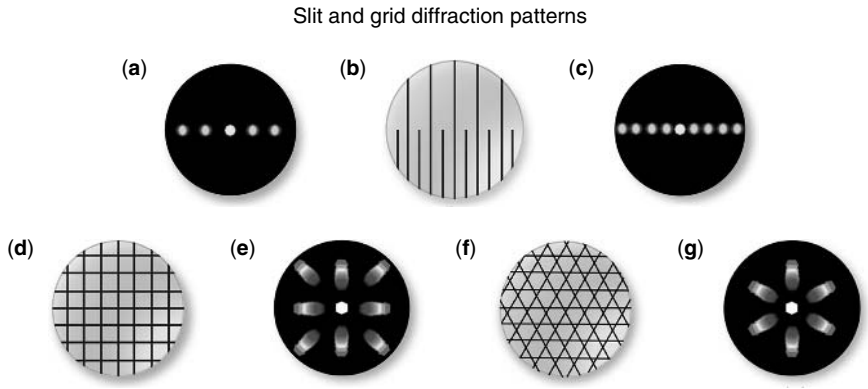


Figure 4. Diffraction patterns generated by narrow and wide slits and by complex grids. (**a**) Orthoscopic image of the grid seen at the rear focal plane of the objective when focused on the wide slit pattern in (**b**). (**b**) Conoscopic image of the grid that has a greater slit width at the top and a lesser width at the bottom. (**c**) Orthoscopic image of the narrow width portion of the grid [lower portion of (**b**)]. (**d**) and (**f**) Conoscopic images of grid lines arranged in a square pattern (**d**) and a hexagonal pattern (**f**). (**e**) and (**g**) Orthoscopic images of patterns in (**d**) and (**f**), respectively.

the *captured* orders in this case, represent, the diffraction pattern of the line grating seen at the rear focal plane of the objective (18).

The fainter diffracted images of the aperture diaphragm are caused by light deviated or diffracted, spread out in fan shape, at each of the openings of the line grating (Fig. 3b). The blue wavelengths are diffracted at a lesser angle than the green wavelengths, which are diffracted at a lesser angle than the red wavelengths.

The blue wavelengths from each slit at the rear focal plane of the objective, interfere constructively to produce the blue area of the diffracted image of each spectrum or order; similarly for the red and green areas (Fig. 3a). Where the diffracted wavelengths are one-half wave out of phase for each of these colors, the waves destructively interfere. Hence, the dark areas between the spectra or orders. All wavelengths from each slit add constructively at the position of the zeroth order. This produces the bright white light you see as the zeroth order at the center of the rear focal plane of the objective (Figs. 2, 3, and 4).

The closer the spacing of a line grating, the fewer the spectra that will be captured by a given objective, as illustrated in Fig. 4a–c. The diffraction pattern illustrated in Fig. 4a was captured by a 40X objective imaging the lower portion the line grating in Fig. 4b, where the slits are closer together (17,18). In Fig. 4c, the objective is focused on the upper portion of the line grating (Fig. 4b) where the slits are farther apart, and more spectra are captured by the objective. The direct light and the light from the diffracted orders continue on, focused by the objective, to the intermediate image plane at the fixed diaphragm of the eyepiece. Here, the direct and diffracted light rays interfere and are thus reconstituted into the real, inverted image that is *seen* by the eye lens of the eyepiece and further magnified. This is illustrated in Fig. 4d–g by two types of diffraction gratings. The square grid illustrated in Fig. 4d represents the orthoscopic image of the grid (i.e., the usual specimen image) as seen through the full aperture of the objective. The diffraction pattern derived from this grid is shown as a conoscopic image that would be seen at the rear focal plane of the objective (Fig. 4e). Likewise, the orthoscopic image of a hexagonally arranged grid (Fig. 4f) produces a corresponding hexagonally arranged conoscopic image of first-order diffraction patterns (Fig. 4g).

Microscope specimens can be considered complex gratings that have details and openings of various sizes. This concept of image formation was largely developed by Ernst Abbe, the famous German microscopist and optics theoretician of the nineteenth century. According to Abbe (his theories are widely accepted now), the details of a specimen will be resolved if the objective captures the zeroth order of the light and at least the first order (or any two adjacent orders, for that matter). The greater the number of diffracted orders that gain admittance to the objective, the more accurately the image represents the original object (2,14,17,18).

Further, if a medium of refractive index higher than that of air (such as immersion oil) is used in the space between the front lens of the objective and the top of the coverslip, as shown in Fig. 5a, the angle of the diffracted orders is reduced, and the fans of diffracted light are compressed. As a result, an oil immersion objective can

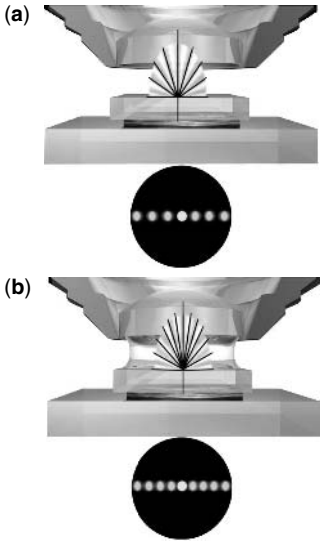


Figure 5. Effect of refractive index of imaging medium on diffracted orders captured by the objective. (**a**) Orthoscopic image of objective back focal plane diffraction spectra when air is the medium between the coverslip and the objective front lens. (**b**) Diffraction spectra when immersion oil of refractive index similar to glass is used in the space between the coverslip and the objective front lens.

capture more diffracted orders and yield better resolution than a dry objective (Fig. 5b). Moreover, because blue light is diffracted at a lesser angle than green or red light, a lens of a given aperture may capture more orders of light when the wavelengths are in the blue region of the visible light spectrum. These two principles explain the classic Rayleigh equation often cited for resolution (2,18–20):

$$d = 1.22(\lambda/2\mathrm{NA}) \quad (1)$$

where d is the space between two adjacent particles (that still allows perceiving the particles as separate), λ is the wavelength of illumination, and NA is the numerical aperture of the objective.

The greater the number of higher diffracted orders admitted into the objective, the smaller the details of the specimen that can be clearly separated (resolved). Hence, the value of using high numerical aperture for such specimens. Likewise, the shorter the wavelength of visible light used, the better the resolution. These ideas explain why apochromatic lenses of high numerical aperture can separate extremely small details in blue light.

Placing an opaque mask at the back of the objective blocks the outermost diffracted orders. This either reduces the resolution of the grating lines, or any other object details, or it destroys the resolution altogether, so that the specimen is not visible. Hence, the usual caution not to close down the condenser aperture diaphragm below the suggested two-thirds to nine-tenths of the objective's aperture.

Failure of the objective to grasp any of the diffracted orders results in an unresolved image. In a specimen that has very minute details, the diffraction fans are spread at a very large angle and require a high numerical aperture objective to capture them. Likewise, because the diffraction fans are compressed in immersion oil or in water, objectives designed for such use can resolve better than dry objectives.

If alternate diffracted orders are blocked out (still assuming the grating as our specimen), the number of lines in the grating appear doubled (a spurious resolution). The important caveat is that actions introduced at the rear of the objective can have a significant effect on the eventual image produced (18). For small details in a specimen (rather than a grating), the objective projects the direct and diffracted light onto the image plane of the eyepiece diaphragm as small, circular diffraction disks known as Airy disks (illustrated in Fig. 6). High numerical aperture objectives capture more of the diffracted orders and produce smaller disks than low numerical aperture objectives. In Fig. 6, Airy disk size is shown steadily decreasing from Fig. 6a through Fig. 6c. The larger disk sizes in Figs. 6a, b are produced by objectives that have lower numerical apertures, whereas the very sharp Airy disk in Fig. 6c is produced by an objective of very high numerical aperture (2,18).

The resulting image at the eyepiece diaphragm level is actually a mosaic of Airy disks that are perceived as light and dark regions of the specimen. When two disks are so close together that their central black spots overlap considerably, the two details represented by these

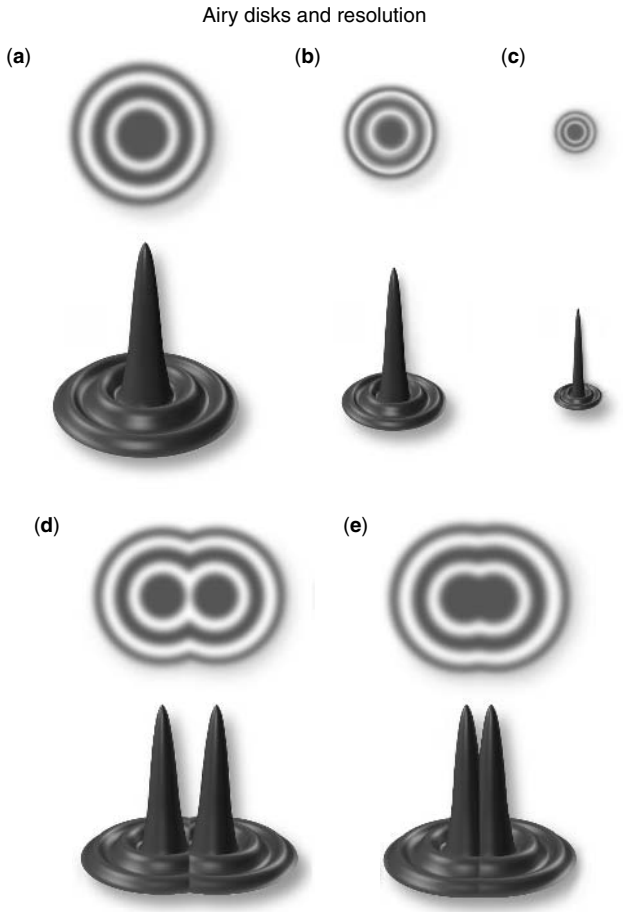


Figure 6. Airy disks and resolution. (a–c) Airy disk size as a function of objective numerical aperture, which decreases from (**a**) to (**c**) as numerical aperture increases. (**d**) Two Airy disks so close together that their ring structures overlap. (**e**) Airy disks at the limit of resolution.

overlapping disks are not resolved or separated and thus appear as one (illustrated in Fig. 6c). The Airy disks shown in Fig. 6d are just far enough apart to be resolved.

The basic principle to remember is that the combination of direct and diffracted light (or the manipulation of direct or diffracted light) is critically important in image formation. The key places for such manipulation are the rear focal plane of the objective and the front focal plane of the substage condenser. This principle is fundamental to most of the contrast improvement methods in optical microscopy (18; and see the section on **Contrast-enhancing techniques**); it is of particular importance at high magnification of small details close in size to the wavelength of light. Abbe was a pioneer in developing these concepts to explain image formation of light-absorbing or *amplitude* specimens (2,18–20).

Köhler Illumination

Proper illumination of the specimen is crucial in achieving high-quality images in microscopy and critical photomicrography. An advanced procedure for microscope illumination was first introduced in 1893 by August

Köhler of the Carl Zeiss corporation to provide optimum specimen illumination. All manufacturers of modern laboratory microscopes recommend this technique because it produces specimen illumination that is uniformly bright and free from glare and thus allows the user to realize the microscope's full potential.

Most modern microscopes are designed so that the collector lens and other optical components built into the base project an enlarged and focused image of the lamp filament onto the plane of the aperture diaphragm of a properly positioned substage condenser. Closing or opening the condenser diaphragm controls the angle of the light rays that emerge from the condenser and reach the specimen from all azimuths. Because the light source is not focused at the level of the specimen, illumination at specimen level is essentially grainless and extended and does not suffer deterioration from dust and imperfections on the glass surfaces of the condenser. The opening size of the condenser aperture diaphragm, along with the aperture of the objective, determines the realized numerical aperture of the microscope system. As the condenser diaphragm is opened, the working numerical aperture of the microscope increases and results in greater light transmittance and resolving power. Parallel light rays that pass through and illuminate the specimen are brought to focus at the rear focal plane of the objective, where the image of the variable condenser aperture diaphragm and the light source are observed in focus simultaneously.

Light pathways illustrated in Fig. 7 are schematically drawn to represent separate paths taken by the specimen-illuminating light rays and the image-forming light rays (17). This is not a true representation of any real segregation of these pathways, but it is a diagrammatic representation presented for visualization and discussion. The left-hand diagram in Fig. 7 demonstrates that the ray paths of illuminating light produce a focused image of the lamp filament at the plane of the substage condenser aperture diaphragm, the rear focal plane of the objective, and the eye point (also called the *Ramsden disk*) above the eyepiece. These areas in common focus are often referred to as conjugate planes, a principle that is critical in understanding the concept of Köhler illumination (2,17–21). By definition, an object that is in focus at one plane is also in focus at other conjugate planes of that light path. Four separate planes in each light pathway (both image forming and illumination), together make up a conjugate plane set. Conjugate planes in the path of the illuminating light rays in Köhler illumination (left-hand diagram in Fig. 7) include the lamp filament, condenser aperture diaphragm (at the front focal plane of the condenser), the rear focal plane of the objective, and the eye point of the eyepiece. The eye point is located approximately one-half inch (1 cm) above the top lens of the eyepiece, at the point where the observer places the front of the eye during observation.

Likewise, the conjugate planes in the image-forming light path in Köhler illumination (right-hand diagram in Fig. 7) include the field diaphragm, the focused specimen, the intermediate image plane (i.e., the plane of the fixed diaphragm of the eyepiece), and the retina of the eye or the film plane of the camera. The presence of conjugate focal planes is often useful in troubleshooting a microscope

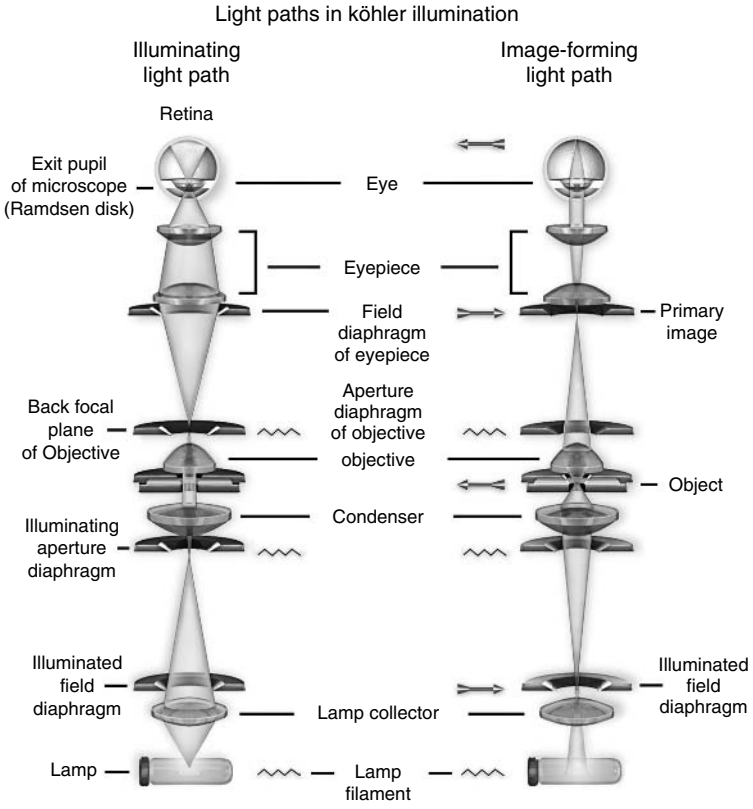


Figure 7. Light paths in Köhler illumination. The illuminating ray paths are illustrated on the left side and the image-forming ray paths on the right. Light emitted from the lamp passes through a collector lens and then through the field diaphragm. The aperture diaphragm in the condenser determines the size and shape of the illumination cone on the specimen plane. After passing through the specimen, light is focused on the back focal plane of the objective and then enters and is magnified by the ocular before passing into the eye.

for contaminating dust, fibers, and imperfections in the optical elements. When such artifacts are in sharp focus, it follows that they must reside on or near a surface that is part of the imaging-forming set of conjugate planes. Members of this set include the glass element at the microscope light port, the specimen, and the graticule (if any) in the eyepiece. Alternatively, if these contaminants are out of focus, then they occur near the illuminating set of elements that share conjugate planes. Suspects in this category are the condenser top lens (where dust and dirt often accumulate), the exposed eyepiece lens element (contaminants from eyelashes), and the objective front lens (usually fingerprint smudges).

In Köhler illumination, light emitted from the tungsten-halide lamp filament passes first through a collector lens located close to the lamp housing and then through a field lens that is near the field diaphragm. A sintered or frosted glass filter is often placed between the lamp and the collector lens to diffuse the light and ensure even intensity of illumination. In this case, the image of the lamp filament is focused onto the front focal plane of the condenser, while the diffuser glass is temporarily removed from the light path. The focal length of the collector lens must be carefully matched to the lamp filament dimensions to ensure that a filament image of the appropriate size is projected into the condenser aperture. For proper Köhler illumination, the image of the filament should completely fill the condenser aperture.

The field lens is responsible for bringing the image of the filament into focus at the plane of the substage condenser aperture diaphragm. A first surface mirror (positioned at a 45° angle to the light path) reflects focused light that leaves the field lens through the field diaphragm and into the substage condenser. The field diaphragm iris opening is a virtual light source for the microscope, and its image is focused by the condenser (raised or lowered) onto the specimen plane. Optical designs for the arrangement of these elements may vary by microscope manufacturer, but the field diaphragm should be positioned at a sufficient distance from the field lens to eliminate dust and lens imperfections from imaging in the plane of the specimen.

The field diaphragm in the base of the microscope controls only the width of the bundle of light rays that reaches the condenser — it does not affect the optical resolution, numerical aperture, or the intensity of illumination. Proper adjustment of the field diaphragm (i.e., focused by adjusting the height of the condenser and centered in the optical path, then opened to lie just outside the field of view) is important for preventing glare, which can reduce contrast in the observed image. Elimination of unwanted light is particularly important when attempting to image specimens that have inherently low contrast. When the field diaphragm is opened too far, scattered light originating from the specimen and light reflected at oblique angles from optical surfaces can act to degrade image quality.

The substage condenser is typically mounted directly beneath the microscope stage in a bracket that can be raised or lowered independently of the stage. The aperture diaphragm opening size is controlled by a swinging arm, a lever, or by rotating a collar on the condenser housing. The most critical aspect of achieving proper Köhler illumination is correct adjustment of the substage condenser. Condenser misalignment and an improperly adjusted condenser aperture diaphragm are the main sources of image degradation and poor quality photomicrography (19).

When properly adjusted, light from the condenser fills the rear focal plane of the objective and projects a cone of light into the field of view. The condenser aperture diaphragm is responsible for controlling the angle of the illuminating light cone and, consequently, the working numerical aperture of the condenser. With respect to the size and shape of condenser light cones, it is important to note, that reducing the size of the field diaphragm only slightly decreases the size of the lower portions of the light cone. The angle and numerical aperture of the light cone remain essentially unchanged as field diaphragm size is reduced (21). Illumination intensity should not be controlled by opening and closing the condenser aperture diaphragm or by shifting the condenser up and down or axially with respect to the optical center of the microscope. It should be controlled only by using neutral density filters placed into the light path or by reducing voltage to the lamp (although the latter is not usually recommended, especially for photomicrography). To ensure maximum performance of the tungsten-halide lamp, refer to the manufacturer's instrument manual to determine the optimum lamp voltage (usually 5–10 volts) and use that setting. Then, adding or removing neutral density filters can easily control the brightness of the illumination without affecting color temperature.

The size of the substage condenser aperture diaphragm opening should coincide with the desired numerical aperture, and the quality of the resulting image should also be considered. In general, the diaphragm should be set to a position that allows two-thirds to nine-tenths (60 to 90%) of the entire light disk size (visible at the rear focal plane of the objective after removing of the eyepiece or using a phase telescope or Bertrand lens). These values may vary due to extremes in specimen contrast.

The condenser aperture diaphragm should be set to an opening size that provides a compromise of resolution and contrast that depends, to a large degree, on the absorption, diffraction, and refraction characteristics of the specimen. This adjustment must be accomplished without overwhelming the image with artifacts that obscure detail and present erroneous enhancement of contrast. The amount of image detail and contrast necessary to produce the best photomicrograph also depends on refractive index, optical characteristics, and other specimen-dependent parameters.

When the aperture diaphragm is erroneously closed too far, resulting diffraction artifacts cause visible fringes, banding, and/or pattern formation in visual images and photomicrographs. Other problems, such as refraction phenomena, can also produce apparent structures in the image that are not real (21). Alternatively, opening the condenser aperture too wide causes unwanted glare and light scattering from the specimen and optical surfaces within the microscope and leads to a significant loss of contrast and washing out of image detail. The correct

setting will vary from specimen to specimen, and the experienced microscopist will soon learn to adjust the condenser aperture diaphragm (and numerical aperture of the system) accurately by observing the image without necessarily viewing the diaphragm in the rear focal plane of the objective. In fact, many microscopists (including the authors) believe that critical adjustment of the numerical aperture of the microscope system to optimize image quality is the single most important step in photomicrography.

When the illumination system of the microscope is adjusted for proper Köhler illumination, it must satisfy several requirements. The illuminated area of the specimen plane must be no larger than the field of view for any given objective/eyepiece combination. The light must also have uniform intensity, and the numerical aperture may vary from a maximum (equal to that of the objective) to a lesser value that depends on the optical characteristics of the specimen. Table 1 contains a list of objective numerical apertures versus the field of view diameter (for an eyepiece of field number 22 with no tube lens present — see discussion on field number) for each objective, from very low to very high magnifications.

Many microscopes are equipped with specialized substage condensers that have a swing-out top lens, which can be removed from the optical path for use with lower power objectives (2–5X). This action changes the performance of the remaining components in the light path, and some adjustment is necessary to achieve the best illumination conditions. The field diaphragm can no longer be used for aligning and centering the substage condenser and is now ineffective in limiting the area of the specimen under illumination. Much of the unwanted glare once removed by the field diaphragm is also reduced because the top lens of the condenser produces a light cone that has a much lower numerical aperture and allows light rays to pass through the specimen at much lower angles. Most important, the optical conditions for Köhler illumination no longer apply.

For low power objectives (2–5X), aligning of the microscope optical components and establishing of Köhler illumination conditions should always be undertaken at a higher (10X) magnification before removing the swing-out condenser lens for work at lower (5X and less) magnifications. Then, the height of the condenser should not be changed. Condenser performance is radically changed when the swing-out lens is removed (18,21). The image of the lamp filament is no longer formed in the aperture diaphragm, which ceases to control the numerical aperture of the condenser and the illumination system. In fact, the aperture diaphragm should be opened completely to avoid vignetting, a gradual fading of light at the edges of the view field.

Table 1. View-field Diameters (FN 22) (SWF 10X Eyepiece)[a]

Objective Magnification	Diameter (mm)
1/2X	44.0
1X	22.0
2X	11.0
4X	5.5
10X	2.2
20X	1.1
40X	0.55
50X	0.44
60X	0.37
100X	0.22
150X	0.15
250X	0.088

[a]Source: Nikon.

Contrast adjustment in low magnification microscopy is achieved by adjusting of the field diaphragm (18,19,21). When the field diaphragm is wide open (more than 80%), specimen details are washed out, and a significant amount of scattering and glare is present. Closing the field diaphragm to a position between 50 and 80% yields the best compromise in specimen contrast and depth of field. This adjustment is now visible at the rear focal plane of the objective when the eyepiece is removed or when a phase telescope or Bertrand lens is inserted into the eye tube. Objectives designed for low magnification have significantly simpler designs than their higher magnification counterparts. This is due to the smaller angles of illuminating light cones produced by low magnification condensers, which require objectives of lower numerical aperture.

Measurement graticules, which must be sharply focused and simultaneously superimposed on the specimen image, can be inserted into any of several conjugate planes in the image-forming path. The most common eyepiece (ocular) measuring and photomicrography graticules are placed in the intermediate image plane, which is positioned at the fixed diaphragm within the eyepiece. It is also theoretically possible to place graticules in any image-forming conjugate plane or in the plane of the illuminated field diaphragm. Stage micrometers are specialized *graticules* placed on microslides, which are used to calibrate eyepiece graticules and to make specimen measurements.

Color and neutral density filters are often placed in the optical pathway to reduce light intensity or alter the color characteristics of the illumination. There are several locations within the microscope stand where these filters are usually placed. Some modern laboratory microscopes have a filter holder sandwiched between the lamp housing and collector lens, which is an ideal location for these filters. Often, neutral density filters along with color correction filters and a frosted diffusion filter are placed together in this filter holder. Other microscope designs provide a set of filters built internally into the body, which can be toggled into the light path by levers. A third common location for filters is a holder mounted on the bottom of the substage condenser, below the aperture diaphragm, that will accept gelatin or glass filters.

It is important not to place filters in or near any of the image-forming conjugate planes to avoid imaging dirt or surface imperfections on the filters, along with the specimen (22). Some microscopes have an attachment for placing filters near the light port at the base (near the field diaphragm). This placement is probably too close to the

field diaphragm, and surface contamination may be either in sharp focus or appear as blurred artifacts superimposed on the image. For the same reasons, it is also not wise to place filters directly on the microscope stage.

MICROSCOPE OBJECTIVES, EYEPIECES, CONDENSERS, AND OPTICAL ABERRATIONS

Finite microscope objectives are designed to project a diffraction-limited image at a fixed plane (the *intermediate* image plane) that is dictated by the microscope tube length and located at a prespecified distance from the rear focal plane of the objective. Specimens are imaged at a very short distance beyond the front focal plane of the objective through a medium of defined refractive index, usually air, water, glycerin, or specialized immersion oils. Microscope manufacturers offer a wide range of objective designs to meet the performance needs of specialized imaging methods (2,6,9,18–21; and see the section on **Contrast-enhancing techniques**) to compensate for cover glass thickness variations and to increase the effective working distance of the objective.

All of the major microscope manufacturers have now changed their design to infinity-corrected objectives. Such objectives project emerging rays in parallel bundles from every azimuth to infinity. They require a tube lens in the light path to bring the image into focus at the intermediate image plane.

The least expensive (and most common) objectives are achromatic objectives, which are corrected for axial chromatic aberration in two wavelengths (red and blue) that are brought into the same focus. Further, they are corrected for spherical aberration in the color green, as described in Table 2. The limited correction of achromatic objectives leads to problems in color microscopy and photomicrography. When focus is chosen in the red–blue region of the spectrum, images will have a green halo (often termed *residual color*). Achromatic objectives yield their best results when light is passed through a green filter (often an interference filter) and using black-and-white film when these objectives are employed for photomicrography. The lack of correction for flatness of field (or field curvature) further hampers achromat objectives. In the past few years, most manufacturers have begun providing flat-field corrections for achromat objectives (see Fig. 8a) and have given these corrected objectives the name plan achromats.

Table 2. Objective Lens Types and Corrections[a]

Type	Corrections for Aberrations: Spherical	Corrections for Aberrations: Chromatic	Flatness Correction
Achromat	*[b]	2[c]	No
Plan achromat	*[b]	2[c]	Yes
Fluorite	3[d]	<3[d]	No
Plan fluorite	3[d]	<3[d]	Yes
Plan apochromat	4[e]	>4[e]	Yes

[a] Source: Nikon Instrument Group.
[b] Corrected for two wavelengths at two specific aperture angles.
[c] Corrected for blue and red — broad range of the visible spectrum.
[d] Corrected for blue, green, and red — full range of the visible spectrum.
[e] Corrected for dark blue, blue, green, and red.

The next higher level of correction and cost is in objectives called fluorites (named for the mineral used for one of the optical components) or semiapochromats illustrated by the center objective in Fig. 8b. This figure depicts three major classes of objectives: achromats that have the least amount of correction, as discussed above; fluorites (or semi-apochromats) that have additional spherical corrections; and apochromats that are the most highly corrected objectives available. Modern fluorite objectives are produced from advanced glass formulations that often contain low-dispersion materials such as fluorspar or newer synthetic substitutes (5). These new formulations allow for greatly improved correction of optical aberration. Similar to achromats, fluorite objectives are also corrected chromatically for red and blue light. In addition, fluorites are also corrected spherically for two or three colors. The superior correction of fluorite objectives compared to achromats enables making these objectives with higher numerical apertures, that result in brighter images. Fluorite objectives also have better resolving power than achromats and provide a higher degree of contrast; this makes them better suited than achromats for color photomicrography in white light.

The highest level of correction (and expense) is found in apochromatic objectives (see Fig. 8c), which are corrected chromatically for at least three colors (red, green, and blue), almost eliminating chromatic aberration, and are corrected spherically for three or four colors. Apochromatic objectives are the best choice for color photomicrography in white light. Because of their high level of correction, apochromat objectives usually have, higher numerical apertures than achromats or fluorites for a given magnification. Many of the newer high-end fluorite and apochromat objectives are corrected for four colors chromatically and four colors spherically.

All three types of objectives formerly suffered from pronounced field curvature and projected images that were curved rather than flat. To overcome this inherent condition, lens designers have produced flat-field corrected objectives that yield flat images. Such lenses are called plan achromats, plan fluorites, or plan apochromats, and although this degree of correction is expensive, these objectives are now routinely used due to their value in photomicrography.

Uncorrected field curvature is the most severe aberration in higher power fluorite and apochromat objectives, and it was tolerated as an unavoidable artifact for many years. During routine use, the view field had to be continuously refocused between the center and the edges to capture all specimen details. The introduction of flat-field (plan) correction to objectives perfected their use for photomicrography and video microscopy, and today these corrections are standard in both general use and high-performance objectives. Correction for field curvature adds a considerable number of lens elements to the objective, in many cases as many as four additional lenses. This significant increase in the number of lens elements for plan correction also occurs in already overcrowded fluorite and

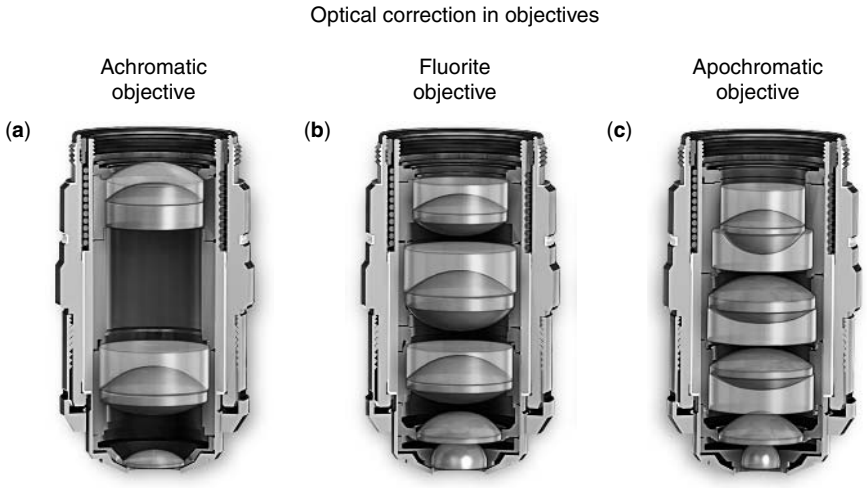


Figure 8. Levels of optical correction for aberration in commercial objectives. (**a**) Achromatic objectives, the lowest level of correction, contain two doublets and a single front lens, (**b**) Fluorites or semiapochromatic objectives, a medium level of correction, contain three doublets, a meniscus lens, and a single front lens, (**c**) Apochromatic objectives, the highest level of correction, contain a triplet, two doublets, a meniscus lens, and a single hemispherical front lens.

apochromat objectives and frequently results in a tight fit of lens elements within the objective barrel (4,5,18).

Before the transition to infinity-corrected optics, most objectives were specifically designed to be used with a set of oculars termed *compensating eyepieces*. An example is the former use of compensating eyepieces with highly corrected high numerical aperture objectives to help eliminate lateral chromatic aberration.

There is a wealth of information inscribed on the barrel of each objective, which can be broken down into several categories (illustrated in Fig. 9). These include the linear magnification, numerical aperture value, optical corrections, microscope body tube length, the type of medium for which the objective is designed, and other critical factors in deciding if the objective will perform as needed. Additional information is outlined here (17):

Optical Corrections. These are usually abbreviated as Achro (achromat), Apo (apochromat), and Fl, Fluar, Fluor, Neofluar, or Fluotar (fluorite) for better spherical and chromatic corrections, and as Plan, Pl, EF, Acroplan, Plan Apo or Plano for field curvature corrections. Other common abbreviations are ICS (infinity-corrected system) and UIS (universal infinity system), N and NPL (normal field of view plan), Ultrafluar (fluorite objective with glass that is transparent down to 250 nanometers), and CF and CFI (chrome-free; chrome-free infinity).



Figure 9. Specifications engraved on the barrel of a typical microscope objective. These include the manufacturer, correction levels, magnification, numerical aperture, immersion requirements, tube length, working distance, and specialized optical properties.

Numerical Aperture. This is a critical value that indicates the light acceptance angle, which in turn determines the light gathering power, the resolving power, and depth of field of the objective. Some objectives specifically designed for transmitted light fluorescence and dark-field imaging are equipped with an internal iris diaphragm that allows adjusting of the effective numerical aperture. Designation abbreviations for these objectives include I, Iris, W/Iris.

Mechanical Tube Length. This is the length of the microscope body tube between the nosepiece opening, where the objective is mounted, and the top edge of the observation tubes where the oculars (eyepieces) are inserted. Tube length is usually inscribed on the objective as the size in number of millimeters (160, 170, 210, etc.) for fixed lengths or as the infinity symbol (∞) for infinity-corrected tube lengths.

Cover Glass Thickness. Most transmitted light objectives are designed to image specimens that are covered by a cover glass (or *coverslip*). The thickness of these small glass plates is now standardized at 0.17 mm for most applications, although there is some variation in thickness within a batch of coverslips. For this reason, some of the high numerical aperture dry objectives have a correction collar adjustment of the internal lens elements to compensate for this variation (Fig. 10). Abbreviations for the correction collar adjustment include Corr, w/Corr, and CR, although the presence of a rotatable, knurled collar and graduated scale is also an indicator of this feature.

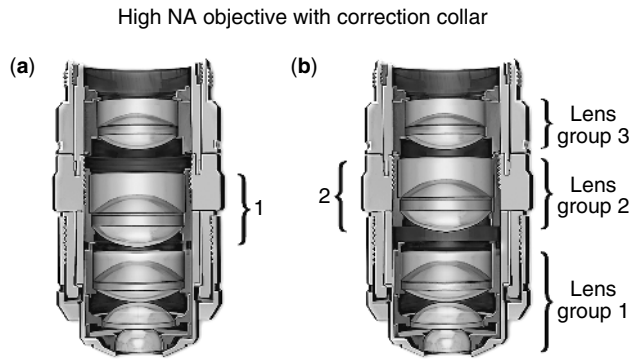


Figure 10. Objective that has three lens groups and correction collar for varying cover glass thicknesses. (**a**) Lens group 2 rotated to the forward position within the objective. This position is used for the thinnest coverslips. (**b**) Lens group 2 rotated to the rearward position within the objective. This position is used for the thickest coverslips.

Working Distance. This is the distance between the objective front lens and the top of the cover glass when the specimen is in focus. In most instances, the working distance of an objective decreases as magnification increases. Working distance values are not included on all objectives and their presence varies, depending on the manufacturer. Common abbreviations are L, LL, LD, and LWD (long working distance); ELWD (extralong working distance); SLWD (superlong working distance), and ULWD (ultralong working distance).

Objective Screw Threads. The mounting threads on almost all objectives are sized to standards of the Royal Microscopical Society (RMS) for universal compatibility. This standard specifies a mounting thread that is 20.32 mm in diameter whose a pitch is 0.706; it is currently used to produce infinity-corrected objectives by Olympus and Zeiss. Leica and Nikon broke from the standard when they introduced of new infinity-corrected objectives that have a wider mounting thread size; thus Leica and Nikon objectives are usable only on their own microscopes. Abbreviations commonly used are RMS (Royal Microscopical Society objective thread), M25 (metric 25-mm objective thread), and M32 (metric 32-mm objective thread).

Immersion Medium. Most objectives are designed to image specimens where air is the medium between the objective and the cover glass. To attain higher working numerical apertures, many objectives are designed to image the specimen through another medium that reduces refractive index differences between glass and the imaging medium. High-resolution plan apochromat objectives can achieve numerical apertures up to 1.40 when the immersion medium is a special oil whose refractive index is 1.51. Other common immersion media are water and glycerin. Objectives designed for special immersion media usually have a color-coded ring inscribed around the circumference of the objective barrel, as listed in Table 3 and described next.

Table 3. Color-coded Rings on Microscope Objectives

Immersion Color Code[a]	Immersion Type
Black	Oil immersion
Orange	Glycerol immersion
White	Water immersion
Red	Special
Magnification Color Code[b]	**Magnification**
Black	1X, 1.25X
Brown	2X, 2.5X
Red	4X, 5X
Yellow	10X
Green	16X, 20X
Turquoise blue	25X, 32X
Light blue	40X, 50X
Cobalt (dark) blue	60X, 63X
White (cream)	100X

[a] Narrow colored ring located near the specimen end of objective.
[b] Narrow band located closer to the mounting thread than the immersion code.

Color Codes. Many microscope manufacturers label their objectives with color codes, according to a DIN standard, to help in rapid identification of the magnification. For example, the dark blue color code on the objective illustrated in Fig. 9 indicates that linear magnification is 60X. This is very helpful when you have a nosepiece turret that contains five or six objectives and you must quickly select a specific magnification. Some specialized objectives have an additional color code that indicates the type of immersion medium necessary to achieve the optimum numerical aperture. Immersion lenses intended for use with oil have a black color ring, and those intended for use with glycerin have an orange ring. Objectives for imaging living organisms in aqueous media, designated *water immersion* objectives, have a white ring, and highly specialized objectives for unusual immersion media often have an engraved red ring. Table 3 lists current magnification and imaging media color codes used by most manufacturers.

Specialized Optical Properties. Microscope objectives often have design parameters that optimize performance under certain conditions. For example, special objectives are designed for polarized illumination (signified by the abbreviations P, Po, Pol, or SF, and/or all barrel engravings are painted red), phase contrast (PH, and/or green barrel engravings), differential interference contrast (DIC), and many other abbreviations for additional applications. The apochromat objective illustrated in Fig. 9 is optimized for DIC photomicrography, and this is indicated on the barrel. The capital H beside the DIC marking indicates that the objective must be used with a specific DIC prism optimized for high-magnification applications.

Some applications do not require objectives designed to be corrected for cover glass thickness. These include objectives used to observe uncovered specimens in reflected light, metallurgical specimens, integrated circuit inspection, micromachinery, biological smears, and

other applications that require observing of uncovered objects. Other common abbreviations found on microscope objective barrels, which are useful in identifying specific properties, are listed in Table 4 (17).

Optical Aberrations

Lens errors or aberrations in optical microscopy are caused by artifacts that arise from the interaction of light with glass lenses (2–5, 19–24). There are two primary causes of aberration: (1) geometrical or spherical aberrations that are related to the spherical nature of the lens and approximations used to obtain the Gaussian lens equation; and (2) chromatic aberrations that arise from variations in the refractive index of the wide range of frequencies in visible light.

In general, optical aberrations induce faults in the features of an image being observed through a microscope. These artifacts were first addressed in the eighteenth century when physicist John Dollond discovered that chromatic aberration was reduced or corrected by using a combination of two different types of glass (flint and crown) to fabricate lenses (13). Later, during the nineteenth century, achromatic objectives that had high numerical aperture were developed, although the lenses still had geometrical problems. Modern glass formulations coupled with advanced grinding and manufacturing techniques have all but eliminated most aberrations from today's microscope objectives, although careful attention must still be paid to these effects, especially when conducting quantitative high-magnification video microscopy and photomicrography (23).

Table 4. Specialized Objective Designations[a]

Abbreviation	Type
Phase, PHACO, PC, Ph 1, 2, 3, etc.	Phase-contrast, using phase condenser annulus 1, 2, 3, etc.
DL, DM, PLL, PL, PM, PH, NL, NM, NH	Phase-contrast: dark low, dark medium, positive low, positive low low, positive medium, positive high contrast (regions that have higher refractive index appear darker); negative low, negative medium, negative high contrast (regions that have higher refractive index appear lighter)
P, Po, Pol, SF	Strain-free, low birefringence, for polarized light
U, UV, Universal	UV transmitting (down to approx. 340 nm), for UV-excited epifluorescence
M	Metallographic (no coverslip)
NC, NCG, OD	No coverslip, no coverglass, ohne decke (German designation for no coverslip)
EPI	Surface illumination (specimen illuminated through objective lens), as contrasted to dia- or transillumination
TL	Transmitted light
BBD, HD, B\D	For use in bright or dark field (hell, dunkel)
D	Dark field
H	Designed primarily for heating stage
U, UT	Designed to be used with universal stage (magnification/NA applied for use with glass hemisphere; divide both values by 1.51 when hemisphere is not used)
DI; MI; TI Michelson	Interferometry; noncontact; multiple beam (Tolanski)

[a]Many of the designation codes are manufacturer-specific.

Spherical Aberration

These artifacts occur when light waves that pass through the periphery of a lens are not brought into identical focus with those that pass closer to the center. Waves that pass near the center of the lens are refracted only slightly, whereas waves that pass near the periphery are refracted to a greater degree and result in producing of different focal points along the optical axis. This is one of the most serious resolution artifacts because the image of the specimen is spread out along the optical axis rather than sharply focused. Spherical aberrations are very important in the resolution of the lens because they affect the coincident imaging of points along the optical axis and degrade the performance of the lens, which seriously affects specimen sharpness and clarity. These lens defects can be reduced by limiting the outer edges of the lens from exposure to light by using diaphragms and also aspherical lens surfaces within the system. The highest quality modern microscope objectives address spherical aberrations in a number of ways, including special lens-grinding techniques, additional lens elements of different curvatures, improved glass formulations, and better control of optical pathways.

Chromatic Aberration

This type of optical defect results from the fact that white light is composed of numerous wavelengths. When white light passes through a convex lens, the component wavelengths are refracted according to their frequency. Blue light is refracted to the greatest extent followed by green and red light, a phenomenon commonly referred to as dispersion. The inability of the lens to bring all of the colors into common focus results in a slightly different image size and focal point for each predominant wavelength group. This leads to color fringes that surround the image.

Lens corrections were first attempted in the latter part of the eighteenth century when Dollond, Lister, and others devised ways to reduce longitudinal chromatic aberration (13). By combining crown glass and flint glass (each type has a different dispersion of refractive index), they succeeded in bringing the blue rays and the red rays to a common focus, near but not identical to the green rays. This combination is termed a lens doublet; each lens has a different refractive index and dispersive properties. Lens doublets are also known as achromatic lenses or

achromats for short, derived from the Greek terms *a* meaning without and *chroma* meaning color. This simple form of correction allows the image points at 486 nm in the blue region and 656 nm in the red region to coincide. This is the most widely used objective lens and is commonly found on laboratory microscopes. Objectives that do not carry a special inscription stating otherwise are likely to be achromats. Achromats are satisfactory objectives for routine laboratory use, but because they are not corrected for all colors, a colorless specimen detail is likely to show a pale green color at best focus (the so-called secondary spectrum) in white light.

A proper combination of lens thickness, curvature, refractive index, and dispersion allows the doublet to reduce chromatic aberration by bringing two of the wavelength groups into a common focal plane. If fluorspar is introduced into the glass formulation used to fabricate the lens, then the three colors red, green, and blue can be brought into a single focal point that results in a negligible amount of chromatic aberration (23). These lenses, known as apochromatic lenses, are used to build very high-quality chromatic aberration-free microscope objectives. Modern microscopes use this concept, and today it is common to find optical lens triplets made of three lens elements cemented together, especially in higher quality objectives. For chromatic aberration correction, a typical 10X achromat microscope objective is built of two lens doublets (Fig. 8a). Apochromat objectives usually contain two lens doublets and a lens triplet (Fig. 8c) for advanced correction of both chromatic and spherical aberrations.

In addition to longitudinal (or axial) chromatic aberration correction, apochromat objectives also exhibit another chromatic defect. Even when the three main colors are brought to identical focal planes axially, the point images of details near the periphery of the field of view are not the same size; for example, the blue image of a detail is slightly larger than the green image or the red image in white light; this causes color ringing of specimen details at the outer regions of the field of view (23), a defect known as lateral chromatic aberration or chromatic difference of magnification. The compensating eyepiece, whose chromatic difference of magnification is just the opposite of that of the objective, is used to correct for lateral chromatic aberration. Because this defect is also found in higher magnification achromats, compensating eyepieces are frequently used for such objectives, too. Indeed, many manufacturers design their achromats so that they have a standard lateral chromatic error and use compensating eyepieces for all their objectives. Such eyepieces often carry the inscription K, C, or Compens. As a result, compensating eyepieces have built-in lateral chromatic error and are not, in themselves, perfectly corrected.

Coverslip Correction

It is possible for the user to introduce spherical aberration into a well-corrected system inadvertently (2,23). For example, when using high magnification and high numerical aperture dry objectives (NA = 0.85–0.95), the correct thickness of the cover glass (suggested 0.17 mm) is critical; hence, a correction collar is included on such objectives to enable adjustment for incorrect cover glass thickness. Similarly, inserting of accessories in the light path of finite-tube-length objectives may introduce aberrations that are apparent when the specimen is refocused, unless such accessories have been properly designed with additional optics. Fig. 10 illustrates how internal lenses operate in an objective designed for coverslip correction.

Other Geometrical Aberrations

These include a variety of effects including astigmatism, field curvature, and comatic aberrations that are corrected by proper lens fabrication.

Curvature of field in the image is an aberration that is familiar to most experienced microscopists. This artifact is the natural result of using lenses that have curved surfaces. When visible light is focused through a curved lens, the image plane produced by the lens is curved. When the image is viewed in the eyepieces (oculars) of a microscope, it either appears sharp and crisp in the center or on the edges of the view field, but not both. Normally, this is not a serious problem when the microscopist is routinely scanning samples to observe their various features. It is a simple matter to use the fine focus knob to correct small deficiencies in specimen focus. However, in photomicrography, field curvature can be a serious problem, especially when a portion of the photomicrograph is out of focus.

Modern microscopes correct field curvature by using specially designed flat-field objectives. These specially corrected objectives, named plan or plano, are the most common type of objective in use today. Plan objectives are also corrected for other optical artifacts such as spherical and chromatic aberrations. In a plan objective that also has been mostly corrected for chromatic aberration, the objective is referred to as a plan achromat. This is also the case for fluorite and apochromatic objectives, which have the modified names, plan fluorite and plan apochromat.

Adding field curvature lens corrections to an objective that has already been corrected for optical aberrations can often add a significant number of lens elements to the objective. For example, the typical achromat objective has two lens doublets and a hemispherical lens, a total of five lens elements. In contrast, a comparable plan achromat objective has three doublets and three single lenses, a total of nine lens elements, which makes it considerably more difficult to fabricate. As we have seen, the number of lens elements increases as lenses are corrected for spherical errors and for chromatic and field curvature aberrations. Unfortunately, as the number of lens elements increases, so does the cost of the objective.

Sophisticated plan apochromatic objectives that are corrected for spherical, chromatic, and field curvature aberrations can contain as many as eighteen to twenty separate lens elements; this makes these objectives the most expensive and difficult to manufacture. Plan apochromatic objectives can cost upward of $3,000 to $5,000 each for high-magnification units that also have high numerical apertures. For most photomicrographic applications, however, it is not absolutely necessary to have the best correction, although this is heavily dependent upon the purpose, the specimen, and the

desired magnification range. When cost is important (when isn't it?), it is often wise to select more modestly priced plan fluorite objectives that have a high degree of correction, especially the more modern versions. These objectives provide crisp and sharp images that have minimal field curvature and are sufficient for most photomicrograpic applications.

Field curvature is very seldom totally eliminated, but it is often difficult to detect edge curvature when using the most plan-corrected objectives, and it does not show up in photomicrographs (19,23). This artifact is more severe at low magnifications and can be a problem when using stereomicroscopes. Manufacturers have struggled for years to eliminate field curvature in the large objectives found in stereomicroscopes. In the past ten years, companies such as Leica, Nikon, Olympus, and Zeiss have made great strides in the quality of optics used to build stereomicroscopes and, although the artifacts and aberrations have not been totally eliminated, high-end models can now produce superb photomicrographs.

Comatic aberrations are similar to spherical aberrations, but they are encountered mainly in off-axis objects and are most severe when the microscope is out of alignment (23). In this instance, the image of a point is asymmetrical and results in a comet-like (hence, the term coma) shape. The comet shape may have its *tail* pointing toward the center of the field of view or away, depending on whether the comatic aberration has a positive or negative value. A coma may occur near the axial area of the light path and/or the more peripheral area. These aberrations are usually corrected along with spherical aberrations by designing lens elements of various shapes to eliminate this error. Objectives that are designed to yield excellent images for wide field of view eyepieces have to be corrected for coma and astigmatism using a specially designed multielement optic in the tube lens to avoid these artifacts at the periphery of the field of view.

Astigmatic aberrations are similar to comatic aberrations; however, these artifacts depend more strongly on the obliquity of the light beam (23). This defect is found at the outer portions of the field of view of uncorrected lenses. The off-axis image of a specimen point appears as a line instead of a point. What is more, depending on the angle of the off-axis rays that enter the lens, the line image may be oriented in either of two different directions, tangentially or radially. Astigmatic errors are usually corrected by designing the objectives to provide precise spacing of individual lens elements as well as appropriate lens shapes and index of refraction. Astigmatism is often corrected in conjunction with the correction of field curvature aberrations.

Eyepieces (Oculars)

Eyepieces work in combination with microscope objectives to magnify the intermediate image further, so that specimen details can be observed. Ocular, an alternative name for eyepiece, has been widely used in the literature, but to maintain consistency during this discussion, we will refer to all oculars as eyepieces. Best results in microscopy require using objectives in combination with eyepieces that are appropriate to the correction and type of objective. Inscriptions on the side of the eyepiece describe its particular characteristics and function.

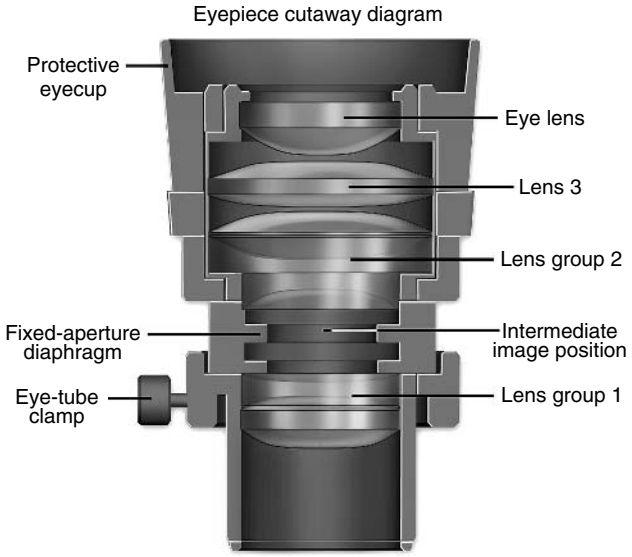


Figure 11. Cutaway diagram of a typical periplan eyepiece. The fixed-aperture diaphragm is positioned between lens group 1 and lens group 2, where the intermediate image is formed. The eyepiece has a protective eyecup that makes viewing the specimen more comfortable for the microscopist.

The eyepiece illustrated in Fig. 11 is inscribed with UW (not illustrated), which is an abbreviation for ultrawide view field. Often, eyepieces will also have an H designation, depending on the manufacturer, to indicate a high eye point focal point that allows microscopists to wear glasses while viewing samples. Other common inscriptions often found on eyepieces include WF for wide field, UWF for ultrawide field, SW and SWF for super-wide field, HE for high eye point, and CF for eyepieces intended for use with CF corrected objectives (2,3,5,19,23,24). As discussed before, compensating eyepieces are often inscribed with K, C, Comp, or Compens, as well as the magnification. Eyepieces used with flat-field objectives are sometimes labeled Plan-Comp. Magnification factors of common eyepieces range from 5–25X, and usually contain an inscription, such as A/24, which indicates that the field number is 24, referring to the diameter (in millimeters) of the fixed diaphragm in the eyepiece. Many eyepieces also have a focus adjustment and a thumbscrew that allows fixing their positions. Now, manufactures often produce eyepieces that have rubber eyecups that position the eyes at the proper distance from the front lens and block room light from reflecting from the lens surface and interfering with the view.

There are two major types of eyepieces that are grouped according to lens and diaphragm arrangement: the negative eyepiece that has an internal diaphragm between the lenses and the positive eyepiece that has a diaphragm below the lenses of the eyepiece. Negative eyepieces have two lenses or lens groups: the upper lens, which is closest to the observer's eye, is called the eye lens and the lower lens (beneath the diaphragm) is often termed the field lens. In their simplest form, both lenses are plano-convex, they have convex sides facing the

specimen. Approximately midway between these lenses is a fixed circular opening or internal diaphragm which, by its size, defines the circular field of view that is observed by looking into the microscope. The simplest kind of negative eyepiece, or Huygenian eyepiece, is used on most routine microscopes fitted with achromatic objectives. Although the Huygenian eye and field lenses are not well corrected, their aberrations tend to cancel each other out. More highly corrected negative eyepieces have two or three lens elements cemented and combined together to make the eye lens. If an unknown eyepiece carries only the magnification inscribed on the housing, it is most likely to be a Huygenian eyepiece, best suited for use with achromatic objectives of 5–40X magnification.

The other main type of eyepiece is the positive eyepiece, commonly known as the *Ramsden* eyepiece, that has a diaphragm below its lenses. This eyepiece has an eye lens and field lens that are also plano-convex, but the field lens is mounted so that the curved surface faces toward the eye lens. The front focal plane of this eyepiece lies just below the field lens at the level of the eyepiece fixed diaphragm, and makes this eyepiece readily adaptable for mounting graticules. To provide better correction, the two lenses of the Ramsden eyepiece may be cemented together.

Simple eyepieces such as the Huygenian and Ramsden and their achromatized counterparts will not correct for a residual chromatic difference of magnification in the intermediate image, especially when used in combination with high magnification achromatic objectives and fluorite or apochromatic objectives. To remedy this in finite microscopy systems, manufacturers produce compensating eyepieces that introduce an equal, but opposite, chromatic error in the lens elements. Compensating eyepieces may be either of the positive or negative type and must be used at all magnifications with fluorite, apochromatic, and all variations of plan objectives (they can also be used to advantage with achromatic objectives of 40X and higher).

In recent years, modern microscope objectives have their correction for chromatic difference of magnification either built into the objectives themselves (Olympus and Nikon) or corrected in the tube lens (Leica and Zeiss), thus eliminating the need for compensation correction of the eyepieces.

Compensating eyepieces play a crucial role in helping to eliminate residual lateral chromatic aberrations inherent in the design of highly corrected objectives. Hence, it is preferable for the microscopist to use the compensating eyepieces designed by a particular manufacturer to accompany that manufacturer's higher corrected objectives. Using an incorrect eyepiece whose apochromatic objective is designed for a finite (160 or 170 mm) tube length microscope results in dramatically increased contrast, red fringes on the outer diameters and blue fringes on the inner diameters of specimen detail. Additional problems arise from a limited flatness of the view field in simple eyepieces, even those corrected with eye lens doublets.

More advanced eyepiece designs resulted in the periplan eyepiece (Fig. 11), which contains seven lens elements that are cemented into a doublet, a triplet, and two individual lenses. Design improvements in periplan eyepieces lead to better correction for residual lateral chromatic aberration, increased flatness of field, and general overall better performance when used with higher power objectives.

Modern microscopes feature vastly improved plan-corrected objectives in which the primary image has much less curvature of field than older objectives. In addition, most microscopes now feature much wider body tubes that have greatly increased the size of intermediate images. To address these new features, manufacturers now produce wide eye-field eyepieces that increase the viewable area of the specimen by as much as 40%. Because the strategies of eyepiece-objective correction techniques vary from manufacturer to manufacturer, it is very important (as stated before) to use only eyepieces recommended by a specific manufacturer for use with their objectives.

Our recommendation is to choose the objective carefully first, then purchase an eyepiece that is designed to work in conjunction with the objective. When choosing eyepieces, it is relatively easy to differentiate between simple and more highly compensating eyepieces. Simple eyepieces such as the Ramsden and Huygenian (and their more highly corrected counterparts) will appear to have a blue ring around the edge of the eyepiece diaphragm when viewed through the microscope or held up to a light source. In contrast, more highly corrected compensating eyepieces have a yellow–red–orange ring around the diaphragm under the same circumstances. Modern noncompensating eyepieces are fully corrected and show no color. Most of the modern microscopes have all corrections in the objectives themselves or have a final correction in the tube lens. Such microscopes do not need compensating eyepieces.

The properties of several common commercially available eyepieces are listed according to type in Table 5 (19,23). The three major types of eyepieces listed in Table 5 are finder, wide field, and superwide field. The terminology used by various manufacturers can be very confusing, and careful attention should be paid to their sales brochures and microscope manuals to ensure that the correct eyepieces are being used with a specific objective. In Table 5, the abbreviations that designate wide-field and superwide-field eyepieces are coupled to their design for high eye point, and are WH and SWH, respectively. The magnifications are either 10X or 15X and the field numbers range from 14 to 26.5, depending upon the application. The diopter adjustment, used to adjust for differences between the eyes, is approximately the same for all eyepieces, and many also contain either a photomask or micrometer graticule.

Light rays emanating from the eyepiece intersect at the exit pupil or eye point (Ramsden disk), where the front of the microscopist's eye should be placed to see the entire field of view (usually 8–10 mm above the eye lens). By increasing the magnification of the eyepiece, the eye point is drawn closer to the upper surface of the eye lens, making it much more difficult for microscopists to use, especially if they wear eyeglasses. Specially designed high-eye-point eyepieces have been manufactured that feature eye-point viewing distances that approach 20–25 mm above the

Table 5. Properties of Commercial Eyepieces

Eyepiece Type	Finder Eyepieces			Superwide Field Eyepieces		Wide Field Eyepieces	
Descriptive abbreviation	PSWH 10X	PWH 10X	35 SWH 10X	SWH 10X H	CROSSWH 10X H	WH 15X	WH 10X H
Field number	26.5	22	26.5	26.5	22	14	22
Diopter adjustment	$-8 \sim +2$	$-8 \sim +2$	$-8 \sim +2$	$-8 \sim +2$	$-8 \sim +2$	$-8 \sim +2$	$-8 \sim +2$
Remarks	$3\frac{1}{4} \times 4\frac{1}{4}$ photomask	$3\frac{1}{4} \times 4\frac{1}{4}$ photomask	35-mm photomask	Diopter correction	Diopter correction crossline		Diopter correction
Diameter of micrometer graticule	—	—	—	—	—	24	24

surface of the eye lens. These improved eyepieces have larger diameter eye lenses that contain more optical elements and usually feature improved flatness of field. Such eyepieces are often designated by the inscription H somewhere on the eyepiece housing, either alone or in combination with other abbreviations, as discussed earlier. We should mention that high-eye-point eyepieces are especially useful for microscopists who wear eyeglasses to correct for near- or farsightedness, but they do not correct for several other visual defects, such as astigmatism. Today, high-eye-point eyepieces are very popular, even among people who do not wear eyeglasses, because the large eye clearance reduces fatigue and makes viewing images through the microscope much more comfortable.

At one time, eyepieces were available in a wide range of magnifications extending from 6.3–30X and sometimes even higher for special applications. These eyepieces are very useful for observation and photomicrography using low-power objectives. Unfortunately, for higher power objectives, the problem of empty magnification (magnification without increased clarity) becomes important when using very high magnification eyepieces and these should be avoided. Today, most manufacturers restrict their eyepiece offerings to those in the 10X to 20X ranges. The diameter of the view field in an eyepiece is expressed as a *field of view number* or *field number* (FN), as discussed before. Information about the field number of an eyepiece can yield the real diameter of the object view field by using this formula (23):

$$\text{View-field diameter} = \text{FN}/(\text{M}_\text{O} \times \text{M}_\text{T}) \qquad (2)$$

where FN is the field number in millimeters, M_O is the magnification of the objective, and M_T is the tube lens magnification factor (if any). Applying this formula to the superwide-field eyepiece listed in Table 5, we arrive at the following for a 40X objective whose tube lens magnification is 1.25:

$$\text{View-field diameter} = 26.5/(40 \times 1.25) = 0.53 \text{ mm}. \qquad (3)$$

Table 1 lists the view-field diameters across the common range of objectives that would occur using this eyepiece.

Care should be taken in choosing eyepiece/objective combinations to ensure the optimal magnification of specimen detail without adding unnecessary artifacts. For instance, to achieve a magnification of 250X, the microscopist could choose a 25X eyepiece coupled to a 10X objective. An alternative choice for the same magnification would be a 10X eyepiece coupled to a 25X objective. Because the 25X objective has a higher numerical aperture (approximately 0.65) than the 10X objective (approximately 0.25) and considering that numerical aperture values define an objective's resolving power, it is clear that the latter would be the better choice. If photomicrographs of the same view field were made with each objective/eyepiece combination described, it would be obvious that the 10X eyepiece/25X objective duo would produce photomicrographs that excelled in specimen detail and clarity compared to the alternative combination.

The numerical aperture of the objective/condenser system defines the *range of useful magnification* for an objective/eyepiece combination (19,22–24). There is a minimum magnification necessary for the detail in an image to be resolved, and this value is usually rather arbitrarily set at 500 times the numerical aperture ($500 \times \text{NA}$). At the other end of the spectrum, the maximum useful magnification of an image is usually set at 1000 times the numerical aperture ($1000 \times \text{NA}$). Magnifications higher than this value will yield no further useful information or finer resolution of image detail and will usually lead to image degradation. Exceeding the limit of useful magnification causes the image to suffer from the phenomenon of *empty magnification* (19), where increasing magnification through the eyepiece or intermediate tube lens only causes the image to become more magnified without a corresponding increase in detail resolution. Table 6 lists the common objective/eyepiece combinations that fall into the range of useful magnification.

Eyepieces can be adapted for measuring by adding a small circular disk-shaped glass graticule at the plane of the fixed-aperture diaphragm of the eyepiece. Graticules usually have markings, such as a measuring rule or grid, etched onto the surface. Because the graticule lies in the same conjugate plane as the fixed-eyepiece diaphragm, it appears in sharp focus superimposed on the image of the specimen. Eyepieces, that use graticules usually contain

Table 6. Range of Useful Magnification (500–1000 × NA of Objective)

Objective (NA)	Eyepieces 10X	12.5X	15X	20X	25X
2.5X (0.08)	—[a]	—[a]	—	x[b]	x[b]
4X (0.12)	—	—	x	x	x
10X (0.35)	—	x	x	x	x
25X (0.55)	x	x	x	x	—
40X (0.70)	x	x	x	—	—
60X (0.95)	x	x	x	—	—
100X (1.42)	x	x	—	—	—

[a] —: Not recommended.
[b] x: Useful range.

a focusing mechanism (helical screw or slider) that allows bringing the image of the graticule into focus. A stage micrometer is needed to calibrate the eyepiece scale for each objective.

Condenser Systems

The substage condenser gathers light from the microscope light source and concentrates it into a cone of light that illuminates the specimen with parallel beams of uniform intensity from all azimuths across the entire view field. It is critical to adjust the condenser light cone properly to optimize the intensity and angle of light that enters the objective front lens. Each time an objective is changed, the substage condenser aperture iris diaphragm must be adjusted correspondingly to provide the proper light cone for the numerical aperture of the new objective.

A simple two-lens Abbe condenser is illustrated in Fig. 12. In this figure, light from the microscope illumination source passes through the aperture or condenser diaphragm, located at the base of the condenser, and is concentrated by internal lens elements, which then project light through the specimen in parallel bundles from every azimuth. The size and numerical aperture of the light cone are determined by adjusting of the aperture diaphragm. After passing through the specimen (on the microscope slide), the light diverges to an inverted cone of the proper angle (2θ in Fig. 12) to fill the front lens of the objective (2,18–24).

Aperture adjustment and proper focusing of the condenser are critically important in realizing the full potential of the objective. Specifically, appropriate use of the adjustable aperture iris diaphragm (incorporated into the condenser or just below it) is most important in securing correct illumination, contrast, and depth of field. Opening and closing this iris diaphragm controls the angle of illuminating rays (and thus the aperture) which pass through the condenser, through the specimen, and then into the objective. Condenser height is controlled by a rack and pinion gear system that allows adjusting

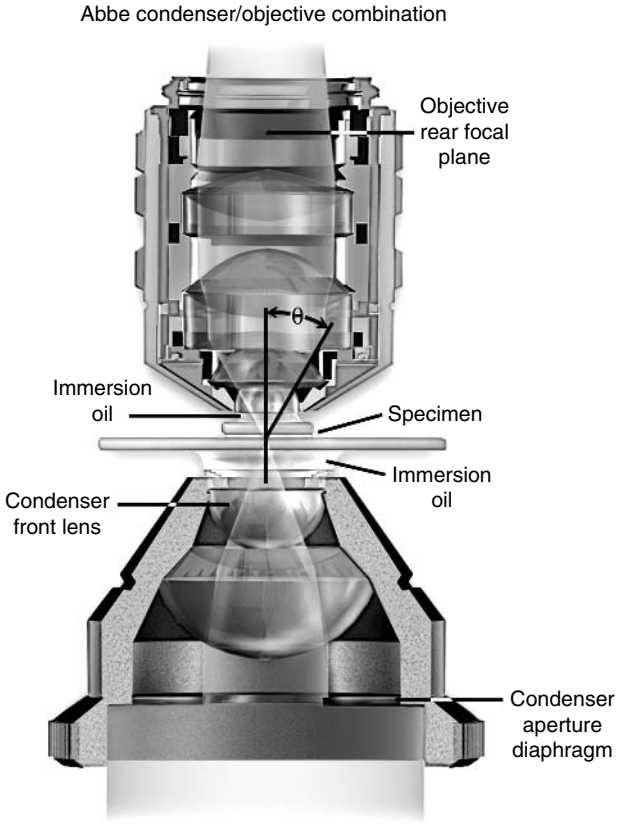


Figure 12. Condenser/objective configuration for optical microscopy. The Abbe two-lens condenser illustrated shows ray traces through the optical train of the microscope. The aperture diaphragm restricts light that enters the condenser before it is refracted by the condenser lens system onto the specimen. Immersion oil is used in the contact beneath the underside of the slide and the condenser top lens and also between the objective and coverslip. The objective aperture angle (θ) controls the amount of light that enters the objective.

the condenser focus to illuminate the specimen properly. Correctly positioning the condenser relative to the cone of illumination and focusing on the specimen are critical in quantitative microscopy and optimum photomicrography. Care must be taken to ensure that the condenser aperture is opened to the correct position with respect to the objective's numerical aperture. When the aperture is opened too much, stray light generated by refraction of oblique light rays from the specimen can cause glare and lower the overall contrast. On the other hand, when the aperture is closed too far, the illumination cone is insufficient to provide adequate resolution, and the image is distorted due to refraction and diffraction from the specimen.

The simplest and least corrected (also the least expensive) condenser is the Abbe condenser that, in its simplest form, has two optical lens elements which produce an image of the illuminated field diaphragm that is not sharp and is surrounded by blue and red color at the edges (Table 7). As a result of little optical correction, the Abbe condenser is suited mainly for routine observation using objectives of modest numerical aperture

Table 7. Condenser Aberration Corrections

Condenser Type	Aberrations Corrected	
	Spherical	Chromatic
Abbe	—(No)	—
Aplanatic	x (Yes)	—
Achromatic	—	x
Aplanatic-Achromatic	x	x

and magnification. The primary advantages of the Abbe condenser are the wide cone of illumination that the condenser can produce as well as its ability to work with long working distance objectives. Manufacturers supply most microscopes with an Abbe condenser as the default, and these condensers are real workhorses for routine laboratory use.

The next highest level of condenser correction is split between the aplanatic and achromatic condensers that are corrected exclusively for either spherical (aplanatic) or chromatic (achromatic) optical aberrations. Achromatic condensers typically contain four lens elements and have a numerical aperture of 0.95, the highest attainable without requiring immersion oil (5; see Table 7). This condenser is useful for both routine and critical laboratory analysis using dry objectives and also for black-and-white or color photomicrography.

The numerical aperture performance, which will be necessary to provide an illumination cone adequate for the objectives, is critical factor in choosing substage condensers (see Table 8). The capability of the condenser numerical aperture should be equal to or slightly less than that of the highest objective numerical aperture. Therefore, if the largest magnification objective is an oil-immersion objective whose numerical aperture is 1.40, then the substage condenser should also have an equivalent numerical aperture to maintain the highest system resolution. In this case, immersion oil would have to be applied between the condenser top lens in contact with the underside of the microscope slide to achieve the intended numerical aperture (1.40) and resolution. Failure to use oil will restrict the highest numerical aperture of the system to 1.0, the highest obtainable using air as the imaging medium (2,5,17–24).

Aplanatic condensers are well corrected for spherical aberration (green wavelengths) but not for chromatic aberration (Table 7). These condensers feature five lens elements and can focus light in a single plane. Aplanatic condensers produce excellent black-and-white photomicrographs by green light generated by a laser source or by using an interference filter that has tungsten-halide illumination.

The highest level of correction for optical aberration is incorporated in the aplanatic–achromatic condenser (2,5). This condenser is well corrected for both chromatic and spherical aberrations and is the condenser of choice for use in critical color photomicrography by white light. A typical aplanatic–achromatic condenser features eight internal lens elements cemented into two doublets and four single lenses.

Engravings on the condenser housing include its type (achromatic, aplanatic, etc.), the numerical aperture, and a graduated scale that indicates the approximate adjustment (size) of the aperture diaphragm. As mentioned before, condensers whose numerical apertures are higher than 0.95 perform best when a drop of oil is applied to the upper lens in contact with the undersurface of the specimen slide. This ensures that oblique light rays from the condenser are not reflected from underneath the slide, but are directed into the specimen without deviation. In practice, this can become tedious and is not commonly done in routine microscopy, but it is essential when working at high resolutions and for accurate photomicrography using high-power (and numerical aperture) objectives.

Table 8. Substage Condenser Applications

Condenser Type	Bright-field	Dark-field	Phase-contrast	DIC	Polarizing
Achromat/ aplanat NA 1.3	• [10–100X]				
Achromat swing-out NA 0.90	• [4–100X]				
Low-power NA 0.20	• [1–10X]				
Phase-contrast Abbe NA 1.25	•	• [up to NA 0.65]	• [10–100X]		
Phase-contrast achromat NA 0.85	•	• [up to NA 0.70]	• [4–100X]		
DIC universal achromat/aplanat	•	• [up to NA 0.70]	• [10–100X]	• [10X, 20X, 100X]	
Dark-field, dry NA 0.80 ~ 0.95		• [4–40X]			
Dark-field, oil NA 1.20 ~ 1.43		• [4–100X]			
Strain-free achromat swing-out NA 0.90	•				• [4–100X]

Another important consideration is the thickness of the microscope slide, which is as crucial to the condenser as coverslip thickness is to the objective. Most commercial producers offer slides that range in thickness between 0.95 and 1.20 mm: the most common is very close to 1.0 mm. A microscope slide 1.20 mm thick is too thick for most high numerical aperture condensers that tend to have a very short working distance. This does not greatly matter for routine specimen observation, but the results can be devastating in precision photomicrography. We recommend choosing microscope slides that are 1.0 ± 0.05 mm thick, and thoroughly cleaning them before use.

When the objective is changed, for example, from 10X to 20X, the aperture diaphragm of the condenser must also be adjusted to provide a light cone that matches the numerical aperture of the new objective. There is a small painted arrow or index mark located on the knurled knob or lever that indicates the relative size of the aperture compared to the linear gradation on this condenser housing. Many manufacturers synchronize this gradation to correspond to the approximate numerical aperture of the condenser. For example, if the microscopist has selected a 10X objective of numerical aperture 0.25, then the arrow would be placed next to the value 0.18–0.20 (about 80% of the objective numerical aperture) on the scale inscribed on the condenser housing.

Often, it is not practical to use a single condenser for an entire range of objectives (2–100X) due to the broad range of light cones that must be produced to match objective numerical apertures. For low-power objectives in the range 2–5X, the illumination cone will have a diameter from 6–10 mm, and the high-power objectives (60–100X) need a highly focused light cone only about 0.2–0.4 mm in diameter. For a fixed focal length, it is difficult to achieve this wide range of illumination cones by using a single condenser (18).

In practice, this problem can be solved in several ways. For low-power objectives (less than 10X), it may be necessary to unscrew the top lens of the condenser to fill the field of view with light. Other condensers produced have a flip-top upper lens to accomplish this more readily. Some manufacturers now produce a condenser that flips over completely when used with low-power objectives. Other companies incorporate auxiliary correction lenses in the light path to secure proper illumination for objectives less than 10X, or produce special low-power and low-numerical aperture condensers. When the condenser is used without its top lens, the aperture iris diaphragm is opened wide and the field diaphragm, now visible at the back of the objective, serves as if it were the aperture diaphragm. Flip-top condensers are manufactured in a variety of configurations, and with numerical apertures range from 0.65–1.40. Those condensers that have a numerical aperture value of 0.95 or less are intended for use with dry objectives. Flip-top condensers that have a numerical aperture greater than 0.95 are intended for use with oil-immersion objectives and they must have a drop of oil placed between the underside of the microscope slide and the condenser top lens when examining critical samples.

In addition to the common bright-field condensers discussed before, there are a wide variety of specialized models suited to many different applications. Substage condensers have a great deal of interchangeability among different applications (Table 8). For instance, the DIC universal achromat/aplanat condenser is useful for bright field, dark field, and phase-contrast, in addition to the primary DIC application. Other condensers have similar interchangeability.

Depth of Field and Depth of Focus

When considering resolution in optical microscopy, the major emphasis is placed on point-to-point resolution in the plane perpendicular to the optical axis. Another important aspect of resolution is the *axial resolving power* of an objective, which is measured parallel to the optical axis and is most often referred to as depth of field (2,4,5,22). Axial resolution, like horizontal resolution, is determined only by the numerical aperture of the objective; the eyepiece merely magnifies the details resolved and projected in the intermediate image plane.

Just as in classical photography, depth of field is determined by the distance from the nearest object plane in focus to that of the farthest plane also simultaneously in focus. Depth of field in microscopy is very short and is usually measured in microns. The term *depth of focus*, which refers to image space, is often used interchangeably with depth of field, which refers to object space. This interchange of terms can lead to confusion, especially when the terms are both used specifically for depth of field.

The geometric image plane might be expected to represent an infinitely thin section of the specimen, but even in the absence of aberrations, each image point is spread into a diffraction figure that extends above and below this plane (2,4,17). The Airy disk, discussed in the section on Image Formation, represents a section through the center of the image plane. This increases the effective in-focus depth of the Z-axis Airy disk intensity profile that passes through slightly different specimen planes.

Depth of focus varies with numerical aperture and magnification of the objective, and under some conditions, high numerical aperture systems (usually of higher magnification power) have deeper focus depths than those systems of low numerical aperture, even though the depth of field is less (4). This is particularly important in photomicrography because the film emulsion or digital camera sensor must be exposed at a plane that is in focus. Small errors made by focusing at high magnification are not as critical as those made by using very low magnification objectives. Table 9 presents calculated variations in the depth of field and image depth in the intermediate image plane for a series of objectives of increasing numerical aperture and magnification.

REFLECTED LIGHT MICROSCOPY

Reflected light microscopy, often referred to as incident light, epi-illumination, or metallurgical microscopy, is the method of choice for fluorescence and for imaging specimens that remain opaque even when ground to a thickness

Table 9. Depth of Field and Image Depth[a]

Magnification	Numerical aperture	Depth of Field (μm)	Image depth (mm)
4X	0.10	15.5	0.13
10X	0.25	8.5	0.80
20X	0.40	5.8	3.8
40X	0.65	1.0	12.8
60X	0.85	0.40	29.8
100X	0.95	0.19	80.0

[a]Source: Nikon.

of 30 microns (25). The range of specimens that fall into this category is enormous and includes most metals, ores, ceramics, many polymers, semiconductors (unprocessed silicon, wafers, and integrated circuits), slag, coal, plastics, paint, paper, wood, leather, glass inclusions, and a wide variety of specialized materials (25–28). Because light cannot pass through these specimens, it must be directed onto the surface and eventually returned to the microscope objective by either specular or diffused reflection. As mentioned before, such illumination is most often referred to as episcopic illumination, epi-illumination, or vertical illumination (essentially originating from above), in contrast to diascopic (transmitted) illumination that passes through a specimen. Today, many microscope manufacturers offer models that permit the user to alternate or simultaneously conduct investigations using vertical and transmitted illumination. Reflected light microscopy is frequently the domain of industrial microscopy, especially in the rapidly growing semiconductor arena, and thus represents a most important segment of microscopic studies.

A typical upright-compound, reflected-light (illustrated in Fig. 13) microscope that is also equipped for transmitted light has two eyepiece viewing tubes and often a trinocular tube head for mounting a conventional or digital/video camera system. Standard equipment eyepieces are usually of 10X magnification, and most microscopes are equipped with a nosepiece that can hold four to six objectives. The stage is mechanically controlled by a specimen holder that can be translated in the x and y directions, and the entire stage unit can be moved precisely up and down by a coarse and fine focusing mechanism. Built-in light sources range from 20- and 100-watt tungsten-halide bulbs to higher energy mercury vapor or xenon lamps that are used in fluorescence microscopy. Light passes from the lamp house through a vertical illuminator interposed above the nosepiece but below the underside of the viewing tube head. The specimen's top surface is upright (usually without a coverslip) on the stage facing the objective, which has been rotated into the microscope's optical axis. The vertical illuminator is horizontally oriented at a 90° angle to the optical axis of the microscope and parallel to the table top; the lamp housing is attached to the back of the illuminator. The coarse and fine adjustment knobs raise or lower the stage in large or small increments to bring the specimen into sharp focus.

Another variation of the upright reflected-light microscope is the inverted microscope — of the Le Chatelier design (25). On the inverted stand, the specimen is placed on the stage, and its surface of interest faces downward. The primary advantage of this design is that samples can be easily examined when they are far too large to fit into the confines of an upright microscope. Also, only the side that faces the objectives need be perfectly flat. The objectives are mounted on a nosepiece under the stage, and their front lenses face upward toward the specimen. Focusing is accomplished either by moving the nosepiece or the entire stage up and down. Inverted microscope stands incorporate the vertical illuminator into the body of the microscope. Many types of objectives can be used with inverted reflected-light microscopes, and all modes of reflected light illumination may be possible: bright-field, dark-field, polarized light, differential interference contrast, and fluorescence. Many inverted microscopes have built-in 35-mm and/or large format cameras or are modular to allow attaching such accessories. Because most families of objectives progress in magnification by a factor of 2, some of the instruments include a magnification changer to provide intermediate magnifications. Contrast filters and a variety of reticles are also often built into reflected-light inverted microscopes. Because an inverted microscope is a favorite instrument for metallographers, it is often referred to as a metallograph (25,27). Manufacturers are largely changing to infinity-corrected optics in reflected light microscopes, but there are still thousands of fixed-tube-length microscopes in use whose objectives are corrected for a tube length between 160 and 210 mm.

In the vertical illuminator, light travels from the light source, usually a 12-volt, 50- or 100-watt tungsten halogen lamp, passes through collector lenses, through the variable aperture iris diaphragm opening, and through the opening of a variable and centerable prefocused field iris diaphragm. The light then strikes a partially silvered plane glass reflector (Fig. 13) or strikes a fully silvered periphery of a mirror that has an elliptical opening for dark-field illumination. The plane glass reflector is partially silvered on the glass side that faces the light source and is antireflection coated on the glass side that faces the observation tube in bright-field reflected illumination. Thus, light is deflected downward into the objective. The mirrors are tilted at an angle of 45° degrees to the path of the light that travels along the vertical illuminator.

The light reaches the specimen, which may absorb some of the light and reflect some of the light, either in a specular or diffuse manner. Light that is returned upward can be captured by the objective in accordance with the objective's numerical aperture and then passes through the partially silvered mirror. In infinity-corrected objectives, the light emerges from the objective in *parallel* (from every azimuth) rays and projects an image of the specimen to infinity (25). The parallel rays enter the body tube lens, which forms the specimen image at the plane of the fixed-diaphragm opening in the eyepiece (intermediate image plane). It is important to note, that in these reflected-light systems, the objective serves a dual function: on the way down, as a matching well-corrected condenser properly aligned; on the way up, as

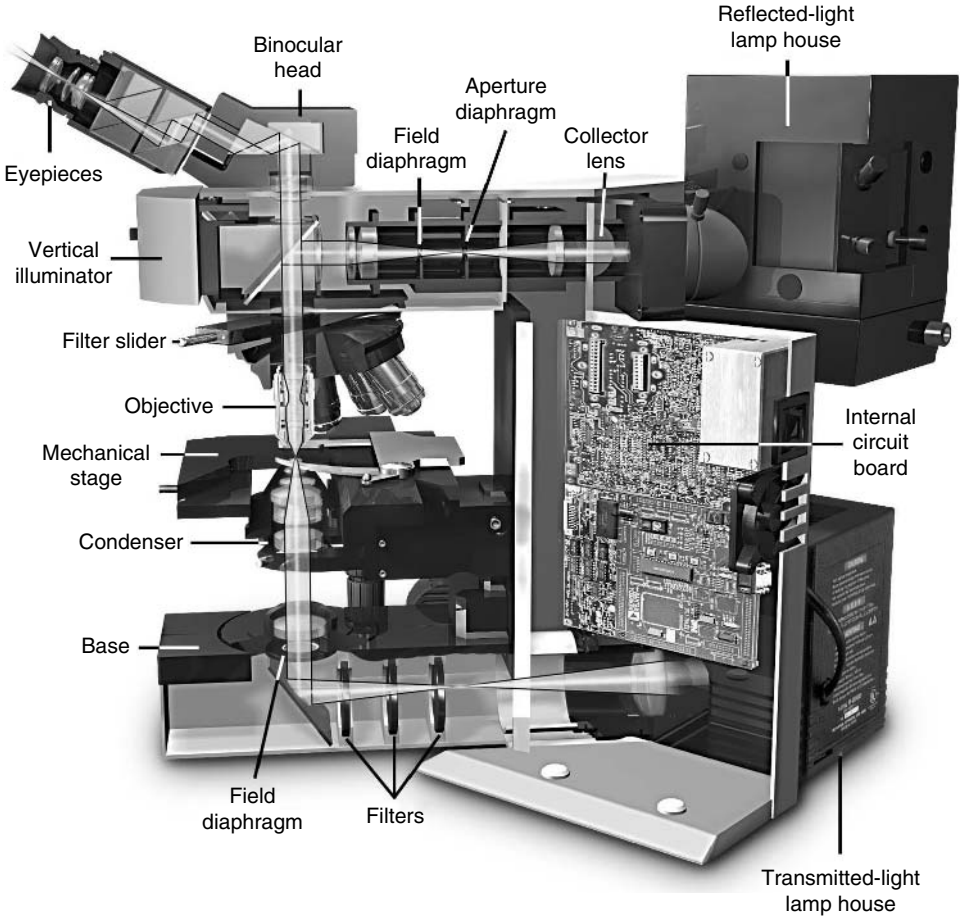


Figure 13. Components of a modern microscope configured for both transmitted and reflected light. This cutaway diagram reveals the ray traces and lens components of the microscope's optical trains. Also illustrated are the basic microscope components including two lamp houses, the microscope built-in vertical and base illuminators, condenser, objectives, eyepieces, filters, sliders, collector lenses, field, and aperture diaphragms.

an image-forming objective in the customary role of an objective that projects the image-carrying rays toward the eyepiece.

Optimal performance is achieved in reflected-light illumination when the instrument is adjusted to produce Köhler illumination. A function of Köhler illumination (aside from providing uniform intensity across the field of view) is to ensure that the objective can deliver excellent resolution and good contrast, even if the source of light is a coil filament lamp.

Some modern reflected-light illuminators are described as universal illuminators because, by using several additional accessories and little or no dismantling, the microscope can easily be switched from one mode of reflected-light microscopy to another. Often, reflectors can be removed from the light path altogether to perform transmitted-light observation. Such universal illuminators may include a partially reflecting plane glass surface (the half-mirror) for bright-field, and a fully silvered reflecting surface that has an elliptical, centrally located clear opening for dark-field observation. The best designed vertical illuminators include condensing lenses to gather and control the light, an aperture iris diaphragm, and a prefocused, centerable iris diaphragm to permit desirable Köhler illumination (2,25).

The vertical illuminator should also provide for inserting of filters for contrast and photomicrography, polarizers, analyzers, and compensator plates for polarized light and differential interference contrast illumination. In vertical illuminators designed for use with infinity-corrected objectives, the illuminator may also include a body tube lens. Affixed to the back end of the vertical illuminator is a lamp house, which usually contains a tungsten-halide lamp. For fluorescence work, the lamp house can be replaced by one that contains a mercury arc lamp. The lamp may be powered by the electronics built into the microscope stand, or in fluorescence, by an external transformer or power supply.

In reflected-light microscopy, absorption and diffraction of the incident light rays by the specimen often lead to readily discernible variations in the image, from black through various shades of gray, or color if the specimen is colored. Such specimens, known as amplitude specimens, may not require special contrast methods or treatment to make their details visible. Other specimens show so little difference in intensity and/or color that their feature

details are extremely difficult to discern and distinguish in bright-field reflected-light microscopy. Such specimens behave much like the phase specimens so familiar in transmitted light work, and they require special treatment or contrast methods that are described in the next section.

Contrast-Enhancing Techniques

Some specimens are considered amplitude objects because they absorb light partially or completely and can thus be readily observed by using conventional bright-field microscopy. Others that are naturally colored or artificially stained by chemical color dyes can also be seen. These stains or natural colors absorb some part of the white light that passes through and transmit or reflect other colors. It is common practice to stain specimens that do not readily absorb light, thus rendering such objects visible to the eye. Often, stains are combined to yield contrasting colors, for example, blue haemotoxylin stain for cell nuclei combined with pink eosin for cytoplasm.

Contrast produced by the absorption of light, brightness, or color has been the classical means of imaging specimens in bright-field microscopy. The ability of a detail to stand out against the background or other adjacent details is a measure of specimen contrast. In a simple formula, contrast can be described as (18)

$$\% \text{ contrast} = [(\mathrm{B_I} - \mathrm{S_I}) \times 100]/\mathrm{B_I}, \qquad (4)$$

where B_I is the intensity of the background and S_I is the specimen intensity. From this equation, it is evident that specimen contrast refers to the relationship between the highest and lowest intensity in the image.

For many specimens in microscopy, especially unstained or living material, contrast is so poor that the object remains essentially invisible regardless of the objective's ability to resolve or clearly separate details. Often, for just such specimens, it is important not to alter them by killing or treatment with chemical dyes or fixatives. This necessity has led microscopists to experiment with contrast-enhancing techniques for more than one hundred years in to improve specimen visibility and to bring more detail to the image. It is a common, but unwise, practice to reduce the condenser aperture diaphragm below the recommended size or to lower the substage condenser to increase specimen contrast. These maneuvers indeed increase contrast, but unfortunately, they also seriously reduce resolution and sharpness.

An early and currently used method of increasing contrast of stained specimens uses color contrast filters, gelatin squares (from Kodak), or interference filters in the light path (18,19). For example, if a specimen is stained red, a green filter will darken the red areas and thus increase contrast. On the other hand, a green filter will lighten any green stained area. Color filters are very valuable aids to specimen contrast, especially when black-and-white photomicrography is the goal. Green filters are particularly valuable for use with achromats, which are spherically corrected for green light, and phase-contrast objectives, which are designed for manipulating wavelength, assume the use of green light, because phase specimens are usually transparent and lack inherent color. Another simple technique for contrast improvement involves selecting a mounting medium whose refractive index is substantially different from that of the specimen. For example, diatoms can be mounted in a variety of contrast-enhancing media such as air whose refractive index is low, or the high refractive index medium Styrax. The difference in refractive indexes improves the contrast of these colorless objects and renders their outlines and markings more visible. The following sections describe many of the more complex techniques used by present-day microscopists to improve specimen contrast.

DARK-FIELD MICROSCOPY

Dark-field illumination requires blocking out the central light rays that ordinarily pass through or around the specimen and allowing only oblique rays to illuminate the specimen. This is a simple and popular method for imaging unstained specimens, which appear as brightly illuminated objects on a dark background. Oblique light rays emanating from a dark-field condenser strike the specimen from every azimuth and are diffracted, reflected, and refracted into the microscope objective (5,18,19,28). This technique is illustrated in Fig. 14. If no specimen is present and the numerical aperture of the condenser is greater than that of the

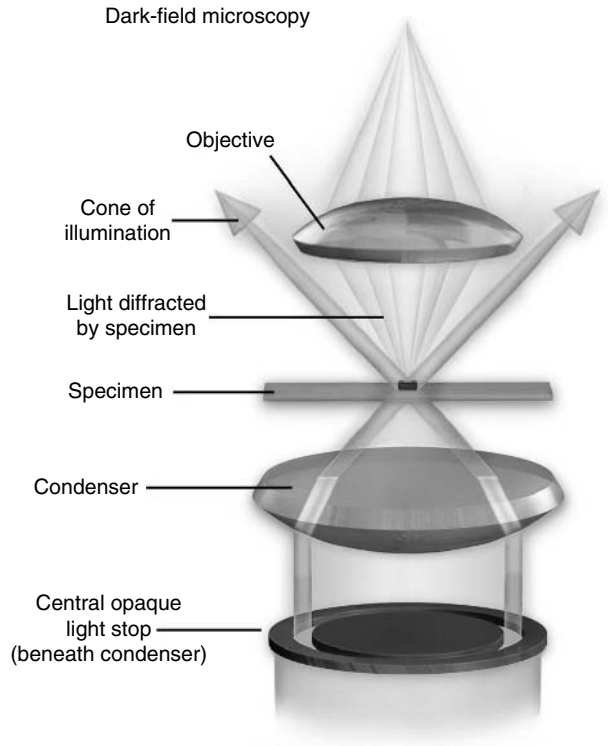


Figure 14. Schematic configuration for dark-field microscopy. The central opaque light stop is positioned beneath the condenser to eliminate zeroth-order illumination of the specimen. The condenser produces a hollow cone of illumination that strikes the specimen at oblique angles. Some of the reflected, refracted, and diffracted light from the specimen enters the objective front lens.

objective, the oblique rays cross and all such rays miss entering the objective because of the obliquity. The field of view will appear dark.

In Fourier optics, dark-field illumination removes the zeroth order (unscattered light) from the diffraction pattern formed at the rear focal plane of the objective (5). Oblique rays, now diffracted by the specimen and yielding first, second, and higher diffracted orders at the rear focal plane of the objective, proceed onto the image plane where they interfere with one another to produce an image of the specimen. This results in an image formed exclusively from higher order diffraction intensities scattered by the specimen. Specimens that have smooth reflective surfaces produce images due, in part, to the reflection of light into the objective (18,19). When the refractive index is different from that of the surrounding medium or when refractive index gradients occur (as in the edge of a membrane), light is refracted by the specimen. Both instances of reflection and refraction produce relatively small angular changes in the direction of light and allow some to enter the objective. In contrast, some light that strikes the specimen is also diffracted and produces a 180° arc of light that passes through the entire numerical aperture range of the objective. The resolving power of the objective is the same in dark-field illumination as that in bright-field conditions, but the optical character of the image is not as faithfully reproduced.

Several pieces of equipment are used to produce dark-field illumination. The simplest is a *spider stop* (Fig. 14) placed just under the bottom lens (in the front focal plane) of the substage condenser (5,18,19,28). Both the aperture and field diaphragms are opened wide to pass oblique rays. The central opaque stop (you can make one by mounting a coin on a clear glass disk) blocks out the central rays. This device works fairly well, even with the Abbe condenser, when using the 10X objective up to 40X or higher objectives that have a numerical aperture no higher than 0.65. The diameter of the opaque stop should be approximately 16–18 mm for a 10X objective of numerical aperture 0.25 to approximately 20–24 mm for 20X and 40X objectives of numerical apertures that approach 0.65. The stop should be placed as close as possible to the aperture diaphragm.

For more precise work and blacker backgrounds, use a condenser designed especially for dark-field, that is, to transmit only oblique rays (28,29). There are several varieties, including dry dark-field condensers that have air between the top of the condenser and the underside of the slide. Immersion dark-field condensers are also available. These require using a drop of immersion oil (some are designed to use water instead) that establishes contact between the top of the condenser and the underside of the specimen slide. The immersion dark-field condenser has internal mirrored surfaces and passes rays of great obliquity free of chromatic aberration that produce the best results and blackest background.

Dark-field objects are quite spectacular to see, and objects of very low contrast in bright-field shine brilliantly in dark-field. Such illumination is best for revealing outlines, edges, and boundaries. A high-magnification dark-field image of a diatom is illustrated in Fig. 15.

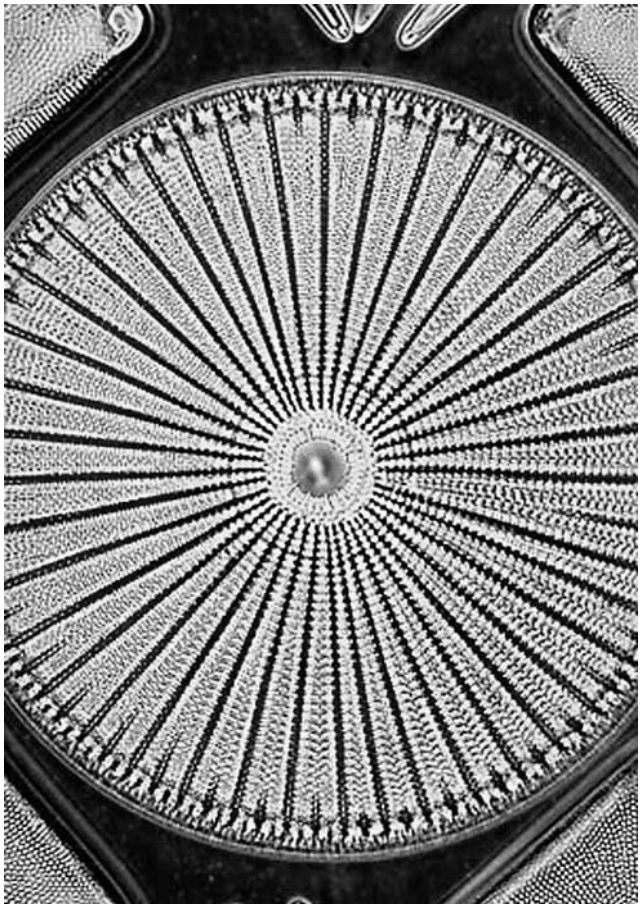


Figure 15. Dark-field photomicrograph of the diatom *Arachnoidiscus ehrenbergi* taken at high magnification using oil immersion optics and a 100X objective.

Rheinberg Illumination

Rheinberg illumination, a form of optical staining, was initially demonstrated by the British microscopist Julius Rheinberg to the Royal Microscopical Society and the Quekett Microscopical Club (England) more than one hundred years ago (18,28). This technique is a striking variation of low to medium power dark-field illumination using colored gelatin or glass filters to provide rich color to both the specimen and background. The central opaque dark-field stop is replaced by a transparent, colored, circular stop inserted into a transparent ring of a contrasting color (illustrated in Fig. 16). These stops are placed under the bottom lens of the condenser. The result is a specimen rendered in the color of the ring whose background has the color of the central spot. A photomicrograph of the spiracle and trachea of a silkworm larva taken by Rheinberg illumination is shown in Fig. 17.

PHASE-CONTRAST MICROSCOPY

Research by Frits Zernike during the early1930s uncovered phase and amplitude differences between zeroth-order and deviated light that can be altered to produce favorable conditions for interference and contrast

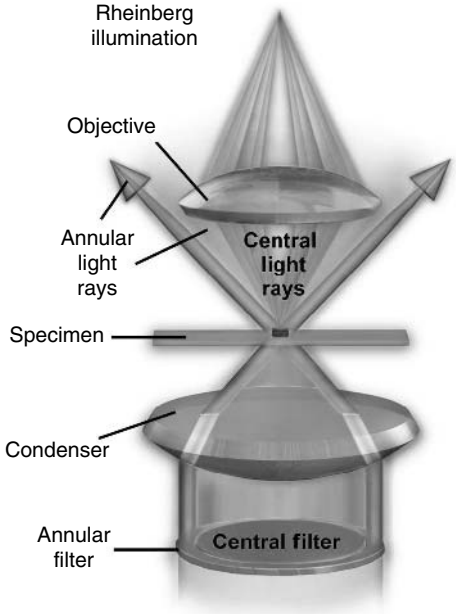


Figure 16. Schematic microscope configuration for Rheinberg illumination. Light passes through the central/annular filter pack before entering the condenser. Zeroth-order light from the central filter pervades the specimen and illuminates it with higher order light from the annular filter. The filter colors in this diagram are a green central and red annular.

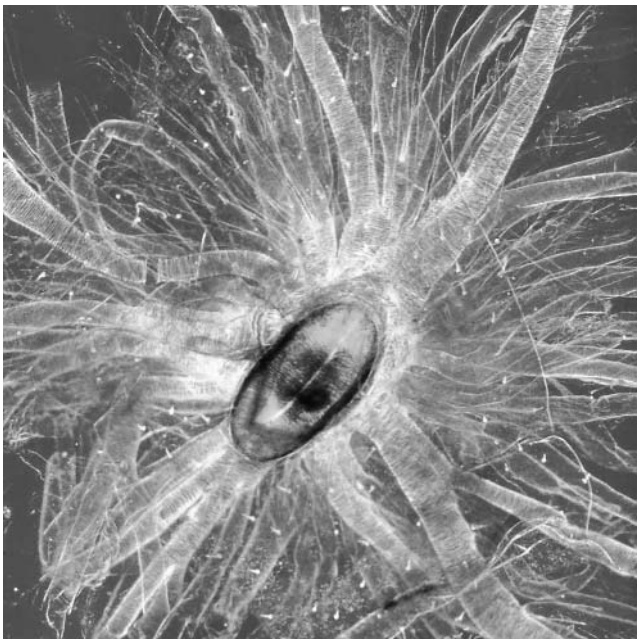


Figure 17. Spiracle and trachea from silkworm larva photographed at low magnification under Rheinberg illumination using a blue central and yellow annulus filter combination and a 2X objective.

enhancement (30,31). Unstained specimens that do not absorb light are called phase objects because they slightly alter the phase of the light diffracted by the specimen, usually by retarding such light approximately one-quarter wavelength compared to the undeviated direct light that passes unaffected through or around the specimen. Unfortunately, our eyes as well as camera film cannot detect these phase differences. To reiterate, the human eye is sensitive only to the colors of the visible spectrum or to differing levels of light intensity (related to wave amplitude).

The direct zeroth-order light in phase specimens, passes undeviated through or around the specimen. However, the light diffracted by the specimen is not reduced in amplitude as it is in a light-absorbing object, but is slowed by the specimen because of the specimen's refractive index or thickness (or both). This diffracted light that lags behind by approximately one-quarter wavelength arrives at the image plane *out of phase* with the undeviated light but, in interference, essentially undiminished in intensity. The result is that the image at the eyepiece level is so lacking in contrast as to make the details almost invisible.

Zernike devised a method, now known as phase, contrast microscopy, for making unstained, phase objects yield contrast images as if they were amplitude objects. Amplitude objects show excellent contrast when the diffracted and direct light are out of phase by one-half wavelength (18). Zernike's method was to *speed up* the direct light by one-quarter wavelength so that the difference in wavelength between the direct and deviated light for a phase specimen would now be one-half wavelength. As a result, the direct and diffracted light arriving at the image level of the eyepiece could produce destructive interference (see the section on **Image formation** for absorbing objects previously described). In such a procedure, the details of the image appear darker against a lighter background. This is called dark or positive phase contrast. A schematic illustration of the basic phase contrast microscope configuration is shown in Fig. 18.

Another possible course, much less often used, is to arrange to *slow down* the direct light by one-quarter wavelength, so that the diffracted light and the direct light arrive at the eyepiece in phase and can interfere constructively (2,5,18). This arrangement called negative or bright contrast, results in a bright image of the details of the specimen on a darker background.

Phase contrast involves *separating* the direct zeroth-order light from the diffracted light at the rear focal plane of the objective. To do this, a ring annulus is positioned directly under the lower lens of the condenser at the front focal plane of the condenser, conjugate to the objective rear focal plane. As the hollow cone of light from the annulus passes undeviated through the specimen, it arrives at the rear focal plane of the objective in the shape of a ring of light. The fainter light diffracted by the specimen is spread over the rear focal plane of the objective. If this combination were allowed, as is, to proceed to the image plane of the eyepiece, the diffracted light would be approximately one-quarter wavelength behind the direct light. At the image plane, the phase of the diffracted light would be out of phase with the direct light, but the amplitude of their interference would be almost the same as that of the direct light (5,18). This would result in very little specimen contrast.

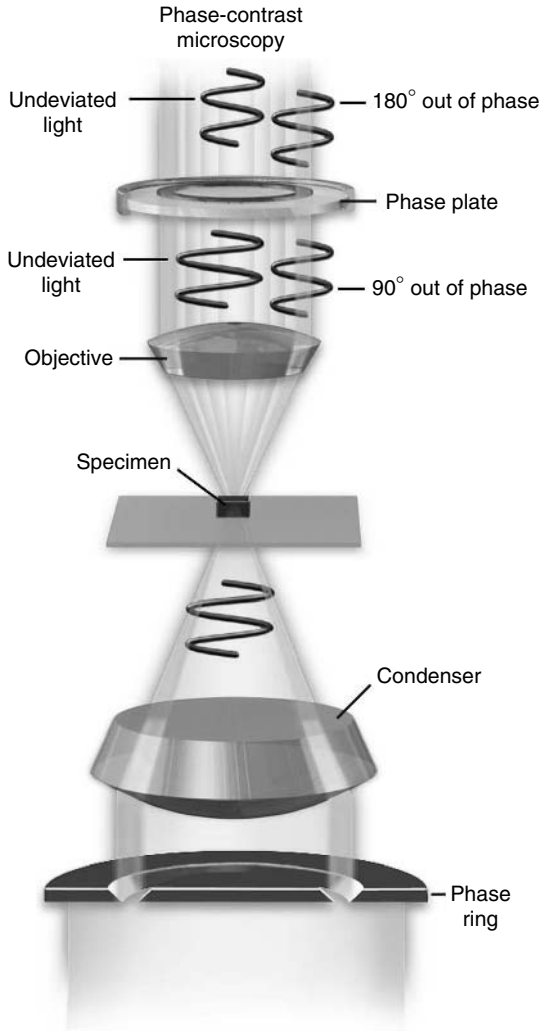


Figure 18. Schematic configuration of phase-contrast microscopy. Light that passes through the phase ring is first concentrated onto the specimen by the condenser. Undeviated and out-of-phase light enters the objective and is retarded by the phase plate before interference at the rear focal plane of the objective.

To speed up the direct undeviated zeroth-order light, a phase plate is installed that has a ring-shaped *phase shifter* attached to it at the rear focal plane of the objective. The narrow area of the phase ring is optically thinner than the rest of the plate. As a result, undeviated light that passes through the phase ring travels a shorter distance in traversing the glass of the objective than the diffracted light. Now, when the direct undeviated light and the diffracted light proceed to the image plane, they are one-half wavelength out of phase with each other. The diffracted and direct light can now interfere destructively so that the details of the specimen appear dark against a lighter background (just as they do for an absorbing or amplitude specimen). This is a description of what takes place in positive or dark phase contrast.

If the ring phase-shifter area of the phase plate is made optically thicker than the rest of the plate, direct light is slowed by one-quarter wavelength. In this case, the zeroth-order light arrives at the image plane in phase with the diffracted light, and constructive interference takes place. The image appears bright on a darker background. This type of phase contrast is described as *negative* or *bright contrast* (2,5,18,19).

Because undeviated zeroth-order light is much brighter than faint diffracted light, a thin absorptive transparent metallic layer is deposited on the phase ring to bring the direct and diffracted light into better intensity balance to increase contrast. Because speeding up or slowing down the direct light is calculated on one-quarter wavelength of green light, the phase image will also appear best when a green filter is placed in the light path (a green interference filter is preferable). Such a green filter also helps achromatic objectives produce their best images because achromats are spherically corrected for green light.

The accessories needed for phase-contrast work are a substage phase-contrast condenser equipped with annuli and a set of phase-contrast objectives, each of which has a phase plate installed. The condenser usually has a bright-field position that has an aperture diaphragm and a rotating turret of annuli (each phase objective of different magnification requires an annulus of increasing diameter as the magnification of the objective increases). Each phase objective has a darkened ring on its back lens. Such objectives can also be used for ordinary bright-field transmitted light work that has only a slight reduction in image quality. A small phase telescope or a Bertrand lens is used to facilitate the centering of these rings in the optical pathway. A photomicrograph of hair cross sections from a fetal mouse taken by phase-contrast illumination is shown in Fig. 19.

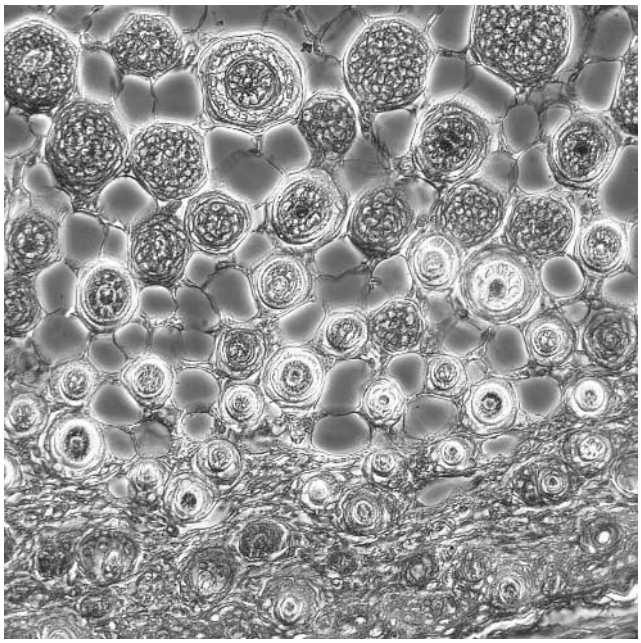


Figure 19. Photomicrograph of a hair thin section from a fetal mouse taken by using phase-contrast optics and a 20X objective.

POLARIZED LIGHT

Many transparent solids are optically isotropic, meaning that the index of refraction is equal in all directions throughout the crystalline lattice. Examples of isotropic solids are glass, table salt (sodium chloride), many polymers, and a wide variety of organic and inorganic compounds (32,33).

Crystals are classified as either isotropic or anisotropic, depending upon their optical behavior and whether their crystallographic axes are equivalent. All isotropic crystals have equivalent axes that interact similarly with light, regardless of the crystal orientation with respect to incident light waves. Light that enters an isotropic crystal is refracted at a constant angle and passes through the crystal at a single velocity without being polarized by interaction with the electronic components of the crystalline lattice.

Anisotropic crystals, on the other hand, have crystallographically distinct axes and interact with light in a manner that depends on the orientation of the crystalline lattice with respect to the incident light. When light enters the optical axis of anisotropic crystals, it acts similarly to interaction with isotropic crystals and passes through at a single velocity. However, when light enters a nonequivalent axis, it is refracted into two rays; each is polarized so that their vibrational directions are oriented at right angles to one another and travel at different velocities. This phenomenon is termed *double* or *birefraction* and is seen to a greater or lesser degree in all anisotropic crystals (32–34).

When anisotropic crystals refract light, the resulting rays are polarized and travel at different velocities. One of the rays, termed the ordinary ray, travels with the same velocity in every direction through the crystal. The other ray travels with a velocity that depends on the propagative direction within the crystal. This light ray is termed the extraordinary ray. The retardation between the ordinary and extraordinary ray increases as crystal thickness increases. The two independent refractive indexes of anisotropic crystals are quantified in terms of their birefringence, a measure of the difference in refractive index. Thus, the birefringence B of a crystal is defined as

$$B = |n_{\text{high}} - n_{\text{low}}|, \tag{5}$$

where n_{high} is the largest refractive index and n_{low} is the smallest. This expression holds true for any part or fragment of an anisotropic crystal except for light waves propagated along the optical axis of the crystal. As mentioned before, light that is doubly refracted through anisotropic crystals is polarized, so that the vibrational directions of the polarized ordinary and extraordinary light waves are oriented perpendicular to each other. Now, we can examine how anisotropic crystals behave under polarized illumination in a polarizing microscope.

A polarizer placed beneath the substage condenser is oriented, so that polarized light that exits the polarizer is plane polarized in a vibrational direction that is east–west with respect to the optic axis of the microscope stand. Polarized light enters the anisotropic crystal where it is refracted and divided into two separate components that vibrate parallel to the crystallographic axes and perpendicular to each other. Then, the polarized light waves pass through the specimen and objective before reaching a second polarizer (usually termed the analyzer) that is oriented to pass a polarized vibration perpendicular to that of the substage polarizer. Therefore, the analyzer passes only those components of the light waves that are parallel to the polarization direction of the analyzer. The retardation of one ray with respect to another is caused by the difference in speed between the ordinary and extraordinary rays refracted by the anisotropic crystal (33,34). A schematic illustration of a microscope configuration for crossed polarized illumination is presented in Fig. 20.

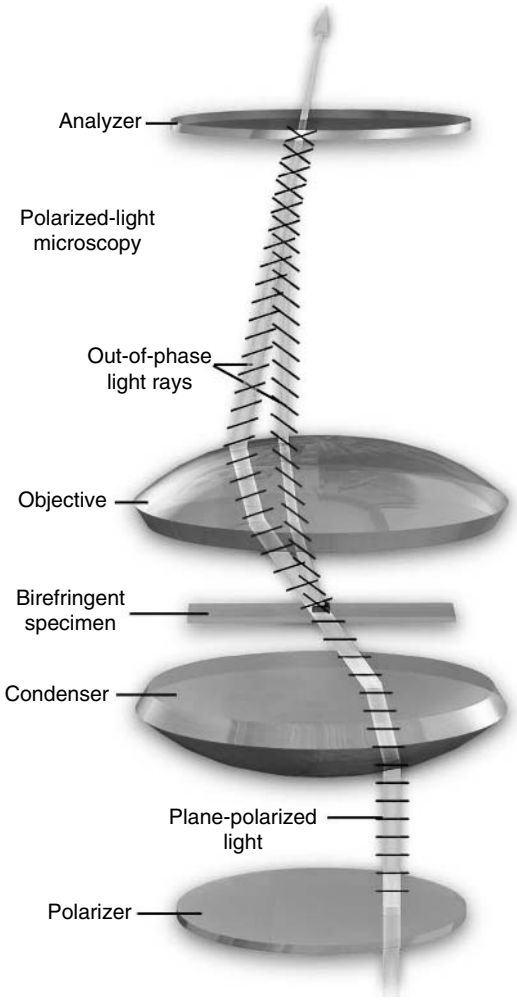


Figure 20. Schematic microscope configuration for observing birefringent specimens under crossed polarized illumination. Light that passes through the polarizer is plane polarized and concentrated onto the birefringent specimen by the condenser. Light rays that emerge from the specimen interfere when they are recombined in the analyzer and produce a myriad of tones and colors.

Now we will consider the phase relationship and velocity differences between the ordinary and extraordinary rays after they pass through a birefringent crystal. These

rays are oriented so that they vibrate at right angles to each other. Each ray encounters a slightly different electrical environment (refractive index) as it enters the crystal, and this affects the velocity at which the ray passes through the crystal (32,33). Because of the difference in refractive indexes, one ray will pass through the crystal at a rate slower than the other ray. In other words, the velocity of the slower ray will be retarded with respect to the faster ray. This retardation can be quantified by the following equation:

$$\text{Retardation } (\Gamma) = \text{ thickness } (t) \times \text{birefringence } (B), \quad (6)$$

or

$$\Gamma = t \times |n_{\text{high}} - n_{\text{low}}|, \quad (7)$$

where Γ is the quantitative retardation of the material, t is the thickness of the birefringent crystal (or material), and B is birefringence as defined before (33). Factors that contribute to the retardation value are the magnitude of the difference in refractive indexes for the environments seen by the ordinary and extraordinary rays and also the sample thickness. Obviously, the greater the difference in either refractive indexes or thickness, the greater the degree of retardation. Early observations made on the mineral calcite indicate that thicker calcite crystals cause greater differences in splitting of the images seen through the crystals. This agrees with Eq. (7) which states that retardation increases with crystal (or sample) thickness.

When the ordinary and extraordinary rays emerge from the birefringent crystal, they are still vibrating at right angles with respect to one another. However, the component vectors of these waves that pass through the analyzer vibrate in the same plane. Because one wave is retarded with respect to the other, interference (either constructive or destructive) occurs between the waves as they pass through the analyzer. The net result is that some birefringent samples (in white light) acquire a spectrum of color when observed through crossed polarizers.

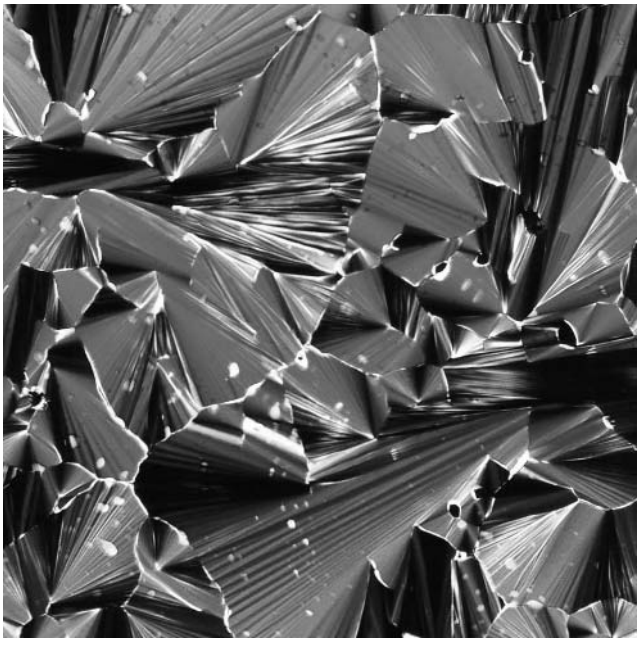


Figure 21. Photomicrograph of high-density columnar-hexatic liquid crystalline calf thymus DNA at a concentration of approximately 450 mg/mL. This concentration of DNA approaches that found in sperm heads, virus capsids, and dinoflagellate chromosomes. The image was recorded using a polarized light microscope and the 10X objective.

Polarized light microscopy requires strain-free objectives and condensers to avoid depolarizing the transmitted light (32–34). Most polarized microscopes are equipped with a centerable stage that has free 360° rotation about the optical axis of the microscope. A Bertrand lens is commonly inserted into the light path so that conoscopic images of birefringent patterns can be observed at the back of the objective. Manufacturers also offer a wide range of compensators and light retarders (full-wave, quarter-wave, and continuous wedge plates) for quantitative measurements of birefringence and for adding color to polarized light photomicrographs. Skilled use of these compensators allows precise chemical identification of minute crystals. Birefringent DNA liquid crystals photographed by crossed polarized illumination are illustrated in Fig. 21.

HOFFMAN MODULATION CONTRAST

The Hoffman modulation-contrast system is designed to increase visibility and contrast in unstained and living material by detecting optical gradients (or slopes) and converting them into variations of light intensity. This ingenious technique, invented by Dr. Robert Hoffman in 1975, employs several accessories that have been adapted to major commercial microscopes (16,18). A schematic illustration of microscope configuration for Hoffman modulation contrast is presented in Fig. 22.

An optical amplitude spatial filter, termed a *modulator* by Hoffman, is inserted at the rear focal plane of an achromat or planachromat objective (although higher correction can also be used). Light intensity that passes through this system varies above and below an average value, which by definition, is then said to be modulated. Objectives useful for modulation contrast can cover the entire magnification range of 10–100X. The modulator has three zones: a small, dark zone near the periphery of the rear focal plane that transmits only 1% of light; a narrow gray zone that transmits 15%; and the remaining clear or transparent zone that covers most of the area at the back of the objective and transmits 100% of the light (5,16,18). Unlike the phase plate in phase-contrast microscopy, the Hoffman modulator does not alter the phase of the light that passes through any of the zones. When viewed by modulation-contrast optics, transparent objects that are essentially invisible in ordinary bright-field microscopy take on an apparent three-dimensional appearance dictated by phase gradients. The modulator does not introduce changes in the phase relationship of light that passes through the system, but it influences the principal zeroth-order maxima. Higher order diffraction maxima are unaffected. Measurements by a Michelson interferometer confirm that the spatial coherency of light

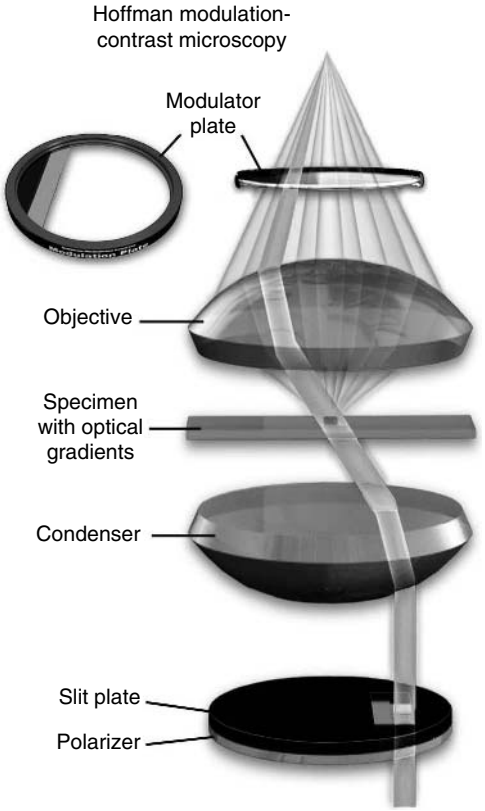


Figure 22. Schematic illustration of microscope configuration for Hoffman modulation contrast. Light that passes through the polarizer and slit is focused onto the specimen by the condenser. After passing through the specimen, light enters the objective and is filtered by the modulator plate in the rear focal plane of the objective.

passed through a Hoffman-style modulator varies (if any) by a factor of less than $\lambda/20$ (5).

Below the stage, a condenser equipped withb a rotating turret is used to hold the remaining components of the Hoffman modulation-contrast system. The turret condenser has a bright-field opening that has an aperture iris diaphragm for regular bright-field microscopy and for aligning and establishing proper conditions of Köhler illumination for the microscope. At each of the other turret openings, there is an off-center slit that is partially covered by a small rectangular polarizer. The size of the slit/polarizer combination differs for each objective of different magnification; hence the need for a turret arrangement. When light passes through the off-axis slit, it is imaged at the rear focal plane of the objective (also termed the Fourier plane), where the modulator has been installed. Like the central annulus and phase ring in phase-contrast microscopy, the front focal plane of the condenser that contains the off-axis slit plate is optically conjugate to the modulator in the objective rear focal plane. Image intensity is proportional to the first derivative of the optical density in the specimen and is controlled by the zeroth order of the phase gradient diffraction pattern.

Below the condenser, a circular polarizer is placed on the light exit port of the microscope (note that both polarizers are below the specimen). The rotation of this polarizer can control the effective width of the slit opening. For example, *crossing* both polarizers at 90° to each other results in *narrowing* the slit, so that its image falls within the gray area of the modulator (16,18). The part of the slit controlled by the polarizer registers on the bright area of the modulator. As the polarizer is rotated, contrast can be varied for best effect. A very narrow slit produces images that have very high contrast and a high degree of coherence. Optical section imaging is also optimized when the slit is adjusted to its narrowest position. When the circular polarizer is oriented so that its vibrational direction is parallel to that of the polarizer in the slit, the effective slit width is at a maximum. This reduces overall image contrast and coherence but yields much better images of thicker objects where large differences in refractive index exist.

In modern advanced modulation-contrast systems, both the modulator and the slit are offset from the optical axis of the microscope (18). This arrangement permits fuller use of the numerical aperture of the objective and results in good resolution of specimen detail. Shapes and details are rendered in shadowed, pseudo-three-dimensional appearance. These appear brighter on one side, gray in the central portion, and darker on the other side, against a gray background. The modulator converts optical phase gradients in details (steepness, slope, rate of change in refractive index, or thickness) into changes in the intensity of various areas of the image at the plane of the eyepiece diaphragm. Resulting images have an apparent three-dimensional appearance and directional sensitivity to optical gradients. The contrast (related to variations in intensity) of the dark and bright areas against the gray gives a shadowed pseudorelief effect. This is typical of modulation-contrast imaging. Rotation of the polarizer alters the contrast achieved, and the orientation of the specimen on the stage (with respect to the polarizer and offset slit) may dramatically improve or degrade contrast.

There are numerous advantages as well as limitations to modulation contrast. Some of the advantages include fuller use of the numerical aperture of the objective that yields excellent resolution of details along with good specimen contrast and visibility. Although many standard modulation-contrast objectives are achromats or planachromats, it is also possible to use objectives that have a higher degree of correction for optical aberration. Many major microscope manufacturers now offer modulation-contrast objectives in fluorite-correction grades, and apochromats can be obtained by special order. Older objectives can often be retrofitted with a modulator made by Modulation Optics, Inc., the company founded by Dr. Robert Hoffman specifically to build aftermarket and custom systems.

In addition to the advantages of using higher numerical apertures that have modulation contrast, it is also possible to do *optical sectioning* by this technique (16,18). Sectioning allows the microscopist to focus on a single thin plane of the specimen without interference from confusing images that arise in areas above or below the plane that is focused on. The depth of a specimen is measured in a direction parallel to the optical axis of the

microscope. Focusing the image establishes the correct specimen-to-image distance and allows interference of the diffracted waves at a predetermined plane (the image plane) positioned at a fixed distance from the eyepiece. This enables viewing diffracting objects separately that occur at different depth levels in the specimen, provided there is sufficient contrast. The entire depth of a specimen can be optically sectioned by sequentially focusing on each succeeding plane. In this system, depth of field is defined as the distance from one level to the next where imaging of distinct detail occurs, and is controlled by the numerical aperture of the objective (5,16,18). Higher numerical aperture objectives exhibit very shallow depths of field and the opposite holds for objectives of lower numerical aperture. The overall capability of an objective to isolate and focus on a specific optical section diminishes as the optical homogeneity of the specimen decreases.

Birefringent objects (rock thin sections, crystals, bone, etc.), that can confuse images in DIC (see next section) can be examined because the specimen is *not* situated between the two polarizers (18). Further, specimens can be contained in plastic or glass vessels without deterioration of the image due to polarization effects, because such vessels are also above both polarizers, not between them. This makes the Hoffman system far more useful than DIC in examining and photomicrographing cell, tissue, and organ culture in plastic containers.

Hoffman modulation contrast can be simultaneously combined with polarized light microscopy to achieve spectacular effects in photomicrography. Figure 23 illustrates a thin section of a dinosaur bone photographed by using a combination of Hoffman modulation-contrast and plane-polarized illumination. Note the pseudo-three-dimensional relief apparent throughout the micrograph and the beautiful coloration provided by polarized light.

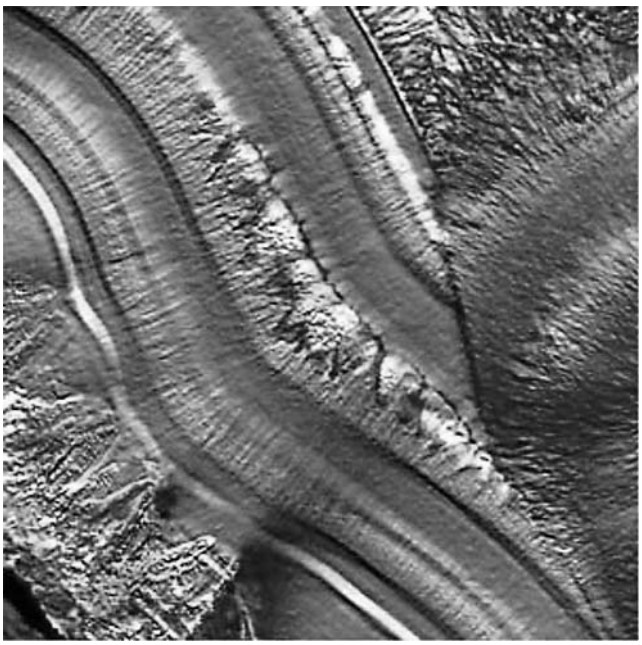


Figure 23. Thin section of a duck-billed dinosaur bone photographed by using a combination of Hoffman modulation-contrast and polarized-light illumination using a 20X objective. Note the pseudo-three-dimensional relief throughout the photomicrograph.

DIFFERENTIAL INTERFERENCE CONTRAST

In the mid 1950s, the French optics theoretician Georges Nomarski improved the method for detecting optical gradients in specimens and converting them into intensity differences (15). Today, there are several implementations of this design, which are collectively called differential interference contrast (DIC). Living or stained specimens, which often yield poor images when viewed in bright-field illumination, are made clearly visible by optical rather than chemical means (2,4,5,15,18–22).

In transmitted-light DIC, light from the lamp is passed through a polarizer located beneath the substage condenser, similarly to polarized-light microscopy. Next in the light path (but still beneath the condenser) is a modified Wollaston prism that *splits* the entering beam of polarized light into two beams that travel in slightly different directions (illustrated in Fig. 24). The prism is composed of two halves cemented together (36). Emerging light rays vibrate at 90 ° degrees relative to each other and have a slight path difference. A different prism is needed for each objective of different magnification. A revolving turret on the condenser allows the microscopist to rotate the appropriate prism into the light path when changing magnifications.

The plane-polarized light that vibrates only in one direction perpendicular to the propagative direction of the light beam enters the beam-splitting modified Wollaston prism and is split into two rays that vibrate perpendicular to each other (Fig. 24). These two rays travel close together, but in slightly different directions. The rays intersect at the front focal plane of the condenser, where they pass, travel parallel, extremely close together, and have a slight path difference, but they are vibrating perpendicular to each other and therefore cannot cause interference. The distance between the rays, called the shear, is so minute that it is less than the resolving ability of the objective (5,18,36).

The split beams enter and pass through the specimen where their wave paths are altered in accordance with the specimen's varying thickness, slopes, and refractive indexes. These variations alter the wave path of both beams that pass through areas of the specimen that lie close together. When the parallel beams enter the objective, they are focused above the rear focal plane where they enter a second modified Wollaston prism that combines the two beams. This removes the shear and the original path difference between the beam pairs. As a result of traversing the specimen, the paths of the parallel beams do not have the same length (optical path difference) for different areas of the specimen.

For the beams to interfere, the vibrations of the beams of different path length must be brought into the same plane and axis by placing a second polarizer (analyzer) above the upper Wollaston beam-combining prism. Then, the light proceeds toward the eyepiece where it can be observed as differences in intensity and color. The design

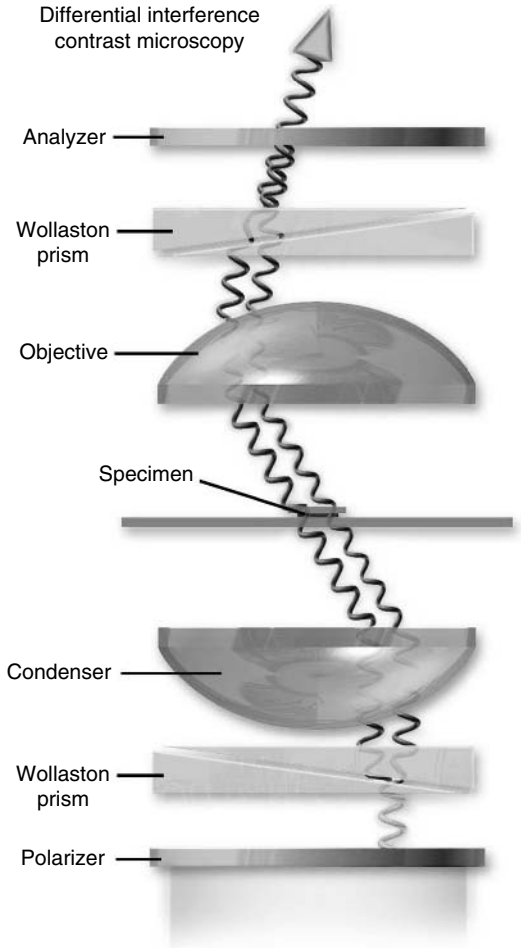


Figure 24. Schematic illustration of microscope configuration for differential interference contrast. Light is polarized in a single vibrational plane by the polarizer before it enters the lower Wollaston prism that acts as a beam splitter. Next, the light passes through the condenser and sample before the image is reconstructed by the objective. The second Wollaston prism at the rear focal plane of the objective acts as a beam combiner and passes the light to the analyzer, where it interferes both constructively and destructively.

makes one side of a detail appear bright (or possibly in color) whereas the other side appears darker (or another color). This shadow effect gives the specimen a pseudo-three-dimensional appearance (18).

In some microscopes, the upper modified Wollaston prism is combined in a single fitting, and the analyzer is incorporated above it. The upper prism may also be arranged so that it can be moved horizontally. This allows varying optical path differences by moving the prism and provide a mechanism for altering the brightness and color of the background and specimen. Because of the prism design and placements, the background will be homogeneous for whichever color has been selected.

The color and/or light intensity effects shown in the image are related especially to the rate of change in refractive index, specimen thickness, or both. Orientation of the specimen can have a pronounced effect on the relief-like appearance; often, rotating of the specimen by 180° changes a *hill* into a *valley* or vice versa. The three-dimensional appearance does not represent the true geometry of the specimen, but it is an exaggeration based on *optical thickness*. It is not suitable for accurate measurement of actual heights and depths.

There are numerous advantages of DIC microscopy compared in particular to phase-contrast and Hoffman modulation-contrast microscopy. In DIC, it is also possible to use the numerical aperture of the system more fully and to provide optical staining (color). DIC also allows the microscope to achieve excellent resolution. The use of the full objective aperture enables the microscopist to focus on a thin plane section of a thick specimen without confusing images from above or below the plane. Annoying *halos*, often encountered in phase-contrast, are absent in DIC images, and common achromat and fluorite objectives can be used for this work. A downside is that plastic tissue culture vessels and other birefringent specimens yield confusing images in DIC. High-quality apochromatic objectives are now designed so that they are suitable for DIC. Figure 25a presents transmitted and reflected light DIC photomicrographs of mouth parts of a blowfly (transmitted DIC), and Figure 25b shows surface defects in a ferrosilicate alloy (reflected DIC). Both photomicrographs were made by using a retardation plate and a 10X objective.

FLUORESCENCE MICROSCOPY

Fluorescence microscopy is an excellent tool for studying material which can be made to fluoresce, either in its natural form (primary or autofluorescence) or when treated with chemicals that fluoresce (secondary fluorescence). This form of optical microscopy is rapidly reaching maturity and is now one of the fastest growing areas of microscopic investigation (6–8).

The basic task of the fluorescence microscope is to permit excited light to irradiate the specimen and then to separate the much weaker reradiating fluorescent light from the brighter excited light. Thus, only the emitted light reaches the eye or other detector. The resulting fluorescing areas shine against a dark background and have sufficient contrast to permit detection. The darker the background of the nonfluorescing material, the more efficient the instrument. For example, ultraviolet (UV) light of a specific wavelength or set of wavelengths is produced by passing light from a UV-emitting source through the exciter filter. The filtered UV light illuminates the specimen, which emits fluorescent light of longer wavelengths while illuminated with ultraviolet light. Visible light emitted from the specimen is then filtered through a barrier filter that does not allow reflected UV light to pass. It should be noted that this is the only mode of microscopy in which the specimen, gives off its own light subsequent to excitation (6). The emitted light reradiates spherically in all directions, regardless of the direction of the exciting light. A schematic diagram of the configuration of fluorescence microscopy is presented in Fig. 26.

Fluorescence microscopy has advantages based upon attributes not as readily available in other optical

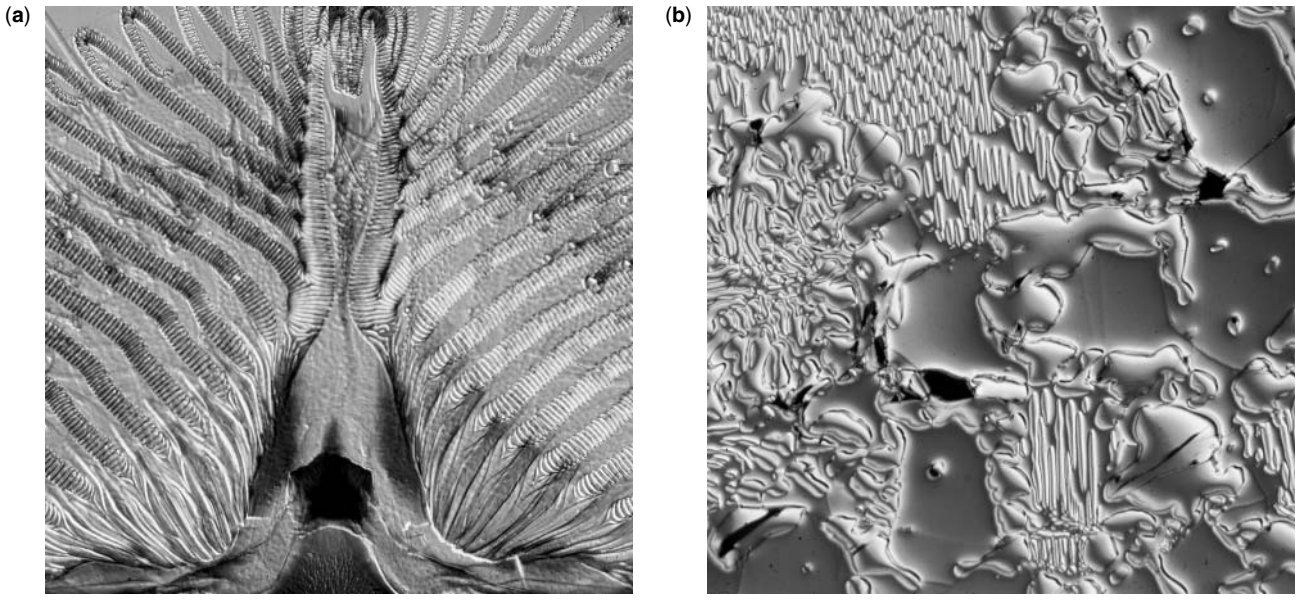


Figure 25. (**a**) Nomarski transmitted light differential interference contrast (DIC) photomicrograph of mouthparts from a blowfly. The image was recorded using a retardation plate and the 10x objective. (**b**) Reflected light differential interference contrast photomicrograph of defects on the surface of a ferro silicon alloy. The image was captured by using a 10X objective and a first-order retardation plate.

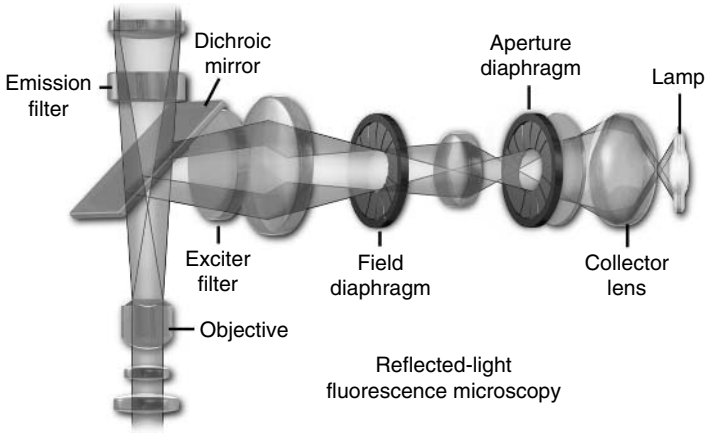


Figure 26. Schematic diagram of the configuration of reflected-light fluorescence microscopy. Light emitted from a mercury arc lamp is concentrated by the collector lens before it passes through the aperture and field diaphragms. The exciter filter passes only the desired wavelengths, which are reflected down through the objective to illuminate the specimen. Fluorescence emitted by the specimen passes back through the objective and dichroic mirror before it is filtered by the barrier emission filter.

microscopic techniques. The use of fluorochromes has made it possible to identify cells and submicroscopic cellular components and entities that have a high degree of specificity amidst nonfluorescing material. What is more, the fluorescence microscope can reveal the presence of fluorescing material with exquisite sensitivity. An extremely small number of fluorescing molecules (as few as 50 molecules per cubic micron) can be detected (6–8). By multiple staining of a given sample, different probes will reveal the presence of different target molecules. Although the fluorescence microscope cannot provide *spatial resolution* below the diffraction limit of the respective objectives, the *presence* of fluorescing molecules below such limits is made visible.

Techniques of fluorescence microscopy can be applied to organic material, formerly living material, or to living material (by using *in vitro* or *in vivo* fluorochromes). These techniques can also be applied to inorganic material (especially when investigating contaminants on semiconductor wafers). There are also a burgeoning number of studies using fluorescent probes to monitor rapidly changing physiological ion concentrations (calcium, magnesium, etc.) and pH values in living cells (1,7,8).

Some specimens fluoresce without the addition of dyes when irradiated with shorter wavelength light (primary or autofluorescence). Autofluorescence has been found useful in plant studies, coal petrography, sedimentary rock petrology, and in the semiconductor industry. In the study of animal tissues or pathogens, autofluorescence is often either extremely faint or has such nonspecificity as to make it of minimal use. Of far greater value for such specimens are the fluorochromes (also called fluorophores) which are excited by irradiating light and whose eventual

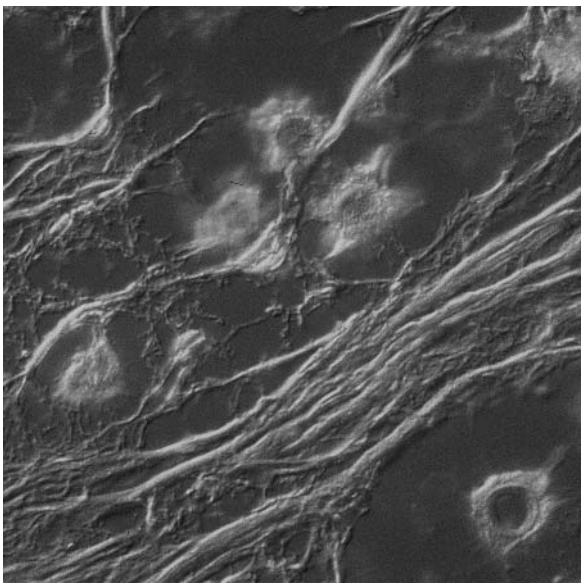


Figure 27. Fluorescence/DIC combination photomicrograph of cat brain tissue infected with *Cryptococcus*. Infected neurons are heavily stained with the DNA-specific fluorescent dye acridine orange. Note the pseudo-three-dimensional effect that occurs when these two techniques are combined. The micrograph was recorded by using a 40X objective.

yield of emitted light is of greater intensity. This is called secondary fluorescence.

Fluorochromes are stains, somewhat similar to the better known tissue stains, which attach themselves to visible or subvisible organic matter. These fluorochromes that can absorb and then reradiate light are often highly specific in their attachment targeting and have significant yields in absorption–emission ratios. This makes them extremely valuable in biological applications. The growth in the use of fluorescence microscopes is closely linked to the development of hundreds of fluorochromes that have known intensity curves of excitation and emission and well-understood biological structure targets (6). When deciding which label to use for fluorescence microscopy, bear in mind that the fluorochrome chosen should have a high likelihood of absorbing the excited light and should remain attached to the target molecules. The fluorochrome should also provide a satisfactory yield of emitted fluorescent light. Figure 27 shows a photomicrograph of cat brain cells infected with *Cryptococcus*. The image was made by using a combination of fluorescence and DIC microscopy, taking advantage of the features of both contrast-enhancing techniques. Infected neurons are heavily stained with the DNA-specific fluorescent dye acridine orange, and the entire image is rendered in a pseudo-three-dimensional effect by DIC.

PHOTOMICROGRAPHY AND DIGITAL IMAGING

The use of photography to capture images in a microscope dates back to the invention of the photographic process (37). Early photomicrographs had remarkable quality, but the techniques were laborious and burdened with long exposures and a difficult process for developing emulsion plates. The primary medium for photomicrography was film until the past decade when improvements in electronic camera and computer technology made digital imaging cheaper and easier to use than conventional photography. This section will address photomicrography on film and by electronic analog and digital imaging systems.

The quality of a photomicrograph, either digital or recorded on film, depends on the quality of the microscopy. Film is a stern judge of how good the microscopy has been before it captures the image. It is essential to configure the microscope using Köhler illumination, to adjust the field and condenser diaphragms correctly to and optimize the condenser height. When properly adjusted, the microscope yields images that have even illumination across the entire field of view and displays the best compromise of contrast and resolution.

Photographic Films

Films for photography are coated with a light-sensitive emulsion of silver salts and/or dyes. When light is allowed to expose the emulsion, *active centers* combine to form a latent image that must be developed by photographic chemicals (19,37–39). This requires exposing the film in a darkened container to a series of solutions where temperature, development time, and appropriate agitation or mixing of the solutions must be controlled. The developing process must then be halted by a *stop* solution. Next, the unexposed emulsion, which consists of silver salts and dyes, is cleared, and the film is fixed, then washed, and dried for use. The development, stop, fixing, and clearing must be done in a darkroom or in lighttight developing tanks, and film must be handled in complete darkness. The temperature, duration, and agitation usually depend on the film used. For example, the Kodachrome K14 process is far more demanding than the E6 process used for Ektachrome and Fujichrome (19,37). The emulsion speed determines how much light must be used to expose the film in a given time period. Films are rated according to their ASA or ISO number, which gives an indication of the relative film speed. Larger ISO numbers indicate faster films; A film of ISO 25 is one of the slowest available a film and of ISO 1600 is one of the fastest. Because the microscope is a relatively stable platform that has good illumination, films in the 50–200 ISO range are commonly used for photomicrography.

Film is divided into a number of categories, depending on whether it is intended for black-and-white or color photography. Color films are subdivided into two types: color films that yield positive transparencies (colors like those of the image being observed) and color negatives (colors complementary to those in the microscope image, e.g., green for magenta). Color films can be further subdivided into two groups: films designed to receive daylight quality light and those designed to receive indoor or tungsten light. As a rule, but with some exceptions, transparency or positive films (slides) have the suffix *chrome*, such as Kodachrome, Ektachrome, Fujichrome, Agfachrome, etc. Color negative films usually end with the suffix *color*, such as Kodacolor, Ektacolor, Fujicolor, Agfacolor, etc. (37). Each of these film types is offered in

a variety of film speeds or ISO ratings. Film packages usually display the ISO of the film and indicate whether the film is intended for daylight or indoor/tungsten balanced illumination. Modern film magazines have a code (termed DX) that allows camera backs to recognize the film speed automatically and set the control device to the proper ISO.

The color temperature of tungsten-halide lamps in modern microscopes varies between 2900 and 3200 K, depending on the voltage applied to the lamp filament. Film manufacturers offer film balanced for this illumination and usually indicate on the magazine that the film is intended for indoor or scientific use. Fuji offers a very nice transparency film, Fujichrome 64T, that has a rather slow emulsion speed intended for tungsten illumination, but is designed to perform well with *push* film processing. To push a film, it is first underexposed by one or several *f* stops (factors of 2), then the development time in the first developer is increased to compensate for this decrease in film density. This technique often will increase the color saturation of the image (19,37).

Daylight-balanced film, by far the most common film available at retailers at a wide variety of ISOs, can also be used with the microscope, provided that an appropriate filter is added to the light path. Manufacturers usually add a daylight conversion light filter to their microscope packages as standard equipment, and high-end research microscopes usually have this filter (called a daylight color temperature conversion filter) housed in the base of the microscope. Almost any color print or transparency film can be used for microscopy, provided that the daylight-balanced filter is in place for those films designed for a 5500-K color temperature or removed if the film is balanced for tungsten illumination (3200 K).

For many applications, securing the photomicrograph almost immediately is a necessity or a great advantage. Here, Polaroid films are unrivaled. These films are available for photomicrography in three sizes: 35-mm Polachrome (color transparency), and Polapan (black-and-white) in $3\frac{1}{4}'' \times 4\frac{1}{4}''$ film packs and $4'' \times 5''$ individual film packets (37,39). Larger formats are available in color (ID numbers 668/669 or 58/59) or black-and-white (ID numbers 667, 52, 665, 55, etc.).

The Polaroid large format films produce a paper negative and a paper positive print. After the positive print has been made, the paper negative is peeled away and discarded. This can be accomplished within a few minutes. The color films, except for the film pack 64T and the large format 64T, are balanced for daylight (5500 K) color temperature. All microscope manufacturers supply adapters for their photomicrographic cameras that accept large format Polaroid film sizes.

Polaroid black-and-white films 665 P/N and 55 P/N require special mention. These films produce a paper positive print and a transparent, polymer-based negative. If the negative is bathed in an 18% sodium sulfite solution for a few minutes, washed, and hung up to dry, it produces a permanent negative of high quality and high resolution suitable for use in a photographic enlarger (37,39).

Polaroid 35-mm films deserve more popular use. They are loaded into the usual 35-mm camera back and exposed in the typical manner according to their rated ISO and color temperature requirements. When the film is purchased, the container includes the film cassette and a small box of developer. The film is processed in a tabletop processor which measures approximately $4'' \times 5'' \times 9''$. All processing is carried out in daylight after the processor has been loaded and closed. The finished strip of positive color or positive black-and-white transparencies is removed and is ready to be examined and to be cut apart for mounting in frame holders for projection. The actual processing takes only five minutes and produces micrographs of quite accurate color and good resolution.

Recently, Polaroid has marketed a relatively inexpensive photomicrographic camera call the MicroCam. This camera is lightweight and can be inserted into one of the eyetubes or phototube of the microscope. It contains an exposure meter, an eyepiece, and an automatically controlled shutter. The camera accepts 339 color film packs or 331 black-and-white film packs. These films, approximately $3'' \times 3''$ in size, are self-developing and do not require peeling apart. Although the films have only about half the resolution of the more common large format Polaroid films, the resulting print is easily obtained and may be quite adequate for many applications.

DIGITAL PHOTOMICROGRAPHY

During the past decade, the quality of digital cameras has greatly improved and today, the market has exploded; at least 40 manufacturers offer a wide variety of models to suit every application. The cameras operate by capturing the image and projecting it directly onto a computer chip without using film. In recent years, the number of pixels of information that can be captured and stored by the best digital cameras approaches, but is still short of, the resolution available from traditional film (37). Digital images offer many opportunities for computer-controlled image manipulation and enhancement as well as the advantage of permanent digital storage. The highest quality digital cameras cost many thousands of dollars.

Selecting of an electronic camera must be preceded by carefully considering of its proposed use. This includes examining of fixed specimens or live specimens, the need for gray-scale or true color images, sensitivity to low light levels as in fluorescence, resolution, speed of image acquisition, use in qualitative or quantitative investigations, and the video feed rate into a computer or VCR (1,2,9,37).

The two general types of electronic cameras available for microscopy are the tube cameras (Vidicon family) or charge-coupled device (CCD) cameras. Either type can be intensified for increased sensitivity to low light. The SIT (silicon intensified target) or ISIT is useful for further intensification of tube cameras, and the ICCD is an intensified CCD camera. CCD cameras can also be cooled to increase sensitivity by giving a better signal-to-noise ratio. These kinds of cameras can be designed to respond to light levels undetectable by the human eye (2,37).

The criteria for selecting of an electronic camera for microscopy include sensitivity of the camera, quantum efficiency, signal-to-noise ratio, spectral response, dynamic

range capability, speed of image acquisition and readout, linearity of response, speed of response in relation to changes in light intensity, geometric accuracy, and ready adaptability to the microscope. A very important criterion for the newer digital CCD cameras is resolution. Current chips range from as few as 64×64 pixels up to 2048×2048 and higher for very specialized applications, but larger arrays are continuously being introduced; at some not too distant time, digital image resolution will rival that of film.

The purchaser must also decide whether the camera will operate at video rate (therefore making it easily compatible with such accessories as video recorders or video printers). Tube cameras and CCD cameras are available for video rate operation. CCD cameras can also deliver slow acquisition or high-speed acquisition of images. Scientific, rather than commercial grade, CCD cameras are the variety most suitable for research.

CCD cameras are usually small. They have no geometrical distortion, no lag, and good linearity of response. These cameras are also more rugged and less susceptible to mishandling than tube cameras. Each pixel of the CCD camera is a well for charge storage of incoming photons for subsequent readout. The depth of these wells relates to the quality of the response. Binning of adjacent pixels by joining them together into *super pixels* can be employed to speed readout in a slow-scan CCD camera (2,37).

An emerging technology that shows promise as the possible future of digital imaging is the active pixel sensor (APS) complementary metal oxide semiconductor (CMOS) *camera on a chip*. Mass production of CMOS devices is very economical and many facilities that are currently engaged in fabrication of microprocessors, memory, and support chips can be easily converted to produce CMOS optical sensors. Although CCD chips were responsible for the rapid development of video camcorders, the technology has remained a specialized process that requires custom tooling outside the mainstream of integrated circuit fabrication. CCD devices also require a substantial amount of support circuitry; it is not unusual to find five to six circuit boards in a typical video camcorder.

In the center of the APS CMOS integrated circuit is a large array of optical sensors that are individual photodiode elements covered with dyed filters and arranged in a periodic matrix. A photomicrograph of the entire die of a CMOS chip is shown in Fig. 28. High magnification views of a single "pixel" element (Fig. 28) reveal a group of four photodiodes that contain filters based on the primary colors red, green, and blue. Each photodiode is masked by a red, green, or blue filter in low-end chips, but higher resolution APS CMOS devices often use a teal (blue-green) filter in place of one of the green filters. Together, the four elements illustrated in Fig. 28 comprise the light-sensitive portion of the pixel. Two green filter masks are used because visible light has an average wavelength of 550 nm, which lies in the green color region. Each pixel element is controlled by a set of three transistors and an amplifier that work together to collect and organize the distribution of optical information. The array is interconnected much like memory addresses and data busses on a dynamic random access memory (DRAM) chip, so that the charge generated by photons that strike each individual pixel can be accessed randomly to provide selective sampling of the CMOS sensor.

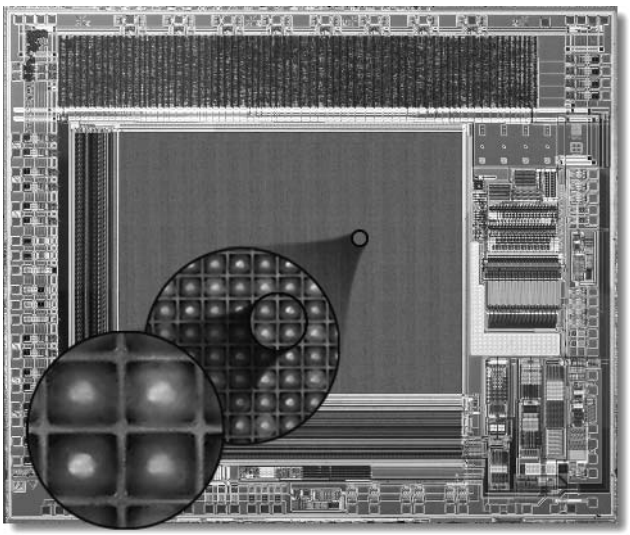


Figure 28. CMOS active pixel sensor used in the new rapidly emerging *camera on a chip* technology. The photomicrograph captures the entire integrated circuit surface and shows the photodiode array and support circuitry. The insets illustrate progressively higher magnification of a set of pixel elements and a single pixel element composed of four dyed photodiodes.

The individual amplifiers for each pixel help reduce noise and distortion levels, but they also induce an artifact known as *fixed-pattern noise* that arises from small differences in the behavior of individual pixel amplifiers. This is manifested by reproducible patterns of *speckle* behavior in the image generated by CMOS active pixel sensor devices. A great deal of research effort has been invested in solving this problem, and the residual level of noise has been dramatically reduced in CMOS sensors. Another feature of CMOS active pixel sensors is the lack of column streaking due to pixel bloom when shift registers overflow. This problem is serious for the CCDs in most video camcorders. Another phenomenon, known as *smear*, which is caused by charge transfer in CCD chips under illumination, is also absent in CMOS active pixel sensor devices.

Assisting the CMOS device is a coprocessor that is matched to the APS CMOS to optimize the handling of image data. Incorporated into the coprocessor is a proprietary digital video processor engine that can perform automatic exposure and gain control, white balance, color matrixing, gamma (contrast) correction, and aperture control. The CMOS sensor and coprocessor perform the key functions of image capture, digital video image processing, video compression, and interfacing to the main computer microprocessor.

The primary concerns about CMOS technology are the rather low-quality, noisy images that are obtained compared with those from similar CCD devices. These are due primarily to process variations that produce a slightly different response in each pixel, which appears in the

image as *snow*. Another problem is that the total amount of chip area that is sensitive to light is less in CMOS devices and thus makes them less sensitive to light. These problems will be overcome as process technology advances, and it is very possible that CMOS devices will eclipse the CCD as the technology of choice in the very near future.

If you have prized transparencies and negatives collected over a long period of time, many of the better camera stores can take these and put them onto a Kodak Photo CD, which is the same size as an audio compact disc (CD). The Photo CD can hold up to 100 images that are stored at several levels of resolution. Digital images recorded on the Photo CD can be displayed on a good monitor by a Photo CD player. If you have a computer and a program such as Adobe Photoshop, Corel Photo Paint, or Picture Publisher and a CD drive, you can open the images on your computer screen, manipulate and/or enhance the images, and then print the images using a digital printer — all without a darkroom! Kodak, Fuji, Olympus, Tektronix, and Sony market dye-sublimation printers that can produce prints virtually indistinguishable from those printed by the usual color enlarger in a darkroom.

Nikon, Olympus, Polaroid, Kodak, Agfa, Microtek, and Hewlett-Packard produce 35-mm negative and positive transparency scanners and flat-bed scanners, that directly scan transparencies or negatives or prints into your computer for storage or manipulation. Images can be stored on the computer hard drive or stored on floppy disks in JPEG or TIFF files. Because the storage of floppy disks is limited to 1.44 megabytes, many micrographers now store images on Zip disks or magneto-optical drive disks; these can hold many images to sizes of 10 megabytes or more. Another popular storage medium that is quickly gaining widespread popularity is the recordable CD-ROM. Magneto-optical disks or CD-ROMs can be given to commercial printers and then printed with stunning color accuracy and resolution.

A photomicrograph is also a photograph. As such, it should reveal information about the specimen, and it should also do so with attention to the aesthetics of the overall image. Always try to compose photomicrographs that have a sense of the balance of the color elements across the image frame. Use diagonals for greater visual impact, and scan the frame for unwanted debris or other artifacts. Select a magnification that readily reveals the details sought. Remember to keep detailed records to avoid repeating mistakes and to help by reviewing images that are several years old. Excellent photomicrographs are within the capability of most microscopists, so pay attention to the details, and the overall picture will assemble itself.

Acknowledgments

The authors thank the Molecular Expressions (http:/microscopy.fsu.edu) web design and graphics team for the illustrations in this article. We also thank George Steares of Olympus America, Inc. for providing advice and microscopy equipment used to illustrate of this manuscript. Assistance from Chris Brandmaier of Nikon USA, Inc. in providing equipment for reflected-light differential interference contrast microscopy is also appreciated. Funding for this research was provided, in part, by the National High Magnetic Field Laboratory (NHMFL), the State of Florida, and the National Science Foundation through cooperative grant No. DMR9016241. We also thank James Ferner, Samuel Spiegel, and Jack Crow for the support of the Optical Microscopy program at the NHMFL.

BIBLIOGRAPHY

1. B. Herman and J. J. Lemasters, eds., *Optical Microscopy: Emerging Methods and Applications*, Academic Press, NY, 1993.
2. S. Inoué and K. R. Spring, in *Video Microscopy: The Fundamentals*, 2nd ed., Plenum Press, NY, 1997.
3. S. Bradbury and B. Bracegirdle, in *Introduction to Light Microscopy*, BIOS Scientific, Oxford, UK, 1998.
4. E. M. Slayter and H. S. Slayter, in *Light and Electron Microscopy*, Cambridge University Press, Cambridge, UK, 1992.
5. M. Pluta, in *Advanced Light Microscopy (3 vols.)*, Elsevier, NY, 1989.
6. M. Abramowitz, in *Fluorescence Microscopy: The Essentials*, Olympus America, Melville, NY, 1993.
7. B. Herman, in *Fluorescence Microscopy*, 2nd ed., BIOS Scientific, Oxford, UK, 1998.
8. F. W. D. Rost, in *Fluorescence Microscopy (2 vol.)*, Cambridge University Press, NY, 1992.
9. G. Sluder and D. E. Wolf, eds., *Methods in Cell Biology, vol. 56: Video Microscopy*, Academic Press, NY, 1998.
10. C. J. R. Sheppard and D. M. Shotton, in *Confocal Laser Scanning Microscopy*, BIOS Scientific, Oxford, UK, 1997.
11. S. W. Paddock, ed., *Methods in Molecular Biology, vol. 122: Confocal Microscopy, Methods and Protocols*, Humana Press, Totowa, NJ, 1999.
12. J. B. Pawley, ed., *Handbook of Biological Confocal Microscopy*, 2nd ed., Plenum Press, NY, 1995.
13. S. Bradbury, in *The Evolution of the Microscope*, Pergamon Press, NY, 1967.
14. *Anticipating the Future*, Zeiss Group Microscopes Business Unit, Jena, Germany, 1996.
15. G. Nomarski, *J. Phys. Radium* **16**, 9S–11S (1955).
16. R. Hoffman, *J. Microscopy* **110**, 205–222 (1977).
17. S. Inoué and R. Oldenbourg, in M. Bass, ed., *Handbook of Optics*, vol. II, 2nd ed., McGraw-Hill, NY, 1995, pp. 17.1–17.52.
18. M. Abramowitz, in *Contrast Methods in Microscopy: Transmitted Light*, Olympus America, Melville, NY, 1987.
19. J. G. Delly, in *Photography Through The Microscope*, 9th ed., Eastman Kodak Co., Rochester, NY, 1988.
20. H. G. Kapitza, in *Microscopy From The Very Beginning*, Carl Zeiss, Oberkochen, Germany, 1994.
21. H. W. Zieler, in *The Optical Performance of the Light Microscope* (2 vols.), Microscope Publications, Chicago, 1972.
22. J. James and H. J. Tanke, in *Biomedical Light Microscopy*, Kluwer Academic, Boston, 1991.
23. M. Abramowitz, in *Optics: A Primer*, Olympus America, Melville, NY, 1984.
24. M. Abramowitz, in *Microscope: Basics and Beyond*, Olympus America, Melville, NY, 1987.
25. M. Abramowitz, in *Reflected Light Microscopy: An Overview*, Olympus America, Melville, NY, 1990.
26. R. McLaughlin, in *Special Methods in Light Microscopy*, Microscope Publications, Chicago, 1977.
27. A. Tomer, in *Structure of Metals Through Optical Microscopy*, ASM International, Boulder, CO, 1990.

28. G. H. Needham, in *The Practical Use of the Microscope*, Charles C. Thomas, Springfield, IL, 1958.
29. C. P. Shillaber, in *Photomicrography in Theory and Practice*, Wiley, NY, 1944.
30. F. Zernike, *Physica* **9**, 686–693 (1942).
31. F. Zernike, *Physica* **9**, 974–986 (1942).
32. E. A. Wood, in *Crystals and Light: An Introduction to Optical Crystallography*, 2nd ed., Dover, NY, 1977.
33. R. E. Stoiber and S. A. Morse, in *Crystal Identification with the Polarizing Microscope*, Chapman & Hall, NY, 1994.
34. P. C. Robinson and S. Bradbury, in *Qualitative Polarized-Light Microscopy*, Oxford Science, NY, 1992.
35. A. F. Hallimond, in *The Polarizing Microscope*, 3rd ed., Vickers Instruments, York, UK, 1970.
36. S. Bradbury and P. J. Evennett, in *Contrast Techniques in Light Microscopy*, BIOS Scientific, Oxford, UK, 1996.
37. M. Abramowitz, in *Photomicrography: A Practical Guide*, Olympus America, Melville, NY, 1998.
38. R. P. Loveland, in *Photomicrography: A Comprehensive Treatise (2 vols.)*, Wiley, NY, 1970.
39. *Photomicrography: Instant Photography Through the Microscope*, Polaroid, Cambridge, MA, 1995.

OVER THE HORIZON (OTH) RADAR

S. J. Anderson
R. I. Barnes
S. B. Colegrove
G. F. Earl
Surveillance Systems Division
Defence Science and Technology Organization
Australia

D. H. Sinnott
Cooperative Research Center for Sensor Signal and Information Processing
Mawson Lakes
South Australia, Australia

FUNCTION, PERFORMANCE AND IMPLEMENTATIONS OF OVER-THE-HORIZON RADAR

In this introductory section, a description is given of over-the-horizon radar in general terms, and current implementations are indicated. Subsequent sections expand in more detail on propagation and scattering issues, operation of OTH radars and system and subsystem design.

Fundamentals of over-the-Horizon Radar

Over-the-horizon radar (OTHR) is an application of radar technology that exploits the phenomenon of electromagnetic wave reflection by the earth's *ionosphere* to allow radar operation beyond the local horizon (1). Excluded by this definition are other radar applications that may give consistent or ephemeral over-the-horizon performance. Two particular examples excluded are *surface wave radar* that exploits the fact that a surface wave in the high-frequency band can be propagated over salt water to ranges beyond the horizon and *microwave radar ducting* in which specific atmospheric conditions can trap an electromagnetic wave within a narrow layer and allow it to propagate to anomalous ranges.

The ionosphere is a solar-induced shell of ionization in the upper atmosphere with which electromagnetic waves interact so that, within a range of frequencies, the waves are reflected to earth. Figure 1 illustrates by a ray model how such an electromagnetic wave may be reflected by the ionosphere and thereby propagate *over the horizon*. An area or object illuminated by this electromagnetic wave returns part of its scattered energy by an (almost) reciprocal ionospherically reflected path and this signal, when received and processed, allows radar operation.

The actual operating range of an OTHR system depends on the frequency used within the band for which ionospheric reflection is significant — low frequencies are reflected more strongly, so they illuminate the surface comparatively close to the radar, and conversely for higher frequencies. The limit on maximum range for this *single-hop* mode of OTHR operation, determined by geometry and practical considerations, is about 3000 km, but *multihop* operation, in which the electromagnetic wave is sequentially reflected between the earth and the ionosphere, can multiply this range. The practicality of multihop radar operation is quite limited, principally by the signal distortion imposed during multiple reflections, geometric spreading, and absorption losses.

For any given frequency of operation, there is a minimum range for which ionospheric reflection to earth is effective. Figure 1 illustrates how rays propagate further into the ionosphere, as the elevation angle increases, until they penetrate the ionosphere and do not return to earth. This establishes the *skip zone* shown. For a number of practical reasons, a minimum skip distance of about 1000 km is typical of implemented systems, and thus the range of OTHR systems is conventionally stated as 1000–3000 km.

The generic techniques of radar and its signal processing, as described elsewhere in this encyclopedia, are, in principle, applicable to OTHR. In practice, severe constraints are imposed by the ionosphere as the propagation medium on the applicability of many techniques. However, the basic approach is in common with other forms of radar: a suitably designed waveform

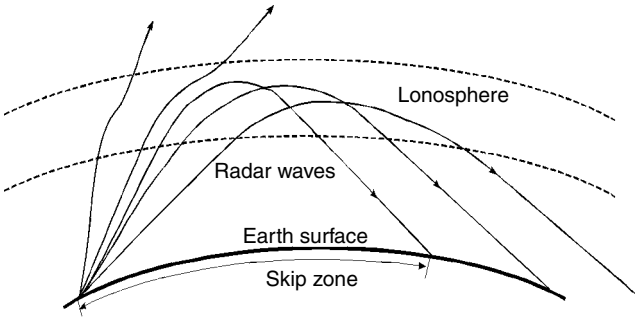


Figure 1. Typical ray paths for HF radio waves transmitted from the earth's surface.

is modulated onto a transmitted carrier wave and processing of the returned echoes allows defining both range (by determining of elapsed time) and direction (by using directional antennas) of the region under radar surveillance. As pointed out later, a third dimension, that of *Doppler*, is available to the OTHR operator and is a crucial parameter in deriving information from the radar returns. In particular, Doppler information allows discriminating of moving targets from stationary (or slowly moving) *clutter*. Clutter is the term used in radar parlance to describe returns from extended areas and is normally seen as contamination to be minimized, by processing in favor of the discrete target return. Of course, in some surface-mapping OTHR applications, this clutter is the object of the surveillance operation.

The ionosphere provides both the support for OTHR operation and the principal limitation on performance. The typical plasma density of the ionosphere means that the operating carrier frequency of an OTHR is constrained to lie approximately within the electromagnetic wave band designated as *high frequency (HF)*, 3–30 MHz, although practical considerations set the lower limit for most OTHR realizations at around 6 MHz. The frequency chosen for a particular application depends on many factors, but in broad terms it is determined by the range at which operations are desired, as explained in more detail in what follows.

Operation in the HF band implies wavelengths from 10–100 m. It follows that transmitting and receiving antenna systems need to be large to realize adequate directive properties, which derive from multiwavelength antenna apertures. The need for angular resolution has driven conventional radar to employ frequencies well above the HF band, where antenna apertures of tens or hundreds of wavelengths are within the dimensions feasible for mechanically steerable antennas. In the HF band, acceptable angular resolution demands antenna systems of hundreds or thousands of meters long. So mechanical steering is out of the question and electronic beam steering, used only in the most demanding (usually military) conventional radar applications, is mandatory for OTHR. An antenna aperture in two dimensions is desirable to realize a pencil-beam and hence two-dimensional angular localization, but to date most OTHR applications have used linear arrays (usually of modestly directional elements) for cost reasons and accepted the consequences of fan rather than pencil beams.

Thus, the *typical* OTHR has very large, ground-based-array antenna systems with electronic steering that make it appear physically quite unlike conventional radars. Another point of departure from conventional radars is that most realizations use separate transmit and receive sites. Whereas most conventional radars use a common antenna in which operation is *monostatic* (the single antenna switches between transmit and receive functions), many OTHR systems use continuous transmission of a repetitive waveform and hence need spatially separated transmit and receive antenna sites. The arrays need to be separated by sufficient distance that the direct leakage to the receiving system of the transmitted signal, by line-of-sight or ground-wave propagation, does not interfere with reception: a separation of 60–100 km is typical. A cost-attractive architecture has a relatively compact, transmitting linear antenna array (100–200 m) and a receiving array about an order of magnitude larger. Then, the transmitter illuminates a region within which a number of much narrower simultaneous receive beams are focused.

Besides being constrained to the broad frequency band necessary for ionospheric propagation, an OTHR must also allow for the fact that the ionosphere is a highly variable and unpredictable medium. For acceptable OTHR performance, continuous, real-time, ionospheric assessment (*sounding*) is necessary, as is the ability of the radar system to adapt to the changes. The day–night variability of the ionosphere requires, typically, a 5 : 1 change in frequency for comparable operational range, but both longer term sunspot-related effects and short-term disturbances are imposed on this diurnal ionospheric cycle.

The large propagative distance and hence very weak returned signals, high external noise in the HF band, and the domination of surface-reflected clutter in the returned radar signal mean that OTH radars rely for their operation on high spectral resolution (Doppler) signal processing. Detection of aircraft with a significant radial velocity component toward or away from the radar is the most common application of OTHR, for which coherent illumination of several seconds suffices to satisfy spectral resolution requirements. If the propagative environment is sufficiently stable to allow coherent illumination over tens of seconds, slower large scatterers, such as ships, can be detected and the fine-grain Doppler characteristics of the sea surface can be mapped. It is artificial to propose hard definitional boundaries for coherent illumination times, but it is common to designate operations as *air* mode (seconds) and *surface* mode (tens of seconds).

Table 1 summarizes the numerical values of major radar design parameters for a typical microwave ground-based military air-defense radar compared with those of a typical OTHR. The table illustrates the distinctiveness of OTHR design.

OTHR Implementations

The principal difficulties in realizing OTHR have been the generation of suitably stable signals and computing power sufficient to attack what is a highly intensive signal processing task. Both of these difficulties paced the evolution of OTHR, although the technology of ionospherically supported propagation beyond the visible horizon has been accessible since the earliest days of radio. It is still heavily used in *shortwave* broadcast services and for point-to-point narrowband communications.

Some limited OTHR systems were operational as early as the 1950s, but it was not until the 1960s that advances in signal generation and computing allowed fielding OTHR systems of substantial capability (2). In the years following, both Western and Soviet systems were developed for long-range defense surveillance, and developments were ongoing. OTHR applications to sea-state sensing have followed, building on enhanced

Table 1. OTHR and Microwave Radar Parameters

Parameter	Microwave Radar	OTHR
Operating frequency	1200–1400 MHz	6–30 MHz
Average transmitted power	4 kW	200 kW
Duty cycle	<16%	100%
Instantaneous bandwidth	1.5 MHz	10 kHz
Waveform repetition frequency	300–1000 Hz	10–50 Hz
Antenna dimensions	Single 8 × 8 m low sidelobe planar array	Separate linear phased arrays: transmitter 100–200 m, receiver 1000–3000 m
Azimuthal coverage	360°	90°
Angular resolution	1.5° (pencil beam)	0.3–1.5° in azimuth,
Cross-range resolution	5 km at 200-km range	10–50 km at 2000-km range.
Instrumented range	10–400 km	1000–3000 km
Range resolution	100 m	15 km
Doppler resolution	25 Hz	1.0–0.01 Hz
Dynamic range (clutter peak to noise floor)	60–70 dB	80 dB
Steering — elevation	Electronic	None (fan beam)
Steering — azimuth	Rotation at 6 rpm	Electronic

knowledge of the ionosphere and access to facilities built with defense funds (3,4).

OTHR PROPAGATION AND THE IONOSPHERE

The Ionosphere and Basic Propagative Principles

The earth's ionosphere, a partial magnetoplasma, is produced by photons and particles that are emitted from the sun and ionize a portion of the neutral upper atmosphere. The lowest altitude at which ionization is encountered is about 60 km, and it extends from there to many earth radii. The region from 90–400 km is the most important for OTHR, because at these altitudes, electromagnetic wave reflection is significant. In this region, the bulk of the ionization is caused by the action of extreme UV and X rays on atoms and molecules of oxygen and nitrogen.

The varying frequency bands of the sun's emission and the physical chemistry of the atmosphere are jointly responsible for relatively distinct ionospheric regions or layers. Those most significant for OTHR operation are designated the D, E, and F regions. The D region does not provide useful reflection of HF waves but is associated with absorption. The E region occurs from 90–150 km, has an ionization peak around 105 km, and provides useful OTHR propagation during the day to about 1800 km. Above this lies the F region, within which there are frequently separate layers. The most important layer for OTHR, because it is the most intense layer in the ionosphere, is designated the F_2 layer which peaks at around 300 km. An important feature of the F_2 layer is that it is sustained throughout the night by ion transport and the low ionization recombination rate.

Free electrons, rather than the far more massive positively charged ions, are the ionospheric constituent of dominant influence on HF propagation. Electron motion induced by the electric component of an incident HF electromagnetic wave leads to a secondary field that may be described as the polarization field of a dielectric of refractive index n less than one, according to the approximation,

$$n^2 \approx 1 - \frac{80.5N_e}{f_c^2} \tag{1}$$

where N_e is the electron density (electrons per m^3) and f_c is the carrier frequency (Hz) of the incident wave. If the electron density gradient is positive with height, the refractive index decreases with height, and an electromagnetic wave from below is deflected increasingly from the normal. If the layer is sufficiently thick that the wave is propagating horizontally before the peak of N_e is reached, then the wave returns to earth. If N_e peaks before the wave is horizontal, then it passes through the layer and is deflected toward the normal in the decreasing electron density environment above the N_e peak. The penetrating rays of Fig. 1 illustrate this. Equation (1) also indicates that waves of lower f_c, for which n departs more from one, are reflected more.

This simplified wave description neglects the influence of the geomagnetic field, which causes the ionospheric refractive index to have tensor properties and exhibit birefringence. The two modes produced through this mechanism, designated O and X, have locally orthogonal polarizations. Further, collision of electrons with other particles contributes a loss mechanism that makes the refractive index complex. Losses are more significant lower in the ionosphere where the increased neutral density causes higher collision frequencies and at certain wave frequencies where the ionosphere appears highly dispersive.

For most situations, OTHR propagation in the ionosphere can be approximated by Hamiltonian ray theory (5) in that the electromagnetic radiation can be thought of as pencils that are bent by ionospheric gradients. Even with this simplification, OTHR propagation may be complicated: for a single target, the separate layers of the ionosphere can present multiple paths, and each of these

paths can contribute O and X modes. Outward and inward propagative paths can be a combination of any of the one-way paths, and so at times, a single target can appear as multiple detections at different ranges because of the plethora of paths of different propagative delay. At longer ranges, target paths can be multihop. It is the job of the radar's *coordinate registration* system to model these effects and establish the true position of targets (6).

Often the O and X modes are not resolved in range by an OTHR. Because the relative phase paths of the two modes can change in time, the combined elliptical polarization appears to rotate in space. This effect, known as Faraday rotation, is pervasive and causes the signal to fade when detected on linearly polarized antennas.

Propagative Morphology

Backscatter ionograms best describe the coverage that is provided by the ionosphere for OTHR. The backscatter ionogram is a representation of the clutter signal that is backscattered from the environment as a function of *group range* (half the transmit-to-receive delay multiplied by the speed of light in a vacuum) and carrier frequency. It is produced by a backscatter sounder, which is a swept frequency device that can measure delay (7). Figure 2 shows a typical daytime response. The clutter is calibrated at the receiver input and is displayed in absolute units. The main lobe extends to 3000 km and is due to the single-hop F_2 propagation.

Because the behavior of the ionospheric layers and the electromagnetic wave propagation that they support are strongly controlled by the sun, the propagation has diurnal, seasonal, and sunspot cycle variations (8). The electron density of the F_2 layer can vary over two orders of magnitude during these cycles, and so an intimate knowledge of the current state of ionospheric propagation is vital for successful operation of an OTHR. It has been shown that existing ionospheric models are inadequate for describing the real variability (9).

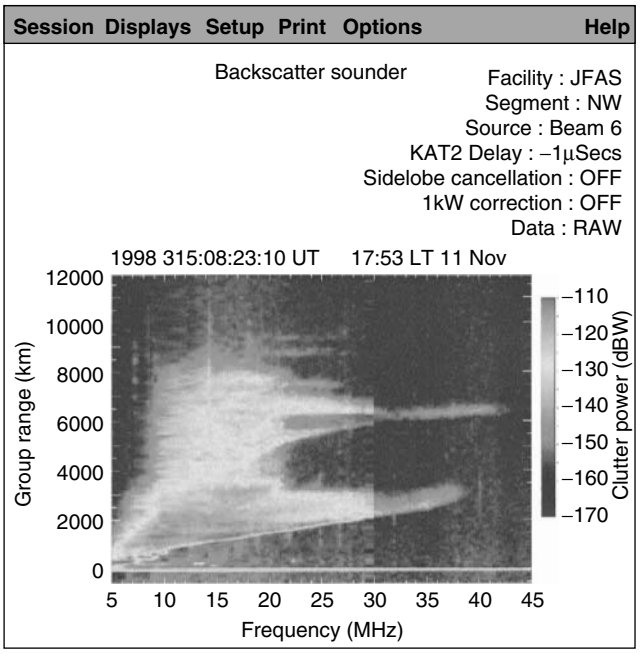


Figure 2. A backscatter ionogram recorded during day time, looking north across the equator. Both single-hop and chordal propagative modes are evident.

Variation in the orientation of the magnetic field naturally separates the ionosphere into three phenomenologically different regions: midlatitude, equatorial (low-latitude) and polar (high-latitude). The equatorial region is characterized by an upwelling of the ionosphere, the *equatorial anomaly*. The shape of the anomaly is such that north–south propagation across it can be chordal (ground to ionosphere to ionosphere to ground).

The polar ionosphere is the most disturbed region. The magnetic field at the poles is vertically oriented and projects far from the surface, where it interacts strongly with the solar wind. Propagation across the poles is fraught with difficulty and can result in deviations of the gross angle of arrival due to walls of ionization associated with polar electrodynamics.

The midlatitude ionosphere is relatively stable and exhibits the largest spatial correlation distance for F_2 layer electron density and peak height measurements. However, it still exhibits many of the disturbances that are detailed below.

Effects on OTHR of Ionospheric Dynamics

The dimensions in OTHR measurement space are group range, angle(s) of arrival, and Doppler shift. Ionospheric effects that perturb these domains introduce errors into the estimate of the target state vector. A number of these effects are catalogued here; some of them are infrequent but have a major impact on OTHR performance, whereas others, more commonly encountered, have more subtle impacts.

The ionosphere is constantly being disturbed by the propagation of atmospheric gravity waves (10). These waves appear as shifting neutral winds in the ionospheric medium. The subsequent motion of electrons is determined by the geomagnetic field and by the drag induced by the wind. Other forces are also involved, but the net effect is that the electron density is alternately enhanced and rarified in a wavelike fashion that leads to calling the modulation a *traveling ionospheric disturbance* (TID) (10). TIDs cause apparent motion of an OTHR-displayed target because they distort the angle of arrival and the Doppler shift of the incoming wave. The distortion of the angle of arrival can be as large as several degrees.

Sporadic E (E_s), is a thin, erratic layer that forms in the E region of the ionosphere and is most prevalent between 90 and 120 km. Its behavior is strongly controlled by the orientation of the magnetic field, and its production mechanism is quite different at mid, low and high latitudes. At midlatitudes, E_s is chiefly a summer daytime phenomenon. E_s, which can be the most intense layer in the ionosphere, has several important effects on OTHR. First, it can provide very good propagation for surface-mode detection because it imparts very little Doppler distortion. Second, it can obscure propagation to the F_2 layer, which has the negative impact of limiting the OTHR maximum range to about 2000 km. Third, because of its erratic temporal and spatial nature, E_s is very hard to

model and consequently can play havoc with the coordinate registration accuracy of an OTHR. An image of E_s patches produced by the Jindalee OTHR in Australia is shown in Fig. 3.

The *terminator* is the boundary between day and night in the ionosphere. Large systematic distortions of the Doppler and the angle of arrival occur with the passage of the sunset and sunrise terminators because of the abrupt change in ionization production, as illustrated in Fig. 4 (11). The terminator is predictable, and its effects can be modeled and subsequently subtracted from the measurements. This is part of the coordinate registration process, but even this level of propagative modeling is difficult.

A more pernicious form of distortion is produced by small-scale ionospheric irregularities. The clutter signal that is received by the radar after propagating through or backscattered by these irregularities may be spread in the Doppler domain. One particular form of spread clutter is actually caused by reflection from *meteor* trails rather than from the ionosphere. Meteor trail lifetimes are such that they produce transitory reflections at HF, which spread their signal energy in Doppler space. Strategies for handling spread clutter exist and include radar waveform parameter adjustment, adaptive signal processing, and antenna pattern control.

A very severe, albeit rare, ionospheric perturbation to OTHR operation is *short-wave fadeout*. This occurs when a disturbance on the sun causes an unusually intense emission of UV and X radiation. The effect in the ionosphere is to increase the *D*-region electron density dramatically, which in turn increases absorption. This effect can be so serious that the radar is disabled for as long as an hour through a lack of sensitivity. A related but less dramatic effect is a *magnetic storm*, caused by the interplay of the earth's magnetic field with the solar wind. This manifests itself at midlatitudes by depressing the intensity of the F_2 layer. Magnetic storm effects can last for days and can be so severe that propagation is limited to around 2000 km via the E and F_1 layers.

Some unusual propagative paths can affect OTHR operation through *range folding* — the aliasing of longer range returns, whereby they are mapped into shorter ranges through the ambiguity inherent in repetitive waveforms. Ionospheric irregularities on the other side of the earth or from the polar regions produce spread clutter this way. Radiation that has traveled around the world and appears on the OTHR display aliased onto short ranges is a particularly quirky example.

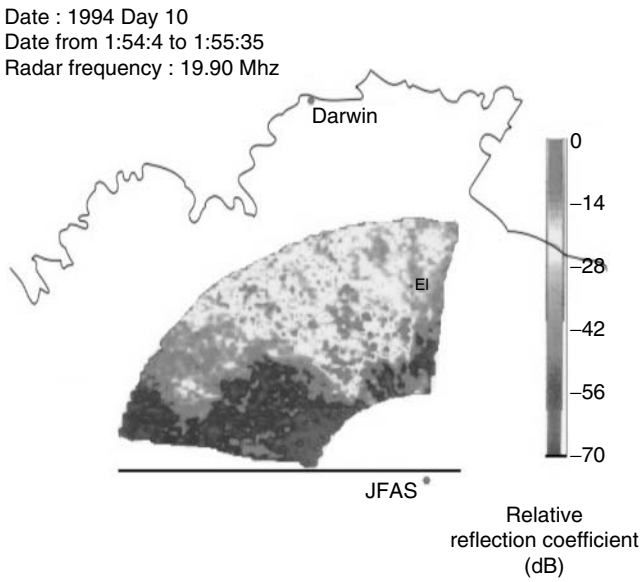


Figure 3. Sporadic *E*-layer patches mapped by the Jindalee OTHR.

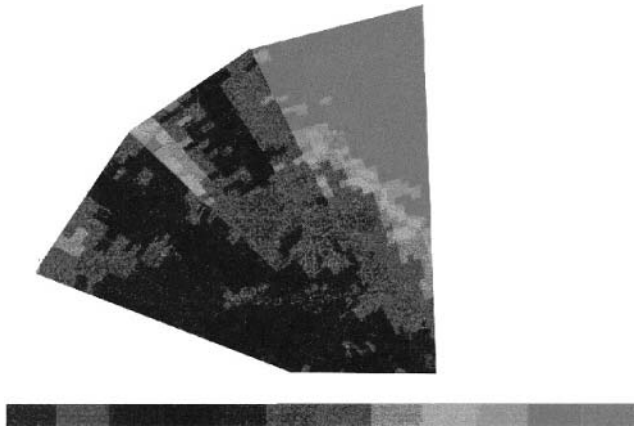


Figure 4. A map of ionospheric Doppler shift recorded at 0647 LT as the dawn terminator crosses the radar footprint.

RADAR SCATTERING PHENOMENOLOGY IN THE HF BAND

As for any radar, the ability of the OTHR sensor to derive information about entities in its field of view depends on the interaction of the probe signals — in this case HF electromagnetic waves — with those entities. In most HF radar applications, the interaction may be viewed as a localized electromagnetic scattering process, where the incident electromagnetic field is taken as a plane wave. In common with the terminology used in other branches of radar, the directly backscattered power is represented by describing an object in terms of its aspect-dependent *radar cross section* (RCS): the area for which the power from the incident plane wave would be sufficient to produce, by omnidirectional radiation, the backscattered power density observed. A *scattering matrix* description generalizes RCS for differing incidence and reflection angles and polarizations.

OTHR wavelengths are sufficiently large that most discrete objects of interest are comparable with or smaller than a wavelength. Scattering by conductors comparable in size to a wavelength is dominated by resonances in which the radar cross section can be a sensitive function of frequency. This means that the strength of the radar return is a poor guide to the size of an object — a wire resonant at the OTHR's carrier frequency can appear to the radar as an echo of strength comparable with that

of a large aircraft. For conductors much smaller than a wavelength, the radar cross section, in the Rayleigh scattering regime varies inversely as the fourth power of the illuminating wavelength (12).

Calculation of HF radar cross sections of metallic targets, such as aircraft and ships, can reasonably assume that the bodies are perfectly conducting. Then, there are many established analytical and numerical methods that can be used to compute the scattered field. The most widely used technique is the method of moments (13,14) for which many computer code implementations are in general use. An example of such calculations is presented in Fig. 5 which shows the squared magnitudes of the elements of the scattering matrix of a Boeing 727 aircraft in free space at a radar frequency of 15 MHz.

As noted previously, a typical OTHR design spatially separates the transmitter site from the receiver site so that they subtend a small but nonzero angle at the target (*quasi-monostatic*). For almost all practical situations, *monostatic* RCS calculations are applicable, despite this departure from exactly monostatic conditions. Figure 5 illustrates this point by including full bistatic coverage in the azimuth, though a single common elevation angle has been adopted. Monostatic solutions correspond to trailing diagonals.

Free-space radar cross sections are normally used to model OTHR scattering behavior of targets above the earth's surface that treat the various multipath scattering processes separately. This approximation fails when the target altitude becomes less than a few wavelengths and a more complicated formulation of the scattering problem is required.

Discrete targets are a major application of OTHR, but there is also a keen interest in echoes from extended targets, such as the earth's surface and the ionospheric

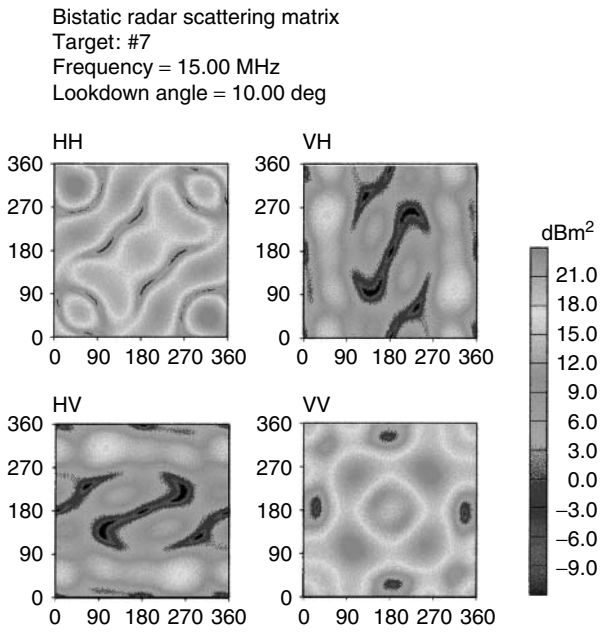


Figure 5. The copolar and cross-polar RCS of a Boeing 727 aircraft in free space. The figures show the full bistatic RCS in the azimuth for a look-down angle of 10° from horizontal.

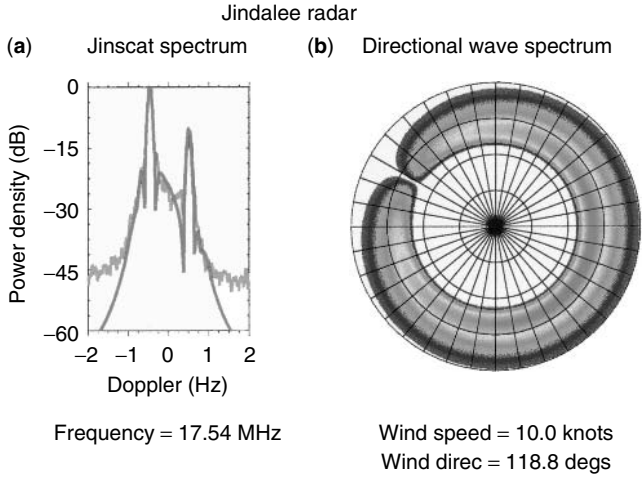


Figure 6. (**a**) The directional wave spectrum corresponding to a Pierson–Moskowitz wave-number spectrum with a simple spreading pattern. (**b**) predicted radar Doppler spectrum for the wave spectrum shown in (**a**) superimposed on a measured spectrum obtained under similar sea conditions.

plasma itself. The ocean surface is of particular interest because of the characteristic Doppler modulation imposed on the radar signals by ocean wave dynamics. The direct problem of scattering from a given ocean surface, specified by its directional wave spectrum, was originally solved for grazing incidence backscatter at vertical polarization (15); subsequently, the theory was extended to handle arbitrary geometries and polarizations, as illustrated in Fig. 6. This example shows (Fig. 6a) a polar representation of a model sea directional wave spectrum, where the wave number is plotted as the radial coordinate and the (sea) wave direction as the angular coordinate and (Fig. 6b) a measured Doppler spectrum, superimposed on a theoretical Doppler spectrum (the best fit solution to the measured Doppler spectrum from a parametric model). The directions in Fig. 6a are chosen so that "12 o'clock" corresponds to the bearing toward the radar. The wave spectral density is color coded, red indicates the strongest components. The two strong peaks in the Doppler spectra are the *Bragg lines* that arise from single resonant (Bragg) scattering from advancing and receding waves whose wavelength satisfies the condition for constructive interference.

Various techniques have been developed to solve the inverse problem and hence obtain an estimate of the directional wave spectrum, each variant relies on different assumptions about the form of the solution (16–18).

OTHR IN OPERATION

Discrete Target Surveillance and Monitoring

Tasking

In target surveillance and monitoring, an OTHR typically sequentially *spotlights* regions of coverage that are small compared to the potential surveillance area. The radar *dwells* on one region for a period of time to integrate coherently (usually seconds) and then moves on to the

next region. For normal air surveillance, an OTHR usually needs to *revisit* a region every 30 seconds or so to maintain tracks. Several effects conspire to constrain the area that an OTHR can survey within the revisit time. The limited angular extent of the transmit and receive patterns, the limited range depth provided by the ionosphere on a constant frequency, and signal processing constraints are the main issues. Therefore, it is very important to have a prioritized set of tasks for an OTHR so that the needs of the tasking agency are attended to as well as possible.

Radar Management

The management of an OTHR to achieve the required tasking most effectively depends on the broader surveillance architecture. For instance, if the radar is part of a network of similar or different sensors, then tasks must be assigned and scheduled across candidate sensors. Some OTHRs have inherent limitations that reduce the range of options and hence make tasking a much simpler exercise.

Setting up and scheduling tasks for a single OTHR requires an operator interface with the radar system. After defining the tasking area, the operator must balance the priorities, and decide on the radar parameters to be used. This usually involves a trade-off between radar sensitivity and shorter revisit times. Depending on the sophistication of the radar, the operator may have to decide on the following parameters:

Aperture Size. Generally, a larger aperture will provide greater sensitivity and resolution. However, if the radar can split apertures and perform multiple tasks at once, using the whole aperture may unnecessarily impact the revisit time.

Waveform Type. Waveform type might be varied for many reasons, for example, some waveforms offer protection against forms of spread Doppler clutter.

Waveform Bandwidth. The waveform bandwidth controls the range resolution of the radar. In general, a higher bandwidth is preferable because this provides a greater chance of ionospheric path resolution and can provide extra sensitivity in a clutter background. However, large bandwidth requires greater processing capability and increases the likelihood of interference from and to other users of the HF spectrum. Ionospheric dispersion provides a limit to the usefulness of increasing bandwidth by setting a physical limit on available bandwidth.

Waveform Repetition Frequency (WRF). This determines the ambiguity depths in range and Doppler. A higher WRF decreases the unambiguous range depth but increases unambiguous Doppler space. The WRF is often forced lower to avoid range-folded clutter, but this opens the possibility of Doppler-aliased clutter that produces regions of OTHR *blindness*s at typical target speeds.

Dwell Coherent Integration Time. Increasing the coherent integration time increases the target *signal-to-noise ratio (SNR)* but impacts the revisit time or the radar coverage potential.

Signal Processing Steps. Basic signal processing produces peaks in azimuth–Doppler–range space corresponding to potential targets. Additional processing can be used to enhance the OTHR target detection capability. For instance, processing to mitigate the effects of transient returns from meteors and lightning may be incorporated, if necessary.

Propagative Paths. It is necessary to know the propagative characteristics as the task is set up. An OTHR is likely to be tasked in ground coordinates, but the radar hardware must be set up in radar coordinates, that is, in steer direction and transmit-to-receive delay. Therefore, coordinate conversions from ground coordinates to radar coordinates are required to ensure that the radar surveys the intended region.

Task Priority. Assigning the various tasks different priorities is a way of altering the task revisit times in a priority-driven system.

Revisit Requirement. Although there may be no means for directly changing this variable, the operator will usually monitor the revisit time for all tasks, especially as a new task is installed, to ensure that the radar is not being overused.

Receiver Attenuation. At night, absorption of electromagnetic waves decreases and incoming signals can be very large, requiring attenuation to ensure that the receiving system does not distort or overload.

Of all the decisions that the operator makes, the choice of operating frequency is probably the most important. Most OTHR systems provide sensor information about the environment to assist operators in their choice. This is the role of the propagative management system (7). The information required to make the best choice of frequency includes the following:

Propagative Sensitivity. The prime concern in frequency selection is to ensure adequate sensitivity. The backscatter sounder in conjunction with a background noise measurement can provide an estimate of the expected propagative strength to an area over the HF band.

Propagative Spectral Purity. In surface mode, and to some extent air mode, the spread in Doppler space impacts detection capability. An adjunct probing Doppler backscatter radar is useful in these circumstances (19).

Available Channels. All HF radars should operate on the principle that they do not interfere with other users of the HF spectrum, and to this end, elaborate spectrum surveillance systems are employed. Even then, there is always some residual background noise which can compromise sensitivity, so radar management should choose a channel that is unoccupied by other HF users and has minimum external noise.

Coordinate Registration Issues. The choice of frequency can impact the accuracy of an OTHR. Use

of some frequencies may place the radar in a situation where targets are detected through multiple paths. This may prove difficult to model analytically and, subsequently, to determine the actual target locations.

In some systems, part or all of the radar management process is automated, and the operator function is one of monitoring to ensure that sensible decisions are made. A description of a representative adaptive radar management scheme that addresses issues such as resource allocation, signal diagnosis, signal correction, task assignment, algorithm selection, and performance evaluation can be found in (20).

Detection and Tracking

Once the radar is surveying a region, the output is processed through the various stages, until it is presented to tracking operators for further vetting. This is usually done at some stage after the data has been transformed into scaled signal power as a function of angle of arrival, range, and Doppler. The graphical presentation format of the signal level data is crucial. Figure 7 shows the *azimuth–range–doppler (ARD)* interface used in the Australian Jindalee OTHR system. Usually, some attempt is made to extract peaks automatically from a *whitened* background (one in which the background noise has been constrained to be of uniform energy density in range and Doppler). Some of the functions that the tracking operators can perform at this stage are as follows:

Assistance to the Tracker. Because of the statistical nature of OTHR trackers and the fact that they are based on incomplete models of the targets, they are less reliable than those for conventional radars. For this reason, most systems allow tracking operators to assist the computer-based tracker. Functions that an operator may perform include track initiation, track deletion, and joining tracks.

Assistance to the Coordinate Registration System. Coordinate registration systems are designed to remove system and environmental effects on target detection in an attempt to locate targets accurately on the earth. They vary in sophistication from the manual setting of ionospheric mirror heights, to three-dimensional numerical ray tracing schemes that have system aspects added and use data-driven, real-time, ionospheric models. Coordinate registration is fraught with difficulty because of the uncertainty in many of the component processes. Therefore, it is important that the system handle uncertainty and provide for operator assistance. Some of the activities that a tracking operator might perform are assistance in processing sensor data for the ionospheric model, overriding the predicted effective layer heights and assistance in associating of propagative paths with tracks.

Provision of Feedback to Radar Management on the Performance of the Task. Tracking operators view the signal level data and the track output. They can provide feedback on the quality of the data, including any evidence of interference, to the radar operator. Tracking operators are also in a good position to determine when tracks of interest are nearing task boundaries and require that the radar manager move tasking regions.

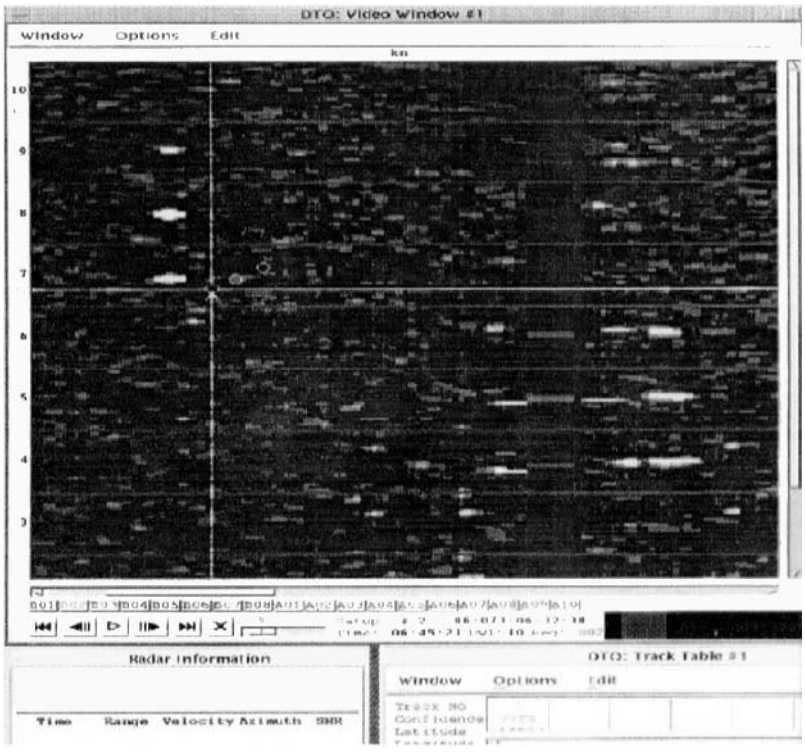


Figure 7. One way of displaying three-dimensional radar data — nesting range within the azimuth along the vertical axis and using the horizontal axis for Doppler. This is the azimuth–range–doppler (ARD) format.

Geophysical Sensing Applications

As noted, OTHRs, designed primarily for defense missions such as detecting ships and aircraft, are nonetheless useful as tools for probing the geophysical environment. The typical spatial resolution of about 10 km means that radar observables are averages on this scale and complement point measurements such as those made by a single wave buoy, for example. In contrast, the achievable Doppler resolution with OTHR is extremely fine, so that very subtle dynamic features of the sea surface or of the midlatitude ionosphere, say, may be revealed. Three examples illustrate the use of OTHR for geophysical observations.

Ionospheric Dynamics

By astute choice of waveform and sampling strategy, it is possible to monitor a region of the ionosphere that spans several million square kilometers at a sampling rate sufficient to observe the evolution of disturbances from atmospheric gravity waves, field-line resonances, day–night terminator passage, equatorial plasma bubbles, and various other phenomena. Figure 8 shows a map of the equivalent vertical velocity of the ionospheric control (reflective) point for a one-hop reflective path; this velocity is inferred from the Doppler shift of ground backscatter by using the effective reflective height model of Breit and Tuve (8). From an analysis of the time evolution of this picture, we can associate the ridge in the lower part of the picture with a propagating solitary wave that endures for more than an hour.

Ocean Wave Height

Of key interest to mariners, offshore oil facilities, oceanographers, and others is the sea roughness, expressed in terms of *root-mean-square (rms)* wave height or alternative measures of surface variance. HF radar observations of the Doppler spectrum of sea clutter are related directly to the directional wave spectrum of sea waves in the gravity wave band, particularly waves of period between about 2 and 20 s. Estimating the directional spectrum yields wave height via simple integration, and this is the preferred method, but other approaches, some based on pattern recognition, perform almost as well at far less computational cost.

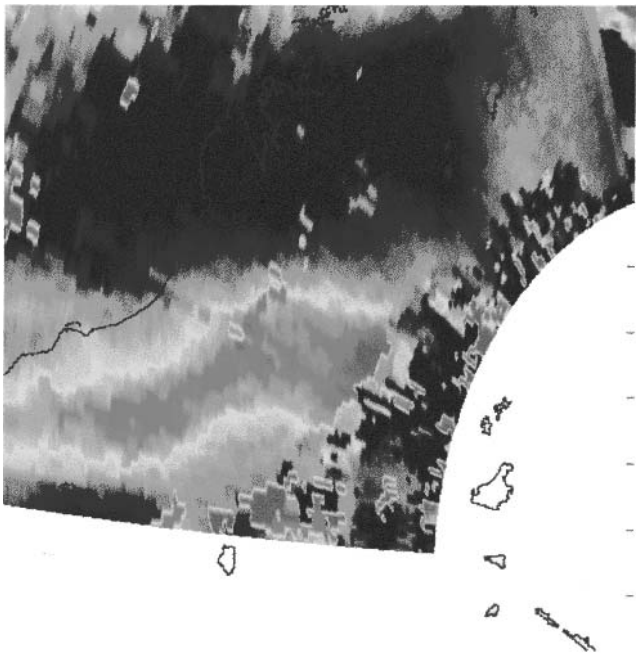


Figure 8. The equivalent vertical velocity of the ionospheric control point for a one-hop reflective path.

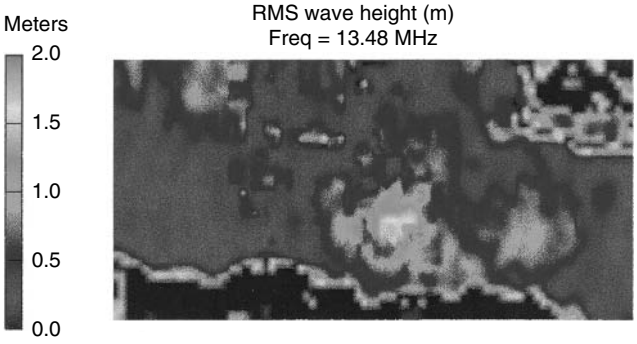


Figure 9. The root-mean-square (rms) wave height measured over a patch of sea off northwestern Australia.

Figure 9 shows the rms wave height for a patch of sea off northern Australia derived by using one such method. The enhanced wave height in the center of the radar footprint coincides with the position of a tropical cyclone.

Surface Wind Fields

Because surface gravity waves on the sea are generated by wind, it is possible to infer the wind field from measurements of the wave distribution. Fairly reliable maps of wind direction (not speed) can be generated from just two point values of the sea clutter Doppler spectrum (21). These are the amplitudes of the two Bragg peaks mentioned earlier that arise from scattering from water waves whose frequencies are typically in the range 0.3–0.5 Hz. Such waves are fairly tightly coupled to the local wind under most conditions, and hence the ratio of wave amplitudes in two opposing directions indicates the generating wind direction, subject to a left–right ambiguity that can be resolved almost everywhere from collateral data or other means. Although far more sophisticated techniques are available, for inferring wind direction, based on the full directional wave spectrum, the present method has the enormous advantage that the spectrum features on which it is based are extremely robust in the presence of signal distortion and contamination.

Figure 10 is a map of wind direction covering a region over the northeastern Indian Ocean. It shows strong curvature, which may be a precursor to cyclone generation.

SYSTEM IMPLEMENTATION

Principal System Architectural Issues

Any sensing system's effectiveness depends critically on the received SNR, and it is a primary OTHR system

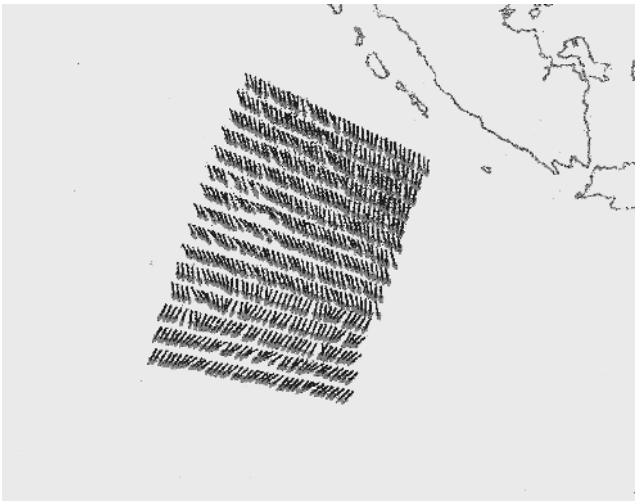


Figure 10. A map of wind direction covering a region over the northeastern Indian Ocean.

design driver. The received SNR of an OTHR depends on a multiplicity of factors. The fact that ionospheric propagation is essential to obtain any signal at all means that the designer has a major parameter, frequency, determined to a large extent by external and time-dependent factors — it is not an independent variable as it is in less medium-dependent realizations of radar. Further, the fact that the operating frequency range is typically 5 : 1 means that the impact of this range on all factors that determine SNR must be recognized.

Another factor peculiar to OTHR, as opposed to conventional radar, is that the external noise environment is usually very high (22). This allows some relaxation of antenna design to effect economies, as long as external noise dominates.

Table 2 shows the dependence of major radar parameters on physical parameters and illustrates the difficult optimization problem posed by OTHR system design. The table assumes a periodic *waveform repetition frequency (WRF)* of f_r Hz that modulates a carrier of frequency f_c Hz where c m/s is the velocity of propagation.

From this list, it follows that the most difficult detection conditions are at low f_c. By making the ratio f_r/f_c constant and the number of samples per dwell fixed, the velocity ambiguity is constant, and the revisit time is approximately constant. The highest f_r that gives the shortest range ambiguity occurs at the maximum frequency where propagative limitations on simultaneous range depth are greatest. Hence, range-ambiguous returns are unlikely. At low frequencies, returns from long ranges are likely, and the WRF is correspondingly low. The aim is to follow this simple relationship of WRF to frequency, but it cannot always be maintained because of the need to avoid extended clutter.

Subdividing of the transmitting and receiving antenna apertures provides added flexibility and robustness. At high frequencies, where full apertures may be unnecessary for sensitivity or resolution, two half apertures operating on different carrier frequencies can enhance the depth of coverage. At low frequencies, approximately 9 dB of extra sensitivity is achieved by using the full aperture.

An approach to antenna geometries can be determined from basic considerations summarized earlier. For a bistatic OTHR, separation of the sites allows basing the design on cost minimization procedures where costs are assigned for transmitting power, transmitting aperture, receiving aperture, and receiver signal processing. The aim in this design approach is to select the worst case conditions and then design to meet a probability of track initiation in a specified time. This allows a trade-off in scanning rate and probability of detection per radar dwell.

Sub System Hardware Design

For an OTHR to attain its fundamental detection capability, careful attention to individual items of equipment is required, taking note of the particular and demanding environment in which it is to operate. Failure to comprehend and manage this environment will result in various forms of signal corruption that limit system performance. Whatever the precise source of the corruption, there are two principal classes of consequences: the radar system noise floor may be artificially raised, or spurious signals may be generated within the radar bandwidth. In either case, the probability of target detection is compromised compared with the system's ultimate potential that is set by external noise levels.

Waveform Generation

The principal design objective for the waveform generation equipment is to maintain spectral purity, so that spurious responses and noise are less than equivalent external noise levels. A database of HF propagative data can be used to establish design objectives in terms of specified radar sensitivity requirements (23). Some quite subtle considerations come into play — an example is the

Table 2. Dependence of OTHR Performance Parameters on Physical Factors

Antenna Gain	Proportional to f_c
Antenna beamwidth	Inversely proportional to f_c
Waveform range ambiguity	$c/(2f_r)$
Waveform velocity ambiguity	$f_r c/(2f_c)$
External noise spectral density	Falls by 30 dB from 6 to 30 MHz.
Target radar cross section (small targets)	Proportional to f_c^4
Depth of coverage on single f_c	Limited by absorption during daytime

recognition that noise is backscattered into the receiving system from the entire radar footprint, not just from the individual range cell under consideration at a given time. To counter this effect requires suppressing the waveform generator noise floor by an additional 20 dB beyond what would be the level if the effect had been overlooked in the design specification.

Power Amplifiers

Logically, a power amplifier can be considered an extension of the waveform generator feeding it, and consequently, the waveform generator specification remains relevant, with one significant exception. Some OTHR architectures call for each of a number, N_T, of power amplifiers to be appropriately phased and fed in parallel, connecting each to its own antenna in an array. Whereas noise originating in the waveform generator will be beam-formed along with the intended signal, noise generated within the power amplifiers is incoherent across the transmitting aperture. Consequently, in terms of the radar system design, the power amplifier noise specification can be relaxed by $10\log(N_T)$ dB with respect to the waveform generator noise specification.

Antennas

Transmitting antenna technology applied to OTHR systems is borrowed from well-established shortwave broadcast application, but has stringent requirements for wide bandwidth and careful control of array mutual coupling effects that can cause operationally unacceptable load interdependencies between power amplifiers. Spatial ambiguities (*grating lobes*) are controlled either by element spacing that does not significantly exceed half a wavelength at the highest frequency of operation or by pseudorandom spacing. The former approach creates excessive mutual coupling at lower frequencies and drives designers to break the band into two or more subbands with separate arrays for each subband. The latter approach offers economies but delivers generally higher and less controllable array side lobes.

Receiving arrays are generally large, so using simple low-cost elements is critical to overall system cost. Control of grating lobes at the highest frequencies determines the maximum interelement spacing and hence the total number of elements required for the desired array aperture. (Alternatively, pseudorandom spacing can be used to break the spatial periodicity which causes grating lobes, but at a cost, as for the transmitting case, of overall side-lobe control.) Although each decibel of mismatch or efficiency loss in the transmitting antenna equates to a decibel loss of radar sensitivity, the receiving antenna operates in such a high external noise field that antenna mismatch and efficiency loss are generally not of prime importance, particularly at the low end of the HF band. This greatly eases the costs of the receiving elements. The criterion is simply that the external noise power available from the antenna should exceed the internal receiver noise by some acceptable quality factor. In fact, excessive coupling of the external environment can compromise system performance through the nonlinear phenomena discussed later.

Wind-driven vibration of the antenna, *Aeolian noise*, is, in principle, a source of contaminating phase modulation of the radar signal, at least one implemented system uses mechanical dampers on its tubular receive array elements to counter any such effects. Measurements have generally shown that wire antennas that have realistic tension do not contribute measurable Aeolian noise.

Receiving System

Design of an OTHR receiving system is a particularly challenging task. Critical parameters and major considerations are summarized in Table 3.

Signal Processing

Modern OTH radars are characterized by a voracious appetite for signal processing capacity: the number of receivers fed from a single receiving antenna array has increased by a factor of about 5 per decade since 1960, and receiver output bandwidth by a similar factor. Accordingly, even the basic processing operations involved in converting from the signal domain (antenna location, time sample) to physical coordinates (range, bearing, velocity) are now demanding tens of gigaflops of processing power. In addition, deeper understanding of the physics involved in propagation and scattering and the development of adaptive filtering algorithms have led to a dramatic increase in the sophistication of signal decomposition and interpretation techniques (26).

Notwithstanding the evolution of signal processing objectives and algorithms, the basic architecture of almost all current processing systems remains the classical pipeline structure. In general, it is computationally more effective to perform waveform compression (range processing) first, before transforming the signal from receiver space to beam space. Then, this projection of the incoming signal onto a spatial grid of range–azimuth cells can be repeated for consecutive occurrences of the elemental waveform until a time series of samples is accumulated for each grid cell.

For almost all practical applications, it is not the time-domain history but the corresponding frequency spectrum associated with the Doppler signature of the scatterers that is required, so that the third signal transformation is the spectrum analysis of the accumulated time series. The suite of algorithms employed for these signal transformations has included, almost without exception, the *fast Fourier transform (FFT)*, especially for Doppler, often for conventional beamforming, and sometimes for range processing, depending on the choice of waveform. In recent times, greater processing capacity has enabled implementing adaptive beam formers in real time, as well as the limited use of superresolution methods in all three domains (27).

Increasingly, the refinement of signal processing techniques is based on more elaborate models of the propagative environment, of target scattering signatures, and of the characteristics of the external noise field. Adoption of these model-based or parametric signal

Table 3. Critical Receiving System Parameters

Noise figure	Must be maintained sufficiently low so that the system sensitivity and hence the radar detection capability are determined by external noise sources of atmospheric and galactic origin.
Out-of-band *intermodulation distortion* (IMD)	Spurious signals can appear at the output of the receiving system as a consequence of the nonlinear interactions of large signals external to the channel under consideration. This phenomenon must be recognized in a balanced approach to receiving system design.
In-band IMD	Fundamentally the same mechanism applies as for out-of-band IMD, but in this case the signal under consideration is itself driving receiver nonlinearity.
Cross-modulation	This too is fundamentally a nonlinear mechanism, but instead of the nonlinearity creating spurious signals close to the frequency of the wanted radar signal, modulation of external unwanted signals is imposed on the wanted radar signal. This corrupts the frequency domain or Doppler space in which targets may have been detected (24).
Reciprocal mixing	In a superheterodyne receiver, the radio frequency is converted to an intermediate frequency in various mixers which, ideally, function purely as frequency translators. Any phase noise or spurious signals on the local oscillator signal can mix with large unwanted signals that appear at the radio frequency input along with the radar signal and appear at the same intermediate frequency as the wanted radar signal from which it cannot be separated (25).
Selectivity	Within the congested HF bands, it is imperative that the filtering employed within the receiver isolate the wanted radar signal from contamination by much larger unwanted signals.
Local sources of interference	The receiver site should be free of sources of interference that could raise the level of noise or interference in which the radar is required to operate. The need to collocate computing and communications equipment with the receiving system poses significant electromagnetic compatibility engineering challenges.

processing methods must be accompanied by careful validation of the model structure and the use of statistical tests to ensure that the model chosen is appropriate to the prevailing situation. For example, the application of some spatial and temporal desmearing techniques, designed to compensate for ionospheric fluctuations, would actually degrade the data if the observed spectral broadening were caused by a different mechanism such as multimode propagation.

Automatic Detection and Tracking

The automatic tracking process in an OTHR uses data supplied from a peak detection process. Peak detection consists of peak selection followed by interpolation (28). In peak selection, the local peaks in SNR are found in amplitude versus range and Doppler. A signal can be declared a local peak only if its SNR is greater than all the surrounding ARD values. By fitting a parabola to each selected peak in ARD space, the location of the center of the signal that produced the peak is interpolated between the cell positions. To limit the number of interpolated peaks sent to the tracking process, a threshold in SNR is applied which is a compromise between enhanced detection of small targets and spurious detections from noise, clutter, and interference. The threshold value is determined by the automatic tracking system's ability to deal with false measurements and prevent the initiation of false tracks.

The peak detections supplied to the automatic tracking process contain ambiguous target identification information, the SNR gives some indication of which peaks are likely to be from targets. However, clutter interference from meteors and the residues from interference suppression can give target-like returns. Further, ionospheric multimode propagation gives multiple peak detections from a single target. Normally, they are resolved in range by correctly selecting of the radar bandwidth.

The automatic tracking system must cope with these uncertainties in the sources of peak detections and allow for closely spaced target peak detections created by multimode propagation. In addition, the tracking system must resolve any velocity ambiguity from the measured Doppler. The most common approach is to track the separate multimode returns in slant range coordinates, group the multimode tracks to form a fused track, and then convert the fused track to ground coordinates. An alternative approach is to have a model of the multimode propagation in the tracking filter, which then tracks in ground coordinates (29). This approach has yet to be demonstrated operationally under conditions that include errors in the ionospheric model, multiple targets, and velocity ambiguity.

A demonstrably useful approach is based on the *probabilistic data association filter (PDAF)* (30). A fixed number of nearest measurements is selected by each filter, and for each measurement, the probability is found that it is either from the target or from clutter and noise. These probabilities control the correction applied to the peaks that update the filter. Possible PDAF extensions include automatic track initiation with multiple velocity ambiguities and tracking in a nonuniform cluttered environment.

LIMITATIONS AND AVENUES FOR FUTURE DEVELOPMENT OF OTHR TECHNOLOGY

It is evident that the extensive coverage and high coverage rate of OTHR is achieved at the cost of employing a propagative channel that may distort and contaminate the signals severely. This has motivated numerous

Table 4. Projected Future Enhancements for OTHR Systems

Enhancement Sought	Strategy for Achievement	Indicated Developments
Improved detection range and greater ability to detect small, fast (airborne) targets	Noise and elevated clutter rejection	Efficient antenna elements (D), polarization filtering (D), spatiotemporal filtering (D), two-dimensional arrays (D), optimum siting (U), lower frequency (U — needs very tall antennas).
Greater ability to detect small, slow (ship) targets	Minimization of sea clutter	Special wave forms (A), multistatic configurations (A), process model optimization (U), optimum siting (U).
Improved spatial precision and resolution	Overcome array diffraction limit	Model-based spectrum analysis (D), extended aperture (U), interferometry (D), stereoscopic observations (A), multistatic configurations to increase time-delay resolution (U), wide bandwidth waveforms (D), superresolution algorithms (D).
Reduced siting costs and expanded options for deployment	Single-site operation	T/R switched pulse waveform (A), highly linear receiver (A), other special waveforms (U), passive radar (A).
Remote operation to enable tasking authority to control radar in real time and view product	High bandwidth communications links and appropriate protocols	Commercial communications technology application (D)
Characterization of scatterer	Exploit static and dynamic signatures	Carrier frequency domain (D), kinematic domain (U), Doppler domain (U), phase signature domain (D).

investigations into the factors that now limit performance and into strategies for overcoming them.

Table 4 lists desired OTHR features or enhancements that are judged of greatest practical importance and associates these with prospective technical solutions. The feasibility of implementation is indicated by D (*demonstrated*), A (deemed *achievable* but not demonstrated) or U (deemed *understood* but subject to confirmation or has major implementation difficulties).

BIBLIOGRAPHY

1. M. I. Skolnik, *Radar Handbook*, 2nd ed., McGraw-Hill, NY, 1989.
2. J. M. Headrick and M. I. Skolnik, *Proc. IEEE* **62**, 664–673 (1974).
3. T. M. Georges, *IEEE Trans. Antennas Propagation* **Ap-28**, 751–761 (1980).
4. Special Issue on High-Frequency Radar for Ocean and Ice Mapping and Ship Location, *IEEE J. Oceanic Eng.* **OE-11**(2) (1986).
5. K. G. Budden, *The Propagation of Radio Waves: The Theory of Radio Waves of Low Power in the Ionosphere and Magnetosphere*, Cambridge University Press, 1985.
6. L. F. McNamara, *The Ionosphere: Communications, Surveillance and Direction Finding*, Orbit, Krieger, Melbourne, Fl, 1991.
7. G. F. Earl and B. D. Ward, *Radio Sci.* **22**(2), 275–291 (1987).
8. K. Davies, *Ionospheric Radio*, Peter Perigrinus, London, 1990.
9. R. I. Barnes et al., *Radio Sci.* **33**(4), 1,215–1,226 (1998).
10. K. Hocke and K. Schlegel, *A Review of Atmospheric Gravity Waves and Travelling Ionospheric Disturbances:* 1982–1995, 1996.
11. S. J. Anderson and M. L. Lees, *Radio Sci.* **23**(3), 265–272 (1988).
12. R. F. Harrington, *Time Harmonic Electromagnetic Fields*, McGraw-Hill, NY, 1961, pp. 292–296.
13. R. F. Harrington, *Field Computation by Moment Methods*, Macmillan, NY, 1968.
14. E. K. Miller, L. Medgyesi-Mitschang, and E. H. Newman, eds., *Computational Electromagnetics; Frequency-Domain Method of Moments*, IEEE Press, NY, 1992.
15. D. E. Barrick, in V. E. Derr, ed., *Remote Sensing of the Troposphere*, U.S. Government Printing Office, Washington, DC, 1972, Chapter 12.
16. B. J. Lipa, *Radio Sci.* **12**, 425–434 (1977).
17. L. R. Wyatt, *Int. J. Remote Sensing* **11**, 1,481–1,494 (1990).
18. S. J. Anderson, Proce. Radarcon-90, Adelaide, Australia, April 1990, pp. 315–323.
19. R. I. Barnes, *IEE Proc. Radar Sonar Navigation* **143**(1), 53–63 (1996).
20. S. J. Anderson, *Proc. IEE* Part F, **139**(2), 182–192 (1992).
21. A. E. Long and D. B. Trizna, *IEEE Trans. Antenna Propagation* **AP-21**, 680–685 (1973).
22. *Characteristics and Applications of Atmospheric Radio Noise Data*, CCIR Report 322-3, International Telecommunications Union, Geneva, Switzerland, 1988.
23. G. F. Earl, *Radio Sci.* **33**(4), 1,069–1,076 (1998).
24. G. F. Earl, *IEEE Trans. Instrum. Meas.* **IM-36**(3) (1987).
25. G. F. Earl, *Consideration of Reciprocal Mixing in HF OTH Radar Design*, HF Radio Systems and Techniques, IEE Conference Publication, 1997, 411–.
26. S. J. Anderson and Y. I. Abramovich, *Radio Sci.* **33**(4), 1,055–1,067 (1998).
27. S. J. Anderson, *Proc. Pacific Ocean Remote Sensing Conf., PORSEC'94, Melbourne, Australia, March 1994.*
28. S. J. Davey, S. B. Colegrove, and D. Mudge, *Advanced Jindalee Tracker: Enhanced Peak Detector, Defence Science and Technology Organization Technical Report No. DSTO-TR-0765*, 1999.
29. G. W. Pulford and R. J. Evans, *Automatica* **32**(9), 1,311–1,316 (1996).
30. Y. Bar-Shalom and E. Tse, *Automatica* **11**, 451–460 (1975).

P

PARTICLE DETECTOR TECHNOLOGY FOR IMAGING

GEOFFREY N. TAYLOR
The University of Melbourne
Parkville, Victoria
Australia

INTRODUCTION: PARTICLE IMAGING

In this article, "particles" refer to charged particles that here enough energy to pass some significant distance through matter. The interaction of particles with matter forms the basis on which particle detectors are designed and operated. The principal process for detection is ionization of matter by the electromagnetic interaction of the moving particle. Ionization produces free charge that can be collected and detected. Electromagnetic radiation from the charged particle moving through matter may also be used to create a detectable signal. Neutral particles can be observed following their interaction with matter, producing one or several charged particles, that are subsequently detected.

The development of particle detectors has occurred within the fields of nuclear and high-energy particle physics. In this regime, the interest lies in reconstructing particles emanating from a particle interaction or from the decay of unstable particles or nuclei. The trajectories of the detected particles are recorded to reconstruct the kinematics (momentum and energy of the particles) of the reactions being observed. The particle trajectories are sampled as the particles pass through the detector elements. The sampled track positions are used to reconstruct the full trajectory.

In addition to the kinematics of such reactions, spatial information is also important, especially when determining the distance that an unstable particle travels before decaying. The decay time distributions of unstable particles follow the exponential statistical time distributions of radioactive decays whose half-lives or lifetimes are characteristic of the particle or the nuclear state being studied. For sufficiently long lifetimes, the combination of track length before decay and the reconstructed kinematics of the decay products are used to determine the decay time of individual decays. The ensemble of data is then used to extract the characteristic lifetime of the particle from the distribution of the individual decays.

The list of particles that can be detected is generally restricted to those whose stability is sufficient for travel some minimum distance in matter. Current limitations of the spatial resolution of electronic detectors limit observable track lengths to a few hundred microns or more. (Stacks of particle-sensitive emulsion, similar to photographic film, have successfully measured decay lengths to a few microns. The rate limitation and lack of timing information, however, limit the applicability of this technique to imaging. The combination of lifetime (τ), extended when viewed from the laboratory by the relativistic factor, $\gamma = E/mc^2$ (where E is the particle energy and m its mass) and normalized particle velocity, $\beta = v/c$ (where v is the particle velocity and c is the speed of light), yields an average distance traveled before decay of $z = \gamma\beta c\tau$.

The most common particles that can be tracked and imaged include electrons, protons, some nuclei, muons, pions, kaons, and neutrinos. High-energy photons (X rays and gamma rays) also have particlelike properties and provide great utility in imaging applications. The source of the particles can be radioactive tracers distributed within the object being imaged or an external source such as a radioactive source, a reactor, an accelerator, or cosmic rays.

For further reading in this area, see (1–7).

INTERACTIONS OF PARTICLES WITH MATTER

Ionization

The passage of moving charged particles through matter results in a coulombic interaction with the atomic electrons in the matter. Energy is transferred from the moving particle to the material. Sufficient energy transfer to the atomic electrons causes ionization of the atom, resulting in an ejected electron and an ionized atom (ion). The electrically charged ionization electrons and ions form the basis of the electrical signal that signifies the passage of the charged particle. In crystalline solids, ionization results in a free electron and "hole," both of which can contribute to electrical signals.

The average rate of energy loss of a high-energy particle, charge z, velocity β, whose mass is much greater than that of the electron is given by the Bethe–Bloch equation:

$$-\left\langle \frac{dE}{dx} \right\rangle = Kz^2 \frac{Z}{A} \frac{1}{\beta^2} \left[\frac{1}{2} \ln \left(\frac{2m_e c^2 \beta^2 \gamma^2 T_{\max}}{I^2} \right) - \beta^2 - \frac{\delta}{2} \right], \quad (1)$$

where Z and A are the atomic number and atomic mass, respectively, and I is the average ionization energy of the material through which the particle is traversing. $K = 4\pi N_A r_e^2 m_e c^2$, N_A is Avogadro's number, $r_e = e^2/4\pi\varepsilon_0 m_e c^2 \approx 2.82\,fm$ is the classical electron radius, and m_e is the electron mass, $m_e c^2 = 0.511 MeV$. $T_{\max}$ is the maximum kinetic energy that the particle can transfer to a stationary electron:

$$T_{\max} = \frac{2m_e c^2 \beta^2 \gamma^2}{1 + 2\gamma \frac{m_e}{M} + \left(\frac{m_e}{M}\right)^2} = \frac{2m_e}{M^2 + m_e^2 + 2m_e E/c^2}, \quad (2)$$

where $\gamma = E/m_0 c^2$ and E, p, v, and M are the total energy, momentum, velocity, and mass of the moving particle. The

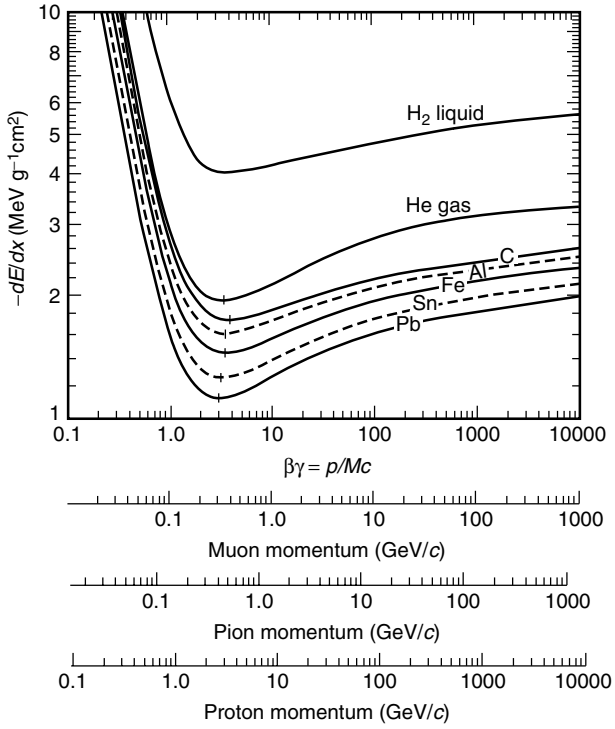


Figure 1. Mean energy loss rate in various materials (7).

distance dx is given in units of areal density, mass per unit area, g cm^{-2}. (Dividing by the material density gives the distance in centimeters.)

Figure 1 shows the mean energy loss rate in various materials. Figure 2 shows the measured ionization energy loss for electrons, pions, kaons, protons, deuterons, and tritons, in the OPAL jet chamber (8), displaying the expected velocity dependence. The different particle masses at given momenta provide different velocities.

At low velocities, the $1/\beta^2$ dependence dominates. Charged particles lose energy at higher rates per distance traveled, as they slow to a stop. This effect causes a (heavy) charged particle to come to a stop rapidly near the end of its trajectory.

The energy loss rate reaches a minimum at intermediate values of $\gamma\beta$ (*minimum ionizing particles* or *MIPs*) before the logarithmic increase becomes important at higher values. This "relativistic rise" due to length contraction in the direction of the particle and elongation of the field transverse to the particle direction is generally small. However, in low density gases, the relativistic rise can be significant before reaching a plateau. The plateau or "density effect" is the result of shielding by polarization of atomic electrons, and hence sets in earlier in higher density material (9). For a detailed description of the determination of ionization energy loss in various materials, see (10).

The energy loss of a charged particle passing through some small amount of material is stochastic. The ionization energy deposited in a thin layer of material follows the "Landau distribution" (11), as shown in Fig. 3. The average value and the shape of the distribution describe the energy deposited in charged particle detectors. In most detectors, the detected signal will follow a similar distribution,

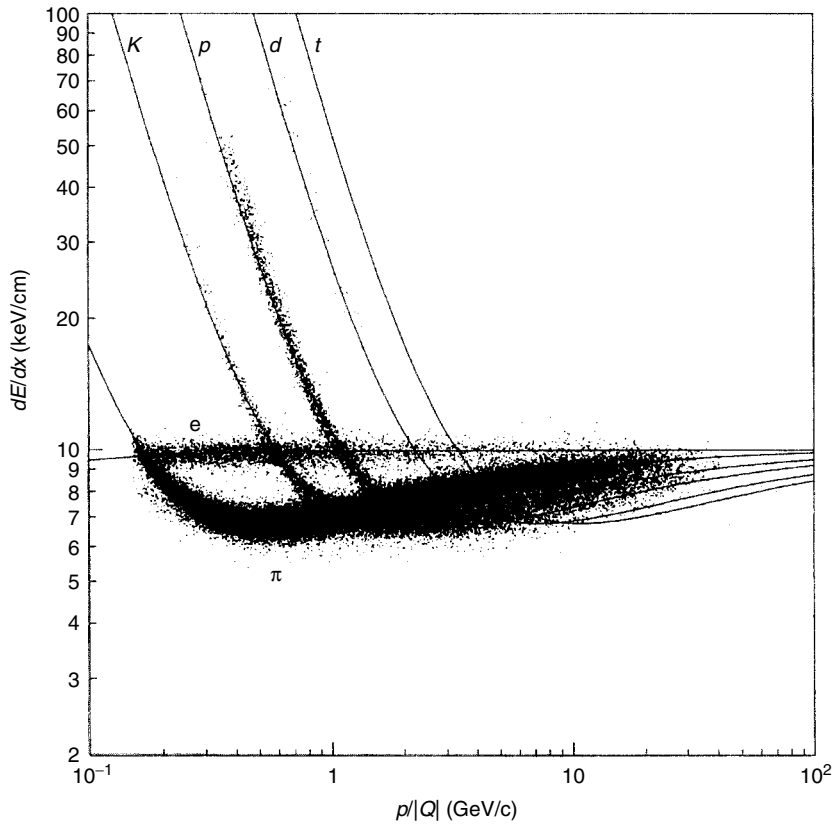


Figure 2. The measured ionization energy loss for electrons, pions, kaons, protons, deuterons, and tritons in the OPAL jet chamber (8), displaying the expected velocity dependence. The different particle masses at given momenta provide different velocities.

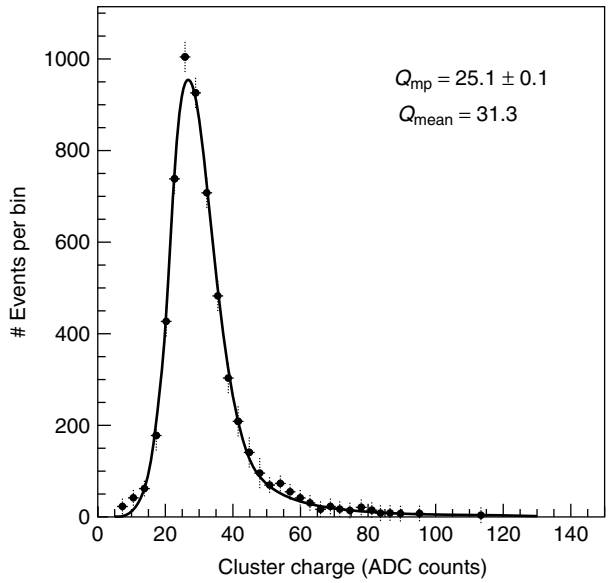


Figure 3. The Landau distribution shows the energy loss distribution measured in a 300-μm thick silicon detector (12).

modified by any nonlinearity, noise, or dynamic range limitations. It is usual to define

$$f(x, \Delta) = \frac{1}{\xi}\omega(\lambda) \tag{3}$$

as the probability that the particle will lose energy Δ after traversing material of thickness x. λ is the difference between Δ and the most probable energy loss Δ^{mp},

$$\lambda = \frac{\Delta - \Delta^{\text{mp}}}{\xi} \tag{4}$$

The Bethe–Bloch equation gives the normalization

$$\xi = \frac{0.154}{\beta^2}\frac{Z}{A}x \times 10^3 \quad (keV) \tag{5}$$

where x is the areal density measured in g cm^{-2}.

The Landau distribution is then described by

$$\omega(\lambda) = \frac{1}{\pi}\int_0^\infty \left(e^{-u\ln u - \lambda u}\sin \pi u\right) du. \tag{6}$$

Figure 3 shows the energy loss distribution measured in a 300-μm thick silicon detector (12).

The Bethe–Bloch equation is not appropriate for electrons. In that case, the equal mass of a moving particle and atomic electrons significantly changes the kinematics. Radiation dominates over ionization for electrons (and positrons) at energies as small as some tens of MeV. The contributions to energy loss for electrons and positrons are shown in Fig. 4. At energies above several hundred GeV, radiation also becomes the dominant energy loss mechanism for muons.

The high energy tail in the Landau distribution is due to "knock-on" electrons or "delta rays" from rare, high-energy transfers to individual atomic electrons, resulting in their rapid ejection. If they have enough energy, these particles also contribute to ionization and thus to additional energy deposition in the material. If the detecting material is thin, however, such as in silicon detectors, the high-energy delta rays escape and do not contribute to the energy detected. In this case, the signal energy detected is less than the energy loss.

Interactions of photons with matter are discussed elsewhere. However, at high energies, electrons and photon interactions are both characterized by material thickness in units of radiation length X_0. For high-energy particles, this is the average distance by which the electron energy has been reduced to $1/e$ of its initial value. Tsai (13) calculated the radiation for a range of materials. A useful fit to these calculations is reported in PDG from O.I.Dahl as

$$X_0 = \frac{716.4 \text{ g cm}^{-2} A}{Z(Z+1)\ln(287/\sqrt{Z})}. \tag{7}$$

The passage of charged particles through matter will also result in coulombic interactions with atomic nuclei. The

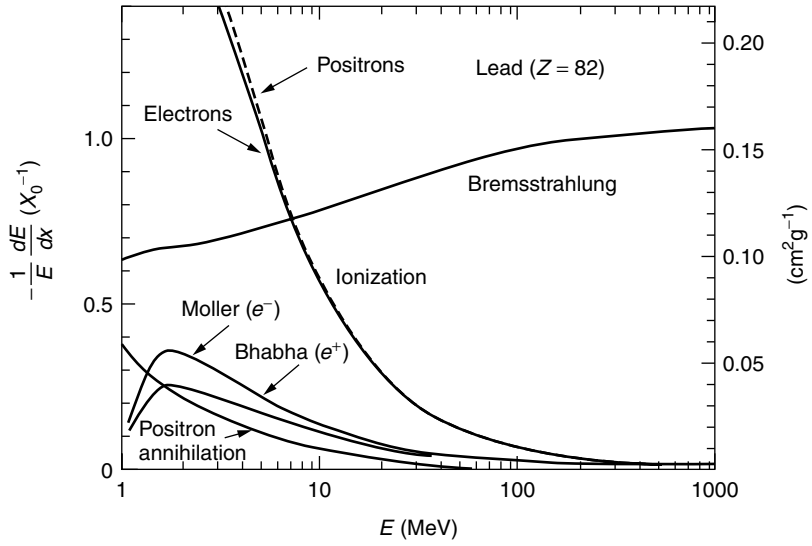


Figure 4. The contributions to energy loss for electrons and positrons (7).

volume occupied by the nuclei is sufficiently restricted that hard scattering of nuclei occurs much less frequently than the interactions between a moving particle and atomic electrons. Furthermore, as distinct from ionization where a heavy moving charged particle is considerably more massive than electrons, the opposite is the case in nuclear coulombic scattering. The large number of interactions in ionization results in significant energy loss but in little change in direction of the heavier moving charged particle — energy loss without significant scattering results. In nuclear coulombic scattering, the massive nuclei behave as scattering centers, but the scattering is typically elastic — scattering without significant energy loss results.

The coulombic force between a moving particle of charge z and momentum p (and velocity $\beta = v/c$, where c is the velocity of light) and a matter nucleus of charge Z results in Rutherford scattering that depends on the impact parameter — the perpendicular distance between the particle trajectory and the matter nucleus. Taking account of the statistical distribution of the possible impact parameters, the "differential cross section" for scattering an angle Θ from the original trajectory is given by the Rutherford scattering formula:

$$\frac{d\sigma}{d\Omega} = 4zZr_e^2\left(\frac{m_e c}{\beta p}\right)^2 \frac{1}{\sin^4(\Theta/2)}, \tag{8}$$

where r_e and m_e are the classical electron radius and electron mass. $d\sigma/d\Omega$ is the differential cross section for scattering into a fraction $d\Omega$ of the unit sphere, centered on the scattering nucleus. The Rutherford formula assumes cylindrical symmetry around the initial particle trajectory. Thus, the scattering depends only on the angle Θ from the initial direction, not on the orientation of the scattered trajectory around the initial direction.

The coulombic scattering cross section is strongly peaked at small scattering angles. Thus, single scatters generally contribute little to a change in the direction of a moving particle. However, passing through a macroscopic amount of matter, a significant directional change can result from the statistical addition of a large number of small scatters. Multiple coulombic scattering (MCS) must be considered when evaluating the trajectories through matter that lead to a loss of information about the original particle trajectory.

Assuming that the scattering angle results from many small angle scatters, the overall scattering angle is, to a good approximation, Gaussian distributed, and the width (for scattering in a plane) is given by

$$\theta_0 \approx \frac{13.6MeV}{\beta cp} z\sqrt{x/X_0}. \tag{9}$$

The characteristic MCS scattering angle is proportional to the square root of the thickness of the material traversed, measured in radiative lengths. It should be pointed out that this Gaussian approximation underestimates the probability of large single scatters. The tail in the scattering distribution is thus somewhat underestimated.

The interactions of photons with matter, discussed elsewhere, are related to processes that affect electrons as they pass through matter.

Magnetic Fields and Particle Trajectories

The classical Lorentz electromagnetic force, which is valid relativistically, results in helical charged particle trajectories in a (uniform) magnetic field. The component perpendicular to the magnetic field follows a circular orbit. Combining this with the longitudinal momentum component results in a helical orbit. The orbit radius is related to the particle momentum and magnetic field by

$$p\cos\lambda = 0.3BRz, \tag{10}$$

where p is the momentum in GeV/c, λ is the pitch of the helical trajectory, B is the magnetic field in tesla, R is the radius of curvature in meters, and z is the particle charge.

From knowledge of the magnetic field, measurement of the perpendicular curvature yields the particle momentum. Energy loss as the particle passes through matter results in a reducing radius. Scattering produces deviations from a helical path and thus contributes to a loss in momentum resolution. These effects must be considered in determining particle momenta.

The positions of particle tracks as they pass through discrete detector elements are sampled to give information from which the full particle trajectory can be reconstructed. The signals must be corrected for any detector biases or misalignments before the data are processed to reconstruct the tracks. The fitting procedure depends on the geometry of the tracking detectors and the presence or otherwise of a magnetic field.

Fitting the track curvature through a magnetic field yields a measure of the momentum. Errors in the measured track positions result in errors in the momentum determined. Assuming a uniform medium and a large number ($\geq$10) N of measured points, equally spaced along a track length L, each of equal measurement error ε, Karimäki (14) determined that the error in the measured track curvature, $k = 1/R$, from detector resolution is

$$\delta k_{\text{res.}} \cong \frac{\varepsilon}{L^2}\sqrt{\frac{720}{N+4}}. \tag{11}$$

The complete covariance matrix is presented in (14) for nonuniform measurement spacing.

An additional error is introduced by multiple coulombic scattering, given by the Particle Data Group as (7)

$$\delta k_{\text{mcs}} \approx \frac{z0.016(\text{GeV/c})}{Lp\beta\cos^2\lambda}\sqrt{\frac{L}{X_0}}, \tag{12}$$

where X_0 is the radiative length of the material traversed.

Assuming that the curvature measurement error and the error due to multiple scattering can be added quadratically, the total curvature error becomes

$$\delta k^2 = \delta k_{\text{res.}}^2 + \delta k_{\text{mcs}}^2. \tag{13}$$

To optimize momentum reconstruction, equal emphasis must be applied to control each term.

A program widely used for estimating expected momentum errors for cylindrically symmetrical geometry is TRACKERR (15).

Fitting tracks to a point of origin (vertex fitting) can provide an additional constraint on the curvature and hence improve the determination of the momentum ("vertex constrained fit"). The fitted tracks can be used to reconstruct the interaction that produced the tracks; to search for separated vertices, indicative of decays or secondary interactions; and to determine the spatial origin of the tracks. For more details, see (16,17).

In high-energy particle physics, measurements of the particle energies can also be carried out by completely absorbing the particles in calorimeters. Electromagnetic and nuclear ("hadronic") cascades result in multiplication of particle tracks. The number of tracks produced is proportional to the energy of the initiating particle. To estimate the number of tracks, active elements (such as scintillator or wire chambers) at discrete points within the cascade region measure the total ionization and its spatial distribution. Typically, a sandwich of many layers of radiator material followed by active volumes is used to contain the complete shower in a sampling calorimeter. A material of high atomic number such as lead is chosen for electromagnetic calorimeters, and iron or a similar material for hadronic calorimeters (18,19). Alternatively, complete absorption calorimeters can be used. Electromagnetic scintillating crystal calorimeters whose scintillation signal is proportional to the energy deposited, or liquid argon or krypton calorimeters that measure total ionization energy can be used to absorb electromagnetic cascades completely.

Particle identification is often important in high-energy physics and in nuclear physics applications. Separation of the various commonly produced particles is possible via their mass differences. For fixed momentum, different masses result in different velocities. Thus, velocity-dependent measurements can thus be applied to distinguish the particle masses. Ionization, particle flight times over fixed distances, and Cerenkov radiation are velocity-dependent effects that can be used for this task.

Scintillation

Ionization energy loss in scintillating materials results in emission of visible light and UV photons in deexcitation of the excited molecules. Collection of this light constitutes particle detection. An appropriate array of scintillators can make particle imaging practical. In particular, the use of scintillating fiber arrays has become a successful high precision tracking technique.

Liquid and solid organic scintillators that have carefully chosen dopants provide fast, reliable, scintillation detectors. Coupling of light guides to plastic scintillators to guide the scintillation light to the photocathode of a photomultiplier is a standard technique for reading out particle-induced signals. Organic scintillators can have very fast signal rise times that result in precise time resolution (typically 150 ps) and short recovery times (10 ns) that are excellent for high rate applications. The positional resolution is determined by the geometry of the scintillator array.

Organic plastic scintillators consist of a base material, usually polystyrene, polyvinyltoluene, or acrylic, and one or more dopants consisting of organic fluors. The base molecules are excited by absorbing approximately 5 eV of energy from ionizing particles. The dopants are chosen to absorb at a wavelength very close to the fundamental excitation of the base molecules (within approximately 1 nm). Energy is transferred quickly from the base molecules to the dopants at high quantum efficiency approaching 100% (19–21). Commonly used combinations result in blue–green or UV–blue emission. A typical 1% doping level gives adequate light emission but also results in significant self-absorption of light that travels through more than about 1 m of the material. Self-absorption can be reduced by using low concentrations (100 ppm) of additional wavelength-shifting dopants.

Cerenkov Radiation

Cerenkov radiation is emitted by charged particles passing through material at a velocity greater than the (phase) velocity of light in the material. Three properties of Cerenkov radiation are important in particle detectors (23): the angle at which the Cerenkov radiation is emitted, the velocity threshold that must be exceeded for radiation to occur, and the number of photons emitted.

The first of these properties is the most important for particle tracking. Cerenkov radiation is emitted in a cone of half-angle θ_c, given by the expression

$$\theta_C = \cos^{-1}(1/n\beta), \tag{14}$$

where n is the index of refraction in the material, and $\beta = v/c$ is the particle velocity.

The cone half-angle, velocity threshold for radiation, and the velocity dependence of the number of photons radiated are all important for velocity determination. Measurements of these quantities can provide particle identification via sensitivity of the velocity to the particle mass for a given momentum (7,23).

Transition radiation, a related process, occurs when a high-velocity particle crosses a material boundary where an abrupt difference exists in electrical permittivity (24). The number of photons emitted is proportional to $(\ln\gamma)^2$, where $\gamma = E/mc^2$. The photon yield is enhanced by creating multiple transitions by using thin foils or expanded polystyrene. The radiation energy is also strongly γ-dependent. Typically, a signal is generated in the soft X-ray region. Transition radiation is particularly interesting for distinguishing electrons from heavier charged particles. For the same momentum, the γ for electrons is significantly higher than for the more massive particles and enhances the transition radiation produced.

PARTICLE DETECTION TECHNIQUES

Scintillation Tracking Devices

Light plastic and liquid scintillators are appropriate for tracking applications to minimize scattering. In X-ray and gamma-ray imaging, heavy inorganic scintillators are

used to absorb photons highly efficiently. Imaging of high-energy photons using a segmented scintillator is discussed elsewhere in this volume.

The positional resolution of scintillator tracking detectors is determined by the geometry of the scintillator. These detectors can be used as the principal tracking device, where fine scintillating fibers have been successfully implemented. The fast response of scintillators makes them attractive as the "trigger" device, where coarser segmentation is sufficient. A signal in trigger detectors can be used to initiate a readout of other, higher precision tracking devices. The timing information provided by the scintillator can provide further selection criteria and can be used to initiate a measurement in devices such as drift chambers (see later) that use timing information to determine track positions.

Techniques for converting the scintillating light signal into an electrical signal must take account of the small light production and collection efficiency and of the quantum efficiency (efficiency in detecting single photons) of the light detection itself (20–70%). For tracking, the light detection device must retain the positional information or the identity of the fiber from which the light was obtained. Position-sensitive photomultipliers are available commercially for this application (25). Hybrid photomultipliers have also been developed for imaging applications. Alternatively an image intensifier can be used to maintain positional information while increasing the signal amplitude and maintaining signal timing. The amplified light output of the image intensifier can be imaged by a CCD device or an array of such devices. The image intensifier can be triggered or gated by an external signal to record only data coincident with a given event.

Bundles of scintillating fibers, plastic and glass, and liquid filled capillaries, coupled to light detection arrays via inactive optical fibers have been successfully used in particle tracking (see 26 and references therein, 27, 28). Figure 5 shows a schematic of such a combination.

Wire Chambers. The signal in a gas particle detector commences with primary ionization of the gas molecules by the track being measured. Primary ionization is characteristic of the ionizing particle track and of the gas. Some of the primary ionization electrons have enough energy to produce further, secondary ionization. The primary charge will drift under the influence of an applied electric field and can be collected for measurement.

Typical gases used in such detectors are the inert gases, carbon dioxide, methane, and ethane. The energy W required to produce an electron–ion pair in such gases is typically 20–40 eV For minimum ionizing particles, the ionization energy loss per track length varies with the density and atomic number of the gas molecules. The resulting ionization varies from approximately 10 electron–ion pairs per centimeter for the lowest atomic weight gases to several hundred pairs for the highest. The relatively low ionization rates lead to a statistical distribution in the primary ionization that is described by Poisson's law. The ionization per track length follows a Landau distribution. The statistical nature of the ionization results in varying efficiency for detection and needs to be considered in thin gas chambers.

The ionization cloud is *drifted* through the gas under the influence of an applied electric field E to a detecting anode. The speed at which the charge cloud drifts under the influence of the field is the drift velocity. The ratio of drift velocity to electric field is the mobility $\mu = v_{\text{drift}}/E$. For ions, the drift velocity is proportional to the field, and the mobility is approximately constant. For electrons, the strong energy and gas dependence of

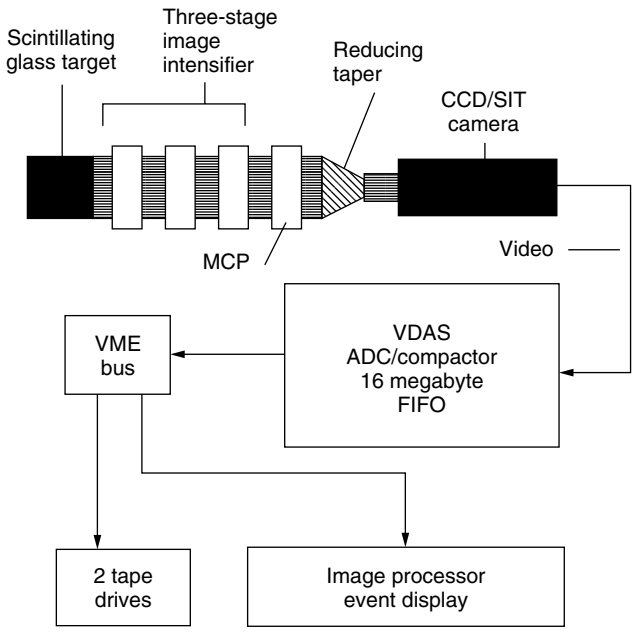


Figure 5. Schematic of a scintillating fiber detector (25).

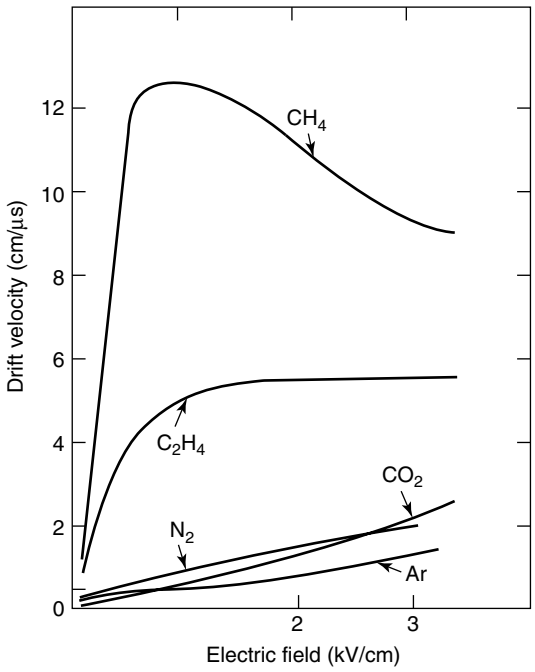


Figure 6. Electron drift velocity in various gases at STP (37).

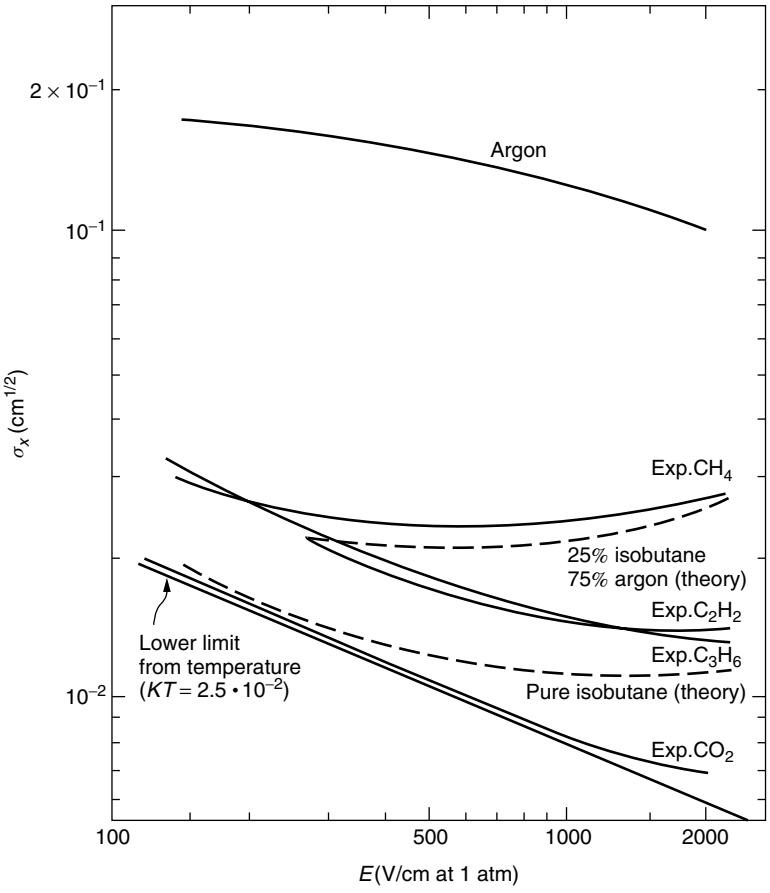


Figure 7. Electron diffusion per centimeter of drift for various gases at STP (28).

the electron–molecule collisional cross section results in a complicated electric field dependence of the drift velocity [see, for example, Biagi (29) and Fig. 6]. Ion mobilities are in the range of 1–10 $cm^2V^{-1}s^{-1}$ for typical chamber gases at STP.

An ionization charge will diffuse from the production region where its density is largest. When drifting the charge to collection anodes, diffusion continues to expand the region occupied by the charge, affecting the precision with which the cloud can be localized. Drifting the ionization charge under the influence of an applied electric field increases the energy of the charge. For electrons in particular, the strong energy dependence of the electron–molecule collisional cross section can result in field-dependent diffusion. An Enhanced collisional cross section results in broadening of the electron cloud (see Fig. 7; 31).

The ability of the detector to determine the position of the original particle track by localizing the primary ionization is limited by the diffusion of the charge cloud before signal collection.

A sufficiently large applied electric field can increase the electron energy to such an extent that collisions with molecules result in further ionization. The critical field in the range 5–20 $kVcm^{-1}$ is required for this ionization multiplication. For a given gas in an electric field above the critical value, there is a characteristic length that electrons will travel, on average, to create an additional electron–ion pair. The inverse of this length is the Townsend coefficient

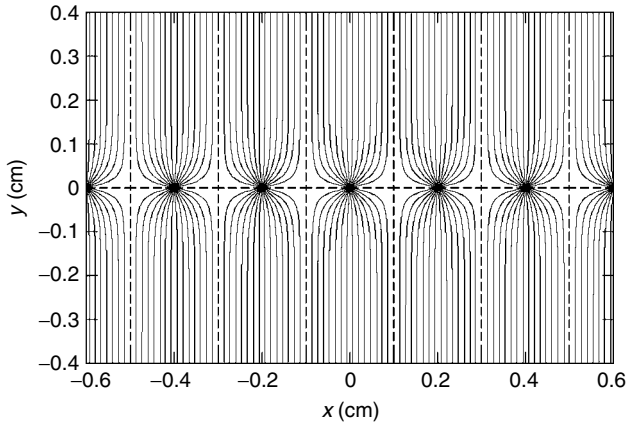


Figure 8. MWPC electric field lines (7). Ionization charge drifts along the field lines to the anode wires. Close to the wires, the field strength increases, accelerating the charge and resulting in charge amplification. The vertical dashed lines indicate virtual boundaries provided by the uniform field.

α. Fast electrons create ionization after a characteristic distance given by $1/\alpha$; the ionized electrons are accelerated in the electric field and create more ionization across the same distance scale. For a uniform electric field, this results in an avalanche of ionization that grows exponentially by $\exp(\alpha z)$, the effective charge gain of the chamber after a distance z.

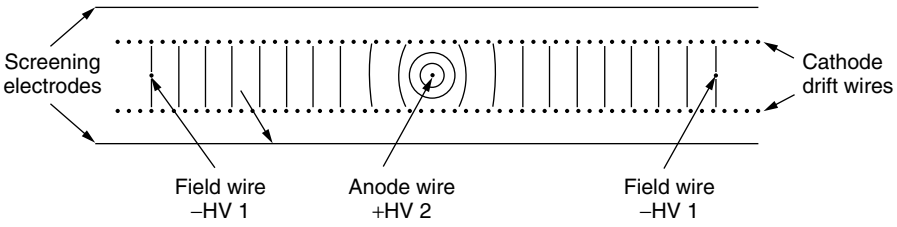


Figure 9. Principle of operation of the drift chamber (30). Particles that perpendicularly to the drift chamber produce ionization electrons that drift toward the anode wire. The cathode wires provide a uniform field through most of this drift region. Measuring the drift time, together with knowledge of the drift velocity obtained in calibration measurements, localizes the position of the track.

Wire chambers use the high field region around thin anode wires to create an avalanche that produces a larger, more easily detected signal. In the *proportional mode*, multiplication occurs in a limited region near the anode. Initial ionization drifts without multiplication across most of the chamber volume. Only once the charge cloud reaches the anode does the avalanche occur. Thus, the collected signal charge is proportional to the original ionization.

The avalanche signal collected on the anode consists of two components: a very fast pulse due to the arrival of the electrons and a much slower signal induced by the motion of ions away from the anode wire that typically lasts some microseconds. The latter provides most of the detectable signal. Electronic filtering of the signal by using differentiating electronics restores the timing information and keeps the fastest component of the signal.

The multiwire proportional chamber (MWPC) (32) achieves a uniform drift field across most of the sensitive volume, and a series of equally spaced, thin anode wires is sandwiched between the cathode planes (see Fig. 8).

The drift chamber takes this concept further and drifs charge across larger distances. The initial ionization charge is localized by measuring the time taken for the charge to drift from its source to the signal electrode. The calibrated drift velocity is applied to extract the distance traveled.

For precision particle imaging, the ionized charged must be localized to maintain positional accuracy. Detailed knowledge of the electric drifting field is required, as well as of the optimized gas mixture and pressure, and the field strength, to minimize diffusion of the electron cloud. Figure 9 shows a schematic of the drift chamber concept (33). A uniformly changing electric potential distribution is set up on the cathode wires away from the anode wires. Close to the anode, field uniformity is lost, where the avalanche field is set up. Geometries vary from planar to cylindrical (see Fig. 10).

The inherent ambiguity in selecting between symmetrically placed tracks on either side of the anode wires can be resolved by staggering the anode wires. The stagger provides a time offset of opposite sign for signals coming from each side of the wires. A consistent fit to the signals can thus resolve the ambiguity.

The time projection chamber (TPC) takes the drift concept to the limit (34). Using a large active volume that has parallel electric and magnetic fields, the charge cloud dispersion is minimized. TPCs that have drift distances of 1–2 m have been successfully operated (35). A proportional mode anode wire or anode pad plane that provides high-speed time digitization of the signals

(a)



(b)

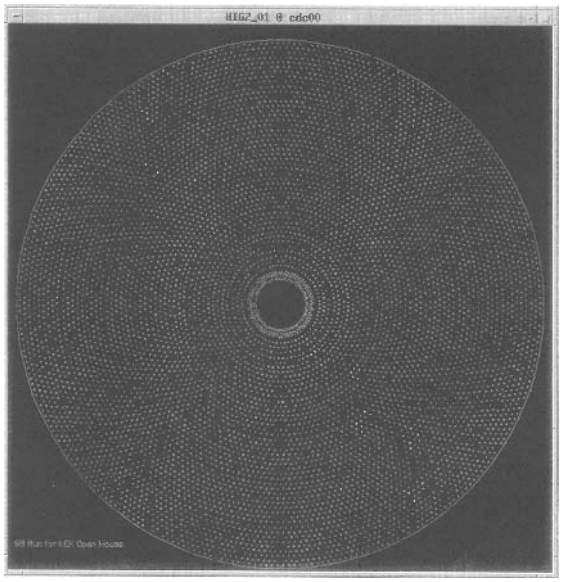


Figure 10. An example of a cylindrical drift chamber from the BELLE experiment (31).

results in three-dimensional reproduction of the original ionization. Two of these coordinates result from the end plane anode geometry, the third from timing.

The end-cap detecting element of the TPC, uses the multistep avalanche chamber to reduce the distorting effects of the anode wire on the drift field shape. The large volume drift space that has a uniform electric field is isolated from the anode field by one or more planes of mesh or wires. Gas amplification can proceed by using the parallel plate amplification region of high field or via the MWPC anode wire avalanche. The additional grids can be used to gate the drift electrons from the amplification region, allowing the possibility of accepting signals only at selected times and reducing the ion buildup in the avalanche region (see Fig. 11).

Drift Tubes. In high rate environments, multiple particle closely spaced tracks degrade the performance of an open structure drift chamber. A drift tube that consists of an anode wire in the center of a conducting cathode tube significantly reduces the volume for multiple track interference. The positional resolution within the tube is determined by the drift time from the track ionization to the central anode. An example of the performance of a large drift tube array can be found in (36).

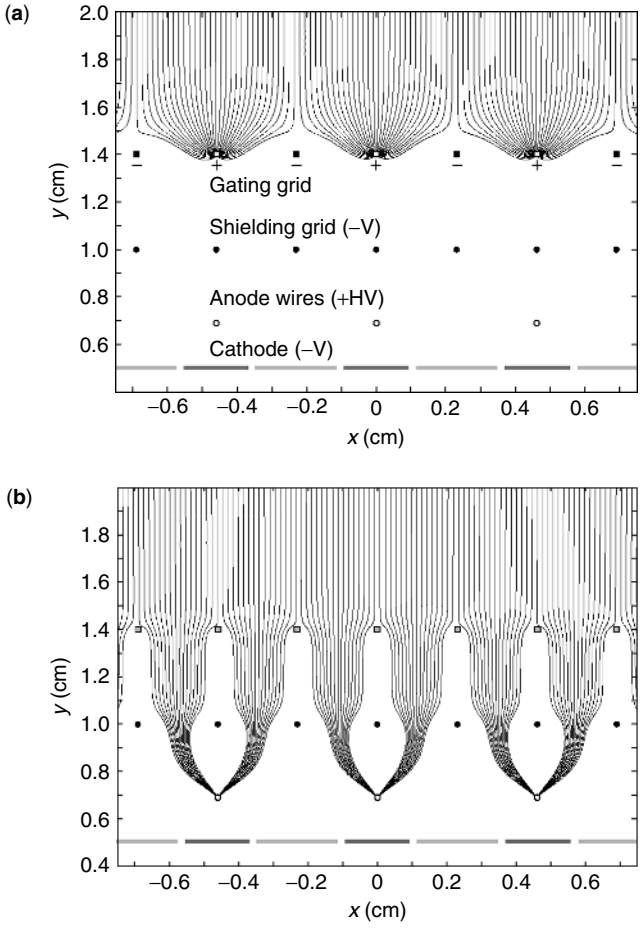


Figure 11. Example of the use of a gating grid to isolate the signal anode until a trigger is received (7). (**a**) Gating grid "closed," field lines end on grid wires; ionization charge from the upper drift volume drifts along field lines and does not reach the anode wires. (**b**) Gating grid "open," field lines guide drifting ionization electrons to the anode wires for detection

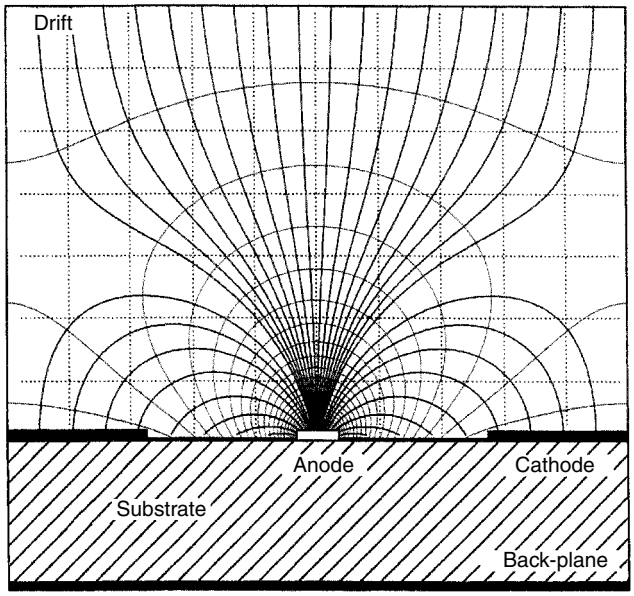


Figure 12. Electric field lines in an MSGC. The ionization charge produced in the upper drift region is drifted toward the anode. The increased field strength in the anode region results in charge amplification for signal detection.

Microstrip Gas Chambers. The microstrip gas chamber (MSGC), is a development of the multiwire proportional chamber. It uses an insulating substrate that has thin conducting strips on the surface and alternating positive and negative potentials on the strips (see Fig. 12).

Two important improvements are achieved by using MSGCs [see, for example, (37,38)]. First, submillimeter pitch can be achieved. In MWPCs, such fine wire spacing is impractical. Second by using a closed spacing (100 μm) of anode and cathode conductors, positive ions are cleared from the anode region much more quickly than by the millimeter or greater spacing in MWPCs. At high particle rates, the slow moving positive ions build up in the anode region, resulting in a space charge that reduces the electric field strength and hence, the avalanche gain and detection efficiency. Significant loss in particle detection efficiency occurs at particle rates of 10^4-10^5 counts/mm^{2}s. To take full advantage of this improvement, the substrate must be slightly conducting in the region between the cathode and anode strips. A semiconductor substrate or the use of surface treatments achieves the necessary surface resistance. Count rates in the millions mm^{-2}s^{-1} have been observed without significant reduction in the gas gain.

Micropattern Gas Detectors. Variants of the MSGC that have micropatterned structures to provide the necessary electric field shaping have been developed. They offer more robust operating conditions than MSGCs which are sensitive to breakdown in the presence of heavily ionizing particles such as alphas or heavy ions. Gas electron

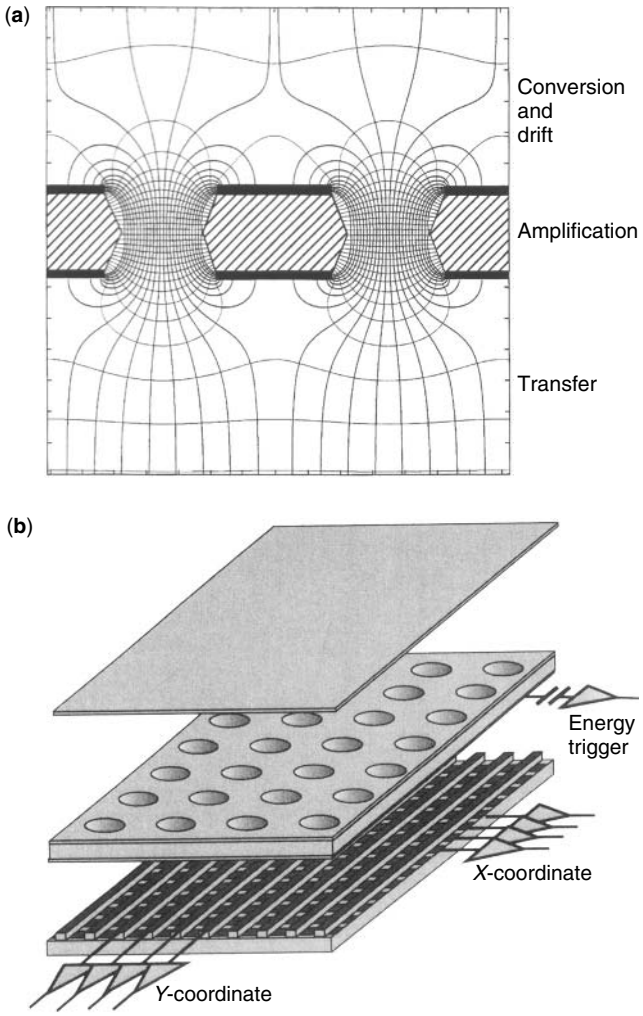


Figure 13. (**a**) Schematic of a single stage GEM detector that has a two-dimensional readout (40). Increased field strength in the amplification region of the detector separates the drift and transfer regions. (**b**) Signals induced in the amplification plate are proportional to the total ionization charge passing from the drift region. Crossed x, y anode planes provide positional information.

multiplier (GEM) devices achieve the necessary anode field shapes by using a perforated metal (cathode) sheet overlying an anode layer (see Fig. 13).

The multistage GEM (43) is an extension of the multistep wire chamber (41) to this technology. This device consists of a double-sided metal on an insulating substrate and an array of 100-mm diameter holes on a 200-mm pitch spacing (see Fig. 14). External electric potentials are arranged to drift electrons created by ionization toward the GEM. High fields in the hole regions generate avalanche conditions. The avalanche-produced electrons pass through the hole in the lower region, where further external fields drift the charge cloud away from the avalanche region to a detection electrode. The separate electrodes for avalanche multiplication and charge detector provide for the possibility of protecting the readout electronics from the effects of the discharge. Multistage GEM detectors allow reducing the avalanche gain on each stage, resulting in safer operating conditions.

Ring-Imaging and Water Cerenkov Counters. Ring-imaging Cerenkov counters (RICH counters) image the cone of Cerenkov light emitted when a particle passes though a "radiator." RICH devices often use a dual radiator, one a liquid or a heavy, high-pressure gas, the second a low-pressure gas or lightweight aerogel. Mirrors are used to focus the light onto a photon-sensitive detector plane. The photons are typically UV. UV-absorbing impurities in the radiators must be tightly controlled.

The large-area photon detector is usually a TPC or drift chamber using a dopant gas that generates photoelectrons from UV absorption. The dopant gas TMAE (tetrakis-dimethylaminoethylene) has been successfully used in this application. The avalanche process itself generates UV photons along with ionization. Designs to prevent feedback from these photons include shields and electrical gating grids. An example of a double radiator RICH counter is shown in Fig. 15 (from 44).

RICH counters are used mostly to identify particles of different mass. However, the Cerenkov rings produced can be used as a tracking signal. Large water Cerenkov detectors, whose capacity is thousands of tons of water, equipped with arrays of photomultiplier tubes, have been used to measure extraterrestrial neutrinos [see Fig. 16 from (45)].

Semiconductor (silicon) Detectors. In forming crystalline solids, atoms are bound into regular arrays. The

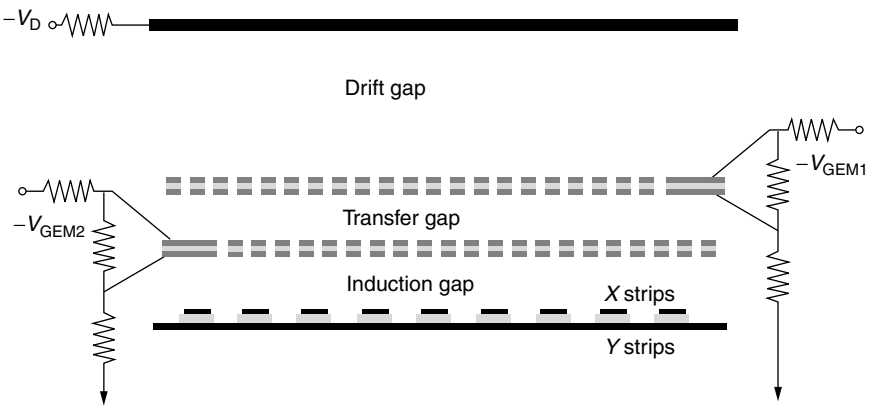


Figure 14. Schematic of a double GEM detector (40). Using two amplification stages, the amplification gain required for detection is the product of the two lower gain processes, resulting in more stable operation.

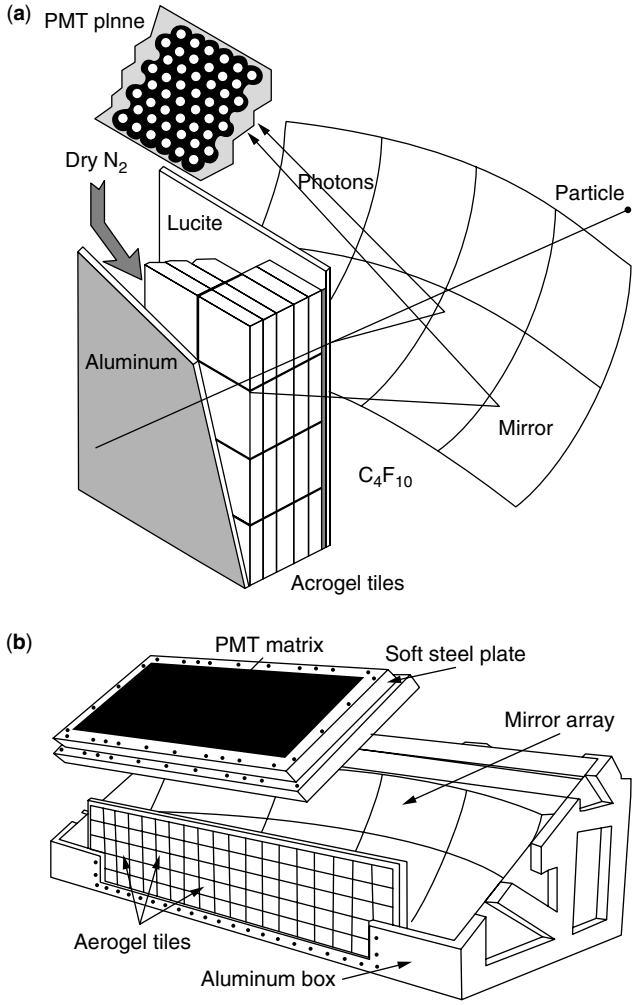


Figure 15. Example of double-radiator RICH counter (41).

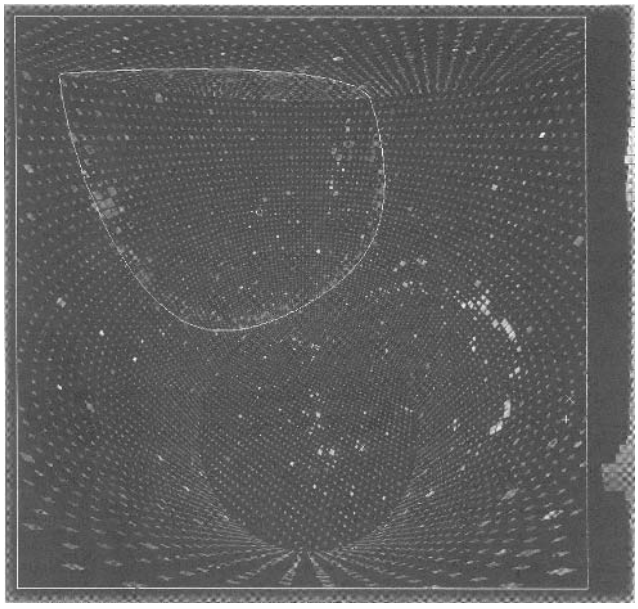


Figure 16. Kamiokande water Cerenkov detector (42). This picture is a depiction of signals from a high-energy interaction in a large water vessel, instrumented with photomultiplier tubes on the inside faces of the vessel. Cerenkov light is emitted from a cone about the axis of high-energy particles. Clearly seen is the ring of signals emanating from a high-energy particle in the water.

atomic electron (line) energies are broadened into energy bands through neighbor interactions. In ("intrinsic") semiconductor crystals, the atomic valence electrons bind neighbor atoms. At zero absolute temperature, all valence electrons are bound in this way, and the valence energy bands are full. Electrical conductivity is zero. The first excited atomic states form the conduction band in semiconductors, and the gap between valence and electron bands is approximately 1 eV (Table 1). Above zero temperature, thermal fluctuations promote some valence electrons into the conduction band, leaving a vacancy or "hole" in the valence band. The conduction band electrons and the holes in the valence band are thus thermally generated current carriers and contribute to limited electrical conductivity. At room temperature, the density of conduction band electrons is approximately 10^{11} cm^{-3} in a crystalline intrinsic semiconductor. The current carrier density follows the Boltzmann temperature dependence and is proportional to $\exp(-E_g/kT)$, where E_g is the energy band gap, k is Boltzmann's constant, and T is the absolute temperature.

Doped or "extrinsic" semiconductors provide control over the electrical properties of semiconductor devices. A (typically small) concentration of impurities is added to pure, crystalline semiconductors, replacing some of the intrinsic atoms in the lattice. The valence of "p-type" impurities is 3, and "n-type" have valence 5. There are only three valence electrons in p-type impurities to bind with neighbors, resulting in an easily filled bond in the neighboring region. Other electrons in the valence band

Table 1.

	C Diamond	Si	Ge (at ($T = 77$K)	GaAs	CdTe	HgI_2
Atomic number, Z	6	14	32	31–33	48–52	53–80
Density (g cm^{-3})	3.36	2.33	5.32	5.35	6.06	6.30
Radiation length (g cm^{-2})	40.3	21.82	12.25	12.19	8.91	7.44
Band gap energy (eV)	6.0	1.1	0.74	1.45	1.47	2.1
Energy per e–h pair,(eV)	13.6	3.61	2.98	4.2	4.43	4.22
Mobility, μ (cm^{2}V^{-1}s^{-1}), e	1,800	1,350	38,000	8,500	1,000	100
Mobility, μ (cm^{2}V^{-1}s^{-1}), h	1,200	480	45,000	400	65	4
Dielectric constant, ε	5.7	11.9	16.2	13.1	10	7.4

can fill this bond, leaving a hole in the valence band. Boron, gallium, and indium are p-type impurities. n-type impurities such as antimony, phosphorous, and arsenic provide additional electrons in the conduction band.

Intrinsic semiconductors have equal densities of "free" holes and electrons. At room temperature, doped semiconductors typically have larger current carrier densities than their intrinsic base materials. However, the product of the concentrations of conduction band electrons n and of valence band holes p is constant:

$$np = n_c n_v \exp(-E_g/kT) \tag{15}$$

where n_c, n_v are the density of states in the conduction and valence bands, respectively.

By fabricating a junction between n-type and p-type materials, the higher concentration of mobile electrons and holes, respectively, causes electrons from the n-type valence band to diffuse across the junction into the p-type region, and holes from the p-type conducting band diffuse into the n-type. These diffusing carriers "recombine" in the junction region. The initially charge neutral regions will then have a fixed lattice charge, positive on the n-type side and negative on the p-type side. The charged region, called the depletion region, produces an electric field across the junction, which opposes the diffusion process and results in an equilibrium depletion width that has a "built-in potential" V_{bi} across the junction.

An externally applied "reverse" bias of appropriately chosen polarity across the junction will move majority carriers away from the junction, increasing the depletion width and increasing the electric potential difference across the junction. The remaining current is due to minority carriers, created primarily by thermal generation in or near the depletion region that drift across the junction under the influence of the electric field.

The width of the depletion region depends on the doping density and the applied voltage V_B as

$$x_d = \sqrt{2\varepsilon\mu_e\rho(V_{bi} + V_B)} \tag{16}$$

where V_{bi} is the built-in junction potential, ε is the permittivity constant, μ_e is the electron mobility, and ρ is the material resistivity. Using a relatively high resistivity material (1–20 kWcm), an applied external bias (typically 50–500V) results in an appreciable depletion width (typically 20 μm for silicon CCDs to 300 μm for silicon strip detectors).

Devices are typically fabricated using an ion-implanted, heavily doped (low resistivity) region close to the surface of a high resistivity bulk semiconductor. Thus, the depletion region grows asymmetrically from applied bias and reaches into the high resistivity bulk region. Detectors become "fully depleted" when the bias is sufficient to deplete the entire thickness of the bulk of charge carriers.

A high-quality, depleted, reversed bias semiconductor diode has a low "dark current" and a large depletion thickness. Ionizing particles passing though the detector create electron–hole pairs along their trajectory in the semiconductor. In the depletion region, this ionization injects a significant charge into this carrier-free region. The electric field across the junction separates the electrons and holes and induces a current pulse through the device. The measurement of the current pulse above the quiescent dark current is the task of low noise electronics using appropriate filtering.

Table 1 shows relevant properties of various semiconductor materials, including the energy required to create electron–hole pairs. Approximately 80 electron–holes pairs are liberated, on average, by a MIP per micrometer of path length in silicon. Thus, for fully depleted silicon detectors 300 μm thick, a MIP passing perpendicularly through the detector generates a signal pulse of approximately 24,000 electron–hole pairs, on average. The ionization process is stochastic, resulting in significant fluctuations in the signal, which must considered when evaluating the signal-to-noise ratio and the margins required for clean signal detection (46).

Strip Detectors. The application of microelectronics techniques to the development of silicon strip detectors has been extremely fruitful for high precision particle tracking (47–52). A large-area p-n diode is segmented into narrow strips along the length of the diode. Each strip is coupled to a separate readout electronic channel (see Fig. 17). The position of a particle incident on the strip detector is localized by determining the strip (or strips) upon which the signal is induced. Measurement of the total charge signal also provides a measure of the total energy deposited in the silicon. Detectors that have a strip pitch of 25 μm are produced routinely. Diffusion of the signal charge can result in spreading the signal across neighboring channels. Analog readout of the signals and interpolation between the strips according to signal size can result in even better resolution.

Using a uniform, approximately planar electric field in the bulk of the detector, the charge is shared between neighboring strips approximately linearly with the relative position of the particle trajectory between the

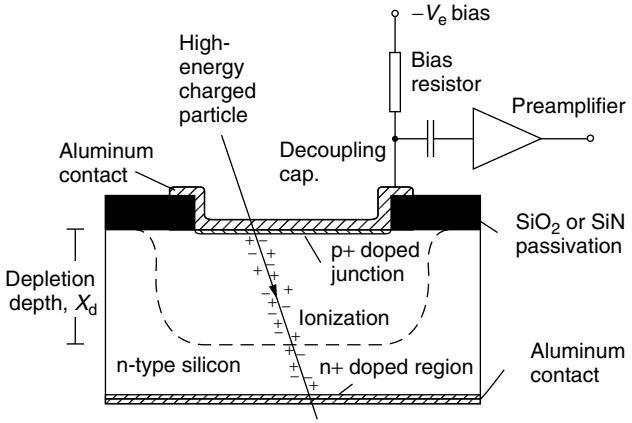


Figure 17. Schematic of a cross section through a p-n silicon detector. The applied bias depletes the silicon of charge carriers in the depletion region. The ionization resulting from a high-energy particle passing through the depletion region drifts to the top and bottom faces of the detector. Electronics detects this charge pulse. Sufficient bias fully depletes the silicon, resulting in maximum efficiency for charged particle detection.

strips. The quantity

$$\eta = \frac{\mathrm{Sig}_2}{\mathrm{Sig}_1 + \mathrm{Sig}_2} \tag{17}$$

gives a good approximation of the relative distance between strips. Sig_1 and Sig_2 are the signal charges from neighboring strips. Corrections for nonlinearity close to the strips can be determined via calibration using tracks of well-defined positions.

Capacitive coupling between the strips can be exploited to reduce the number of independent electronic channels needed to read out the signals. This coupling results in the distribution of the charge deposited amongst several strips. By reading out every second channel (or even fewer) using analog electronics, the center of the charge signal can be determined by interpolation between neighboring readout channels. Signal-to-noise ratios in excess of 10 are required to maintain the positional resolution of the configuration when all strips are read out. The intermediate strips need to be held at the same potential as the readout strips to avoid distorting the planar electric field. Distribution of the bias potential to the strips must be via high-resistance connections. The number of intermediate strips must be optimized. The advantage of fewer electronics channels must be weighed against the increased capacitive load on the readout electronics and hence, on the larger amplifier noise. Furthermore, some signal charge is lost via capacitive coupling to the back-plane, an effect that increases as the number of intermediate strips increases.

The bias applied to deplete the detector will generally need to be decoupled from the amplifier inputs. This can be achieved by using external decoupling networks; however, a more convenient approach is to provide the decoupling capacitors directly on the strips. This is achieved by applying an insulating SiO_2 or Si_3N_4 layer over the metallized diode strips, and metallization is on top for connection to the readout (53). [See Fig. 18 (54).] Capacitance in excess of 100 pF can be generated in this way. Such a decoupling capacitance that is much larger than the silicon sensor strip capacitance ensures practically complete charge collection by the amplifier circuits. Biasing is provided by polysilicon bias resistors fabricated on the detector. Alternatively *punch-through* (or *reach-through* or *foxfet*) biasing can be applied (55).

The simple p-n diode detectors, described so far, collect holes on the readout strips. The electrons drift to the back-plane. A second coordinate can be obtained by segmenting the back-plane into strips orthogonal to the p-side strips. However, the strip isolating oxide contains a positive charge and results in electron accumulation at the oxide–silicon interface, eventually shorting the segmented strips. Various techniques have successfully alleviated this problem. Implantation of p-doped "blocking strip" in the center of the isolating oxide strips breaks the otherwise continuous accumulation layer. Alternatively, a biased metal gate over the isolation oxide can control the accumulation, or a compensating p-doped region can be implanted in the back side before the oxide layer is grown (56).

Double metal processing of the silicon, where a second set of strips isolated from the first metal layer by an oxide is added, offers considerable practical convenience. Using an appropriately designed layout, the crossed strips from the top and bottom sides can both be read out from the same end of the silicon detector. This is particularly useful when cascading several detectors to a single set of readout electronics. [see Fig. 19 (57).]

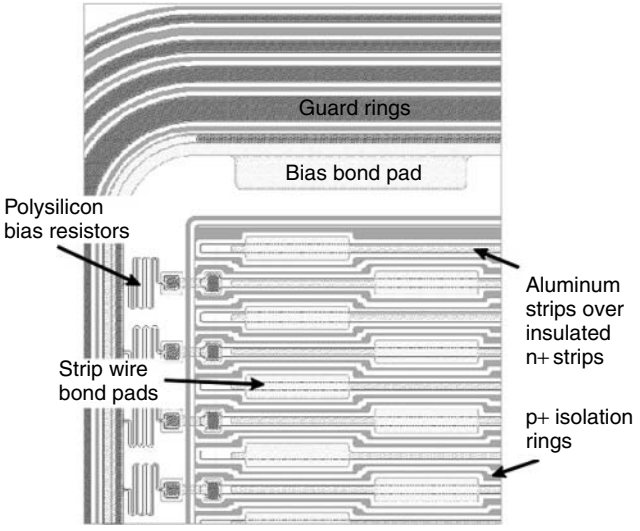


Figure 18. Schematic view of a corner of the n+ side of a capacitively coupled silicon microstrip detector (51). Shown are the polysilicon bias resistors connecting the ion-implanted n+ strips to the bias ring and aluminum readout strips over the implanted regions. These are isolated from the silicon by an insulating layer, usually of silicon dioxide, silicon nitride, or polyimide, that creates a decoupling capacitor between the bias silicon strips and the readout electronics.

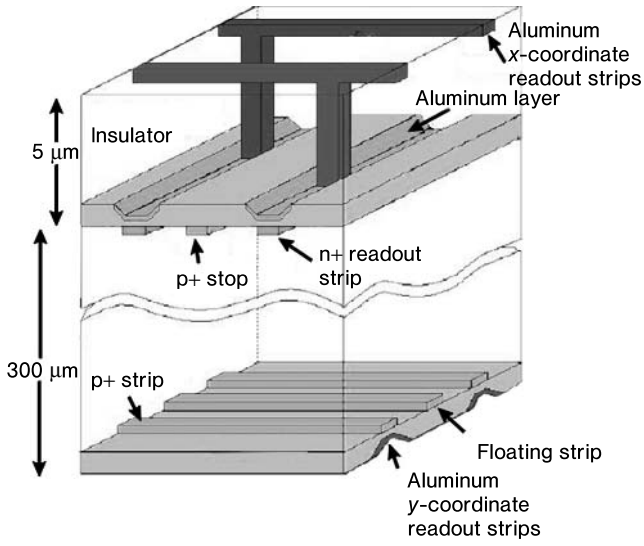


Figure 19. Schematic of a double-sided silicon strip detector, using "double metal technology" on the top layer. The additional aluminum layer allows connecting both coordinate directions at the end of the detector and offers considerable simplification in building a system of such detectors.

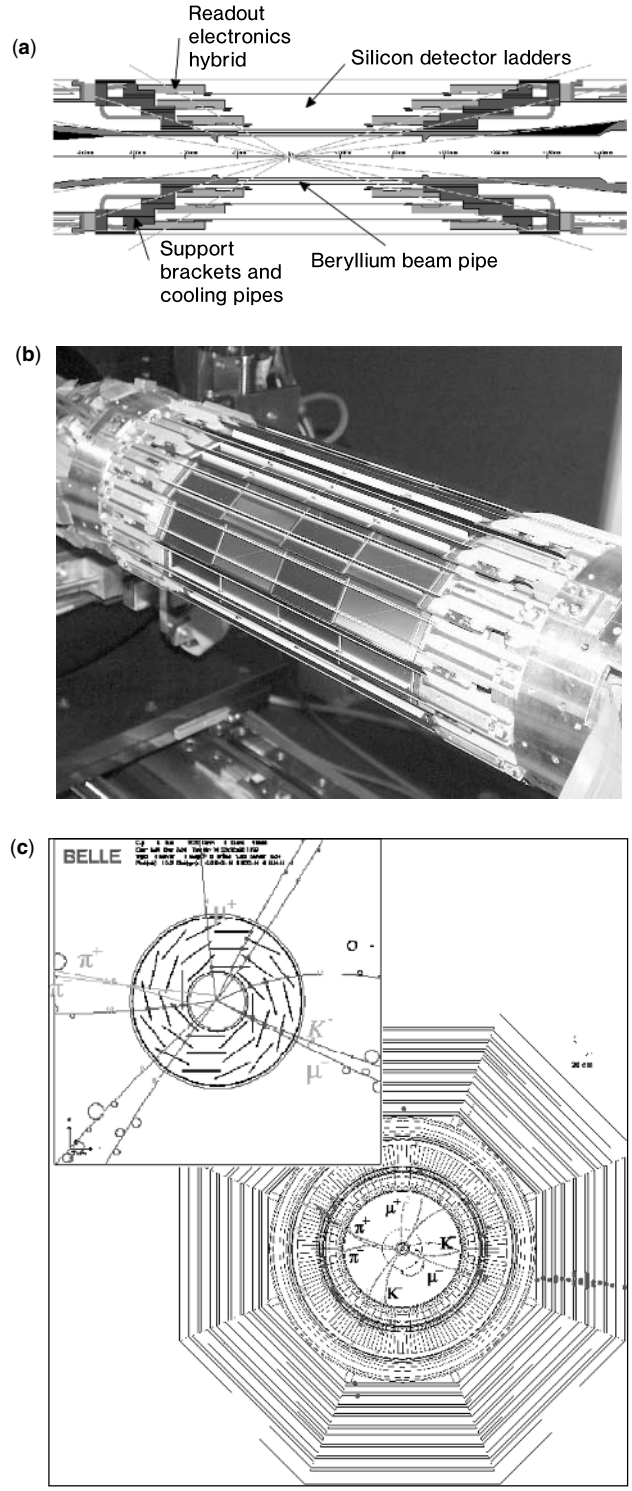


Figure 20. (**a**) Schematic cross section of the BELLE silicon vertex detector (55). (**b**) Photograph of the BELLE silicon vertex detector. (**c**) Tracks originating from a particle interaction, reconstructed using the BELLE vertex detector together with reconstructed interaction using the entire BELLE detector.

The silicon strip detector has matured principally through its development and use for measuring tracks emanating from short-lived particles containing heavy quarks in high-energy physics experiments. Track resolutions less than 10 μm are routinely achieved, resulting in determining the position of the particle decay vertices to a precision of 10–100 μm. [See Fig. 20 (57)].

Charge-Coupled Devices. Application of the highly successful optical CCD technology for particle detection and imaging has been successful based on some modifications to these structures. The CCD architecture results in routing all elements of the device through a single output channel, making the readout time long. Furthermore, the device is sensitive to signals from later arriving particles during the readout phase. Therefore, the CCD is an appropriate solution for low rate environments (58).

Silicon Drift Detectors. Silicon drift detectors (59) are fabricated where p-n junctions are on both sensor surfaces. The junctions are in the form of strips. These devices are depleted from both surfaces. An electric field gradient parallel to the device surface is provided by graded potentials on the strips. Electrons generated by ionizing particles are collected in the middle of the silicon where the potential is minimum, and are drifted out of the sensor under the influence of the parallel field component to readout electrodes formed by a segmented ion-implanted n+ region. [See Fig. 21 (60] The combination of readout electrode segmentation and drift time provides a two-dimensional localization of the original ionization charge.

Optimum fabrication and operation of these devices demands low-defect, high-uniformity silicon wafers to provide good charge collection, precision on-detector voltage dividers to provide good drift field uniformity, and precise temperature control to maintain stable drift velocity. Spatial resolution of 50 μm from timing and 30 μm from anode segmentation have been demonstrated (61).

Pixel Detectors. True, two-dimensional, segmented semiconductor detectors that have active, individual element readout electronics go beyond CCDs in speed, but also in complexity. The advent of very high rate, high particle multiplicity interactions in recent years

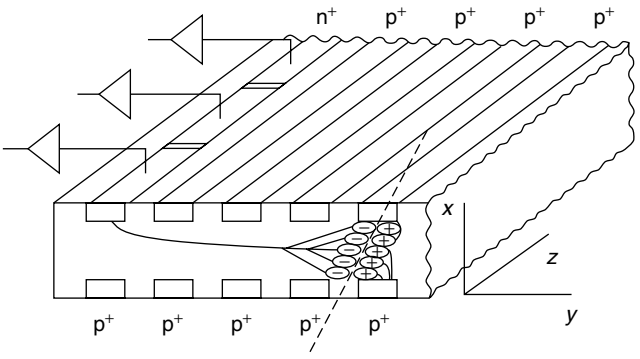


Figure 21. Schematic of the silicon drift detector. Ionization from the charged track is drifted under the influence of the electric field, provided by the p+ strips to which a graded electric potential is applied, to the segment n+ strip attached to the electronics readout.

has driven the development of these new active pixel devices (62).

Both monolithic and hybrid devices are being developed. The latter are considerably more advanced; separating the sensor from the electronic chips allows parallel development of the two components. It also facilitates the use of different sensor materials. Silicon sensors are the most mature technology and offer good performance for charged particle detection. Compound semiconductors of higher atomic number such as GaAs are being pursued for X- and gamma ray applications. Diamond sensors offer very high radiative hardness.

Rectangular pixels of areas as low as 55×55 μm have been produced (63). The small element size has the further advantage of reduced input capacitance to the readout amplifier, resulting in considerably better noise performance. Both binary and analog electronics are being developed, resulting in tracking precision of 10–20 μm in the former and approaching 1 μm in the latter. The hybrid approach uses metal bump bonding to attach the electronic chip to the segmented sensor. The metal typically used is indium or some alloy that has similar characteristics.

Electronic Signal Detection. The necessary signal processing and electronic readout to measure and record the detected signals and produce is an essential part of the particle detection process. The reader is directed to the literature for details of this subject (64–66).

Radiation Damage

Wire Chamber Aging. Prolonged operation of wire chambers exposed to large particle fluxes leads to deposits on the anode wires. These aging effects are strongly affected by the gas mixture used and on the operating conditions, especially the gas gain. High gain operation results in larger charge exposure of the anode wires that results in forming polymerized layers and "whiskers." The build-up on the anode wires results in reduced field strength in the multiplication region and hence reduced signal gain. Whiskers can lead to electrical breakdown. Operation in a lower gain mode can extend the lifetime of wire chambers (67,68).

Radiative Damage in Semiconductor Detectors. Radiative damage in silicon detectors is of two basic categories: bulk damage caused by displacement of atoms from the lattice and surface damage due to charge buildup at the interface between silicon and oxide layers.

Bulk damage depends upon the nonionizing energy deposited in the lattice. The damage rate is thus radiation type and energy-dependent. The displacement damage results in increased dark current, signal charge trapping, and changes in the effective charge carrier density.

The increase in dark current density is proportional to the radiative particle fluence:

$$\Delta I_{\text{dark}} = \alpha \Phi. \tag{18}$$

Annealing of the damage that produces the current increase reduces the effective damage constant α. After long-term annealing (approximately 1 month at room temperature), it reaches a value of $\alpha \approx 3 \times 10^{-17}$ A cm^{-1} for high energy protons and pions and $\alpha \approx 2 \times 10^{-17}$ A cm^{-1} for 1 MeV neutrons.

The effective carrier concentration in initially n-type silicon changes approximately as

$$N = N_0 e^{-\delta \Phi} - \beta \Phi \tag{19}$$

where $N_0$is the initial carrier concentration of the n-type silicon, $\delta \approx 6 \times 10^{-14}$ cm^2 is the donor removal constant, and $\beta \approx 0.03$ cm^{-1} is the acceptor creation constant (69). Acceptor creation dominates At high fluences. Finally, the effective carrier concentration changes sign, typically at fluences above 3×10^{13} cm^{-2} (70). The increasing effective carrier concentration that develops at higher fluences requires a large bias to deplete the detector fully.

Surface damage depends predominantly on the ionization created in the device that is trapped at interfaces between insulating and semiconducting layers. Such a charge buildup can result in increased surface currents, can result in FET structures failing, and cause strip isolation structures to become conducting. These effects depend strongly on the processes used in the fabrication and on device geometry and features.

Other semiconductors have been developed in view of potentially improved radiative insensitivity. GaAs and diamond detectors have been constructed. GaAs has been found relatively resistant to neutron radiation but relatively more sensitive than silicon to high-energy pion and proton irradiation.

Diamond detectors are being developed as an extremely radiation-tolerant alternative to silicon. Crystal growth optimization for improved charge signal collection has produced encouraging results for this material which is tolerant to radiative fluences well above those where significant degradation sets in for silicon (71).

BIBLIOGRAPHY

1. B. Rossi, *High Energy Particles*, Prentice-Hall, 1952.
2. K. Kleinknecht, *Detectors for Particle Radiation*, Cambridge University Press, 1986.
3. W. R. Leo, *Techniques for Nuclear and Particle Physics*, Springer, Berlin, 1987.
4. T. Ferbel, ed., *Experimental Techniques in High Energy Physics*, Addison-Wesley, Menlo Park, 1987.
5. G. Charpak, ed., *Research on Particle Imaging Detectors*, World Scientific, Singapore, 1995.
6. C. Grupen, *Particle Detectors*, Cambridge University Press, 1996.
7. Particle Data Group, D. E. Groom et al., *Eur. Phys. J.* **C15**, 1 (2000). also available at url:http:/pdg.lbl.gov/.
8. R. Akers et al., *Z. Phys.* **C67**, 203 (1995).
9. R. M. Sternheimer, M. J. Berger, and S. M. Seltzer, *Atom. Data Nucl. Data Tables* **30**, 261–271 (1984).
10. S. M. Seltzer and M. J. Berger, *Int. J. Appl. Radiat. Isotopes* **35**(7), 665–676 (1984).

11. L. D. Landau, *J. Exp. Phys (USSR)* **8**, 201 (1944).
12. M. Roco, *Nucl. Instrum. Methods* **A418**, 91–96 (1998).
13. Y. S. Tsai, *Rev. Mod. Phys.* **46**, 815 (1974).
14. V. Karimaki, *Nucl. Instrum. Methods* **A410**, 284–292 (1998).
15. W. Innes, Computer Program TRACKERR, Version 2.04, 1998. Available from ftp:/ftp.SLAC.Stanford.edu/software/trackerr.
16. R. Frühwirth, in *CERN School of Computing*, Bad Herrenalb, Germany, 20 Aug.–2 Sep. 1989, C. Verkerk, (ed.), CERN-90-06, CERN Geneva, 1990.
17. R. Bock, H. Grote, and D. Notz, in W. Regler and R. Frühwirth, eds., *Data-Analysis Techniques for High-Energy Physics*, 2d, Cambridge Mono. Part. Phys. Nucl. Phys. Cosmol., **11**, Cambridge University Press, 2000.
18. R. Wigmans, *Calorimetry: Energy Measurement in Particle Physics*, Oxford Clarendon Press, 2000.
19. R. Wigmans, *Annu. Rev. Nucl. Part. Sci.* **41**, 133–185 (1991).
20. J. B. Birks, *The Theory and Practice of Scintillation Counting*, Pergamon, London, 1964.
21. T. Foerster, *Ann. Phys.* **2**, 55 (1948).
22. I. Berlmann, *Handbook of Fluorescence Spectra of Aromatic Molecules*, Academic Press, NY, 1971.
23. W. W. M. Allison and P. Wright, in T. Ferbel, ed., *Experimental Techniques in High Energy Physics*, Addison-Wesley, 1987.
24. B. Dolgoshein, *Nucl. Instrum. Methods* **A326**, 434 (1993).
25. Hamamatsu Photonics, Japan. (*http://www.hpk.co.jp*).
26. R. C. Ruchti, *Annu. Rev. Nucl. Sci.* **46**, 281–319 (1996).
27. P. Annis et al., *Nucl. Instrum. Methods* **A409**, 629 (1998).
28. B. B. Beck et al., *Nucl. Instrum. Methods* **A412**, 191 (1998).
29. S. Biagi et al., *Nucl. Instrum. Methods* **A283**, 716 (1989).
30. V. Palladino et al., *Nucl. Instrum. Methods* **128**, 323 (1975).
31. G. Charpak et al., *Nucl. Instrum. Methods* **62**, 262 (1968).
32. G. Charpak, F. Sauli, and W. Duinker, *Nucl. Instrum. Methods* **108**, 413 (1973).
33. G. Alimonti et al., *Nucl. Instrum. Methods* **A453**, 71–77 (2000).
34. D. R. Nygren and J. N. Marx, *Phys. Today* **31**, 46 (1978).
35. W. Blum and L. Rolandi, *Particle Detection with Drift Chambers*, Springer-Verlag, 1994.
36. W. Riegler et al., *Nucl. Instrum. Methods* **A443**, 156–163 (2000).
37. A. Oed, *Nucl. Instrum. Methods* **A263**, 351 (1988).
38. R. Bellazzini and A. M. Spezziga, *Rivista del Nuovo Cimento* **17**, 1 (1994).
39. F. Sauli and A. Sharma, *Annu. Rev. Nucl. Part. Sci.* **49**, 341–388 (1999).
40. F. Sauli, New Developments in Gaseous Detectors CERN-EP/2000-108, 2000.
41. G. Charpak et al., *Phys. Letters* **78B**, 523 (1978).
42. F. Sauli, *Nucl. Instrum. Methods* **A386**, 531 (1997).
43. S. Bachmann et al., CERN-EP/2000-116, 31 August, 2000.
44. N. Akopov et al., preprint DESY-2000-190 (physics/0104033).
45. Y. Fukuda et al., *Phys. Lett.* **B81**, 1,562–1,567 (1998).
46. H. Bichsel, *Rev. Mod. Phys.* **60**, 663–699 (1988).
47. G. Lutz, *Semiconductor Radiation Detectors: Device Physics*, Springer-Verlag, NY, 1998.
48. J. Kemmer, *Nucl. Instrum. Methods* **169**, 499 (1980).
49. E. H. M. Heijne et al., *Nucl. Instrum. Methods* **178**, 331 (1980).
50. J. B. A. England et al., *Nucl. Instrum. Methods* **185**, 43 (1981).
51. J. T. Walker et al., *Nucl. Instrum. Methods* **226**, 200 (1984).
52. H. W. Kraner et al., *IEEE Trans. Nucl. Sci.* **30**, 405–414 (1983).
53. M. Caccia et al., *Nucl. Instrum. Methods* **A260**, 124 (1987).
54. P. Allport, private communication.
55. J. Kemmer and G. Lutz, *Nucl. Instrum. Methods* **A273**, 588 (1988).
56. G. Lutz and A. S. Schwarz, *Annu Rev. Nucl. Part. Sci.* **45**, 295–335 (1995).
57. Belle Collaboration, internal note.
58. G. Alimonti et al., *Nucl. Instrum. Methods* **A453**, 71–77 (2000).
59. C. Damerell et al., *Nucl. Instrum. Methods* **A288**, 236 (1990).
60. P. Rehak et al., *Nucl. Instrum. Methods* **A235**, 224 (1985).
61. P. Weilhammer, *Nucl. Instrum. Methods* **A453**, 60–70 (2000).
62. E. H. M. Heijne, CERN EP/2000-137, 2000.
63. B. Mikulec, M. Campbell, G. Dipasquale, C. Schwarz, and J. Watt, *Nucl. Instrum. Methods* **A458**, 352–359 (2001) (Medipix collaboration).
64. V. Radeka, *IEEE Trans Nucl. Sci.* **NS-21**, 51 (1974).
65. F. S. Goulding and D. A. Landing, *IEEE Trans. Nucl. Sci.* **NS-29**, 1,125 (1982).
66. H. Spieler, *IEEE Trans. Nucl. Sci.* **NS-29**, 1,142 (1982).
67. J. A. Kadyk, *Nucl. Instrum Methods* **A300**, 436–479 (1991).
68. J. Vavra, *Nucl. Instrum. Methods* **A252**, 547–563 (1986).
69. G. Lindstroem et al., *Nucl. Instrum. Methods* **A426**, 1 (1999).
70. R. Wunstorf, PhD Thesis, University of Hamburg, 1992.
71. RD48 Collaboration: S. Schnetzer et al., *IEEE Trans. Nucl. Sci.* **46**, 193–200 (1999).

PHOTOCONDUCTOR DETECTOR TECHNOLOGY

ANDREW MELNYK
Rochester, NY

INTRODUCTION

Photoelectric detectors are semiconducting devices that convert optical signals into electrical signals. They rely on the quantum nature of matter, whereby incident photons, ranging from high energy gamma rays through visible light to infrared rays, excite electrons in matter to produce an electrical charge. The operation of a photoelectric detector involves two steps; (1) conducting charge generation by an incident light photon and (2) charge transport and/or multiplication by whatever current gain mechanism may be present. Photoelectric detectors are classified as photoemission and photoconductive detectors. Photoemission

occurs when the energy of the photon $h\nu$ is sufficient to eject an electron from the surface of the solid into vacuum, and the phenomenon is the basis for photomultiplier tubes. Photoconductivity, sometimes referred to as internal photoemission, is observed in a solid when the absorbed photon raises an electron from a nonconducting energy state to a conducting energy state where it contributes to electrical conductivity.

The origin of the field of photoconductivity is attributed to Willougby Smith (1) who noted changes in the resistivity of selenium resistors when they were exposed to light. Systematic treatment of the photoconductivity process was initiated in the 1920s by the investigations of Gudden and Pohl and the Göttingen school (2), principally using zinc sulfide, diamond, and alkali halides. The investigation of photoconductivity in the first half of the twentieth century benefited greatly from the rapid growth of experimental and theoretical interest in condensed matter physics and chemistry. The period from the 1920s to the 1950s saw a rapid expansion in the understanding of photoconductivity. The quantum theory of matter, on the one hand, and industrial interest in devices, on the other, were the prime motivators. The experimental effort, especially in the 1940s and 1950s, was strongly influenced by the television industry and focused primarily on covalently bonded semiconductors, although some notable exploratory studies of the phenomenon in molecular solids were carried out. The detailed quantitative studies, concepts, and measurements of photoconductivity developed by the middle of the twentieth century are discussed in the now classic books by Bube (3), Rose (4), and Ryvkin (5). In the second half of the twentieth century, the explosive growth of the semiconductor industry led to the development of a wide variety of specialized photodetectors based on crystalline semiconductors of covalently bonded silicon and germanium, as well as compounds of elements from the III–V and II–IV columns of the periodic table of elements. During the same period, the photocopying and later the laser printing industries led to the development of very large area photodetectors. These initially used amorphous chalcogenide glasses but now almost exclusively use organic electro-optical materials. For a comprehensive list and detailed discussion of these organic photoconductors, the reader is referred to the book by Borsenberger and Weiss (6).

This section on photoconducting detectors explains the fundamental concepts of photoconductivity and describes the elementary photoconducting devices, their design, materials, and how they work. Except for two, photodectors incorporated in specific imaging devices and an explanation of their operation are not considered here because they are discussed in other sections on specific applications such as UV or IR detectors, CCD imagers, etc. The exceptions are electrophotographic and vidicon-type television camera photoconductors that can be understood only as imaging devices. In contrast to other imaging photodetectors, electrophotographic photoconductors are not covered in any other sections. For this reason and because they differ significantly in their design, materials and operation from the other photodetectors, most of this section is devoted to a detailed discussion of electrophotographic photoconductors.

PHOTOCONDUCTOR CONCEPTS

Band Theory of Matter

Before discussing photoconductors, it is useful to review the modern band theory of solids, which provides the basis for understanding the electrical and photoelectrical properties of solids.

The discrete, outer electron energy levels of an atom within a solid are broadened into bands of allowed energies several electron volts (eV) wide. Each band actually contains N discrete levels, but because N, the number of atoms per cubic centimeter, is about 10^{22}, these levels are separated by 10^{-22} eV and hence are essentially continuous. Between these allowed energy bands are ranges of forbidden energies called band gaps. The widths of the bands and band gaps and their occupancy by electrons depends solely on the chemical elements of the solid and their atomic arrangement.

The electronic nature of the solid is determined by the electron occupancy of the last or highest filled band. If the band is partially filled by electrons, it is called a conduction band, and the solid is a metal. Such a solid is a very good conductor because even a very small electrical field applied to the solid will excite electrons to adjacent unoccupied higher energy levels and hence accelerate them and cause an electrical current to flow. The energy level to which the band is filled is called the Fermi energy level. The other possibility is that the highest occupied band is completely filled. This band is called the valence band, and the empty one just above it is the conduction band. The electrons are not free to move because there are no empty energy levels in the valence band for them to be excited to. The electrons must be excited to the empty conduction band to provide a conducting charge, which requires energies equal to the band gap or larger. Thermal energy kT is one source of excitation and at room temperature provides 0.025 eV of energy. Because the probability of excitation is proportional to the negative exponential of the energy required beyond kT, the degree of conductivity depends on the size of the band gap. If the band gap is many eV, the material is an insulator, but if the band gap is 1 eV or smaller, the material is a semiconductor.

When an electron is excited from a filled valence band to an empty conduction band, it creates a vacancy that behaves like a positive conducting charge, called a hole. The term “charge carrier” is used to designate either an electron or a hole that can contribute to conductivity. Semiconductors whose charge-carrier densities (number per volume) are thermally excited across the band gap are called intrinsic semiconductors. The Fermi level in an intrinsic semiconductor is in the middle of the band gap. Crystalline silicon (Si), whose band gap is about 1 eV, is the prime example of a semiconductor and is the workhorse of the electronic industry.

Doping of selected atom impurities is used to enhance and control the conductivity of semiconductors. We will use Si as a model to explain doping, although the concept is quite general. The Si atom has four valence electrons, which create covalent bonds to four adjacent Si atoms in a crystal. If a phosphorous (P) atom that has five valence electrons replaces a Si atom, the extra unbonded electron will be very weakly attached to the phosphorous atom. Thus, it can easily be thermally excited into the conduction band, leaving a positively charged phosphorous ion, which is just a deeply trapped hole. This makes phosphorous an electron donor; a very low concentration of phosphorous atoms of the order of parts per million (ppm) in Si will make it an n-doped semiconductor (n for negative charge), where electrons are the majority charge carriers. The few holes that will still be produced by thermal excitation from the valence band are called the minority charge carriers. Due to the extra electrons provided by phosphorous atoms, the Fermi level is moved up from the middle of the band gap closer to the conduction band; the amount of shift depends on the concentration of phosphorous doping. In contrast, boron (B), which has three valence electrons, doped into Si is an electron acceptor and makes it a p-doped semiconductor, where holes in the valence band are the majority charge carriers.

If n-doped and p-doped semiconductors are brought into contact, the Fermi levels are mismatched. This mismatch will cause charge to flow; holes from the p side and electrons from the n side result in regions of space charge (p region depleted of holes and n of electrons) that produces a voltage drop equal to the difference in Fermi levels, about 1 volt for Si. The matching of the Fermi levels results in a bending of the bands in a width equal to the depletion regions.

Conductivity and Photoconductivity

At the most elementary level, a photoconductor consists of a slab of material, assumed for simplicity to be a semiconductor, between two ohmic contacts (Fig. 1). The electron current density J (current per unit area) flowing in the slab is given by the density of electrons n and their net average velocity v,

$$J = env = en\mu E, \tag{1}$$

where e is the electronic charge and, in the ohmic regime where the current is proportional to the applied voltage, v is equal to the product of electron mobility μ and electrical field E. The product $en\mu$ is the conductivity of the slab. For a semiconductor that has electron and hole charge carriers of charge densities n and p and corresponding mobilities, the current density becomes

$$J = e(n\mu_n + p\mu_p)E. \tag{2}$$

Equation (2) is a general expression for the conductivity, and if n and p are the charge densities and are not illuminated, Eq. (2) is the dark conductivity. Photoconductivity is the increase in electrical conductivity due to the additional charge carriers produced by light. In practical photoconductors, dark conductivity is negligible compared to photoconductivity, but in general if we designate n and p as the additional light-produced charge carriers, the photoconductivity of the slab in Fig. 1 is given by $e(n\mu_n + p\mu_p)$.

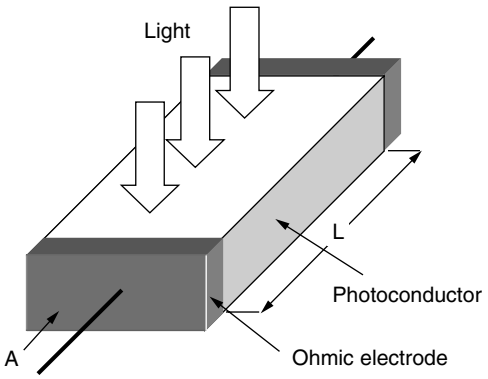


Figure 1. Schematic diagram of a photoconductor that consists of a slab of semiconducting material between two ohmic contacts.

Because photogeneration drives the carrier concentration out of equilibrium, the charge carrier densities are governed by the generation rate per unit volume g of the electrons and holes and their loss due to recombination or trapping. The generation rate is equal to the product of photon absorption per unit volume and the photogeneration efficiency η, which is defined as the number of electron–hole pairs produced per absorbed photon and is often referred to as the quantum efficiency. Photogenerated charge carriers decay at a time constant τ, defined as the lifetime of the charge carrier before it becomes trapped or recombines. At a steady state or equilibrium photoconductivity, the generation rate is equal to the recombination rate; hence the density of photogenerated charge carriers is the product of g and τ. So the photocurrent density expression becomes

$$J = eg(\tau_n\mu_n + \tau_p\mu_p)E. \tag{3}$$

The transit time for the charge carrier to travel the length of the slab L is

$$t_T = \frac{L}{v} = \frac{L}{\mu E} = \frac{L^2}{\mu V} \tag{4}$$

Provided that the light incident and hence the photon absorption are uniform over the length of the slab, Eq. (3) can be rewritten as

$$J = egL\left(\frac{\tau_n}{t_{Tn}} + \frac{\tau_p}{t_{Tp}}\right) = egLG \tag{5}$$

where egL is the total charge generation rate per cross-sectional area and is usually called the primary current density and

$$G = \frac{\tau_n}{t_{Tn}} + \frac{\tau_p}{t_{Tp}} \tag{6}$$

is called the gain of the photoconductor. The gain or current multiplication can be understood as follows. Ohmic

contacts or electrodes require that when a charge carrier leaves the solid at one contact, it is replenished at the opposite contact and maintains the neutrality of the slab charge. Thus each photogenerated charge carrier makes τ/t_T transits before it is lost, resulting in what is referred to as secondary current density. For devices that have very long lifetimes and very short electrode spacings, gains up to 10^6 are possible. The response or signal rise time of a photoconductor is determined by the transit time t_T. To achieve short transit times, small electrode spacings and high electrical fields are used.

In summary, the operation of a photoconductor requires charge generation by light and charge transport; hence the two key material properties of a photoconductor are the photogeneration efficiency η and the charge carrier mobility μ. Most single crystal semiconductors free of impurities and defects have intrinsic generation efficiencies close to unity and very high mobilities, for example, about $10^3\ cm^2\,V^{-1}\,sec^{-1}$ for Si and $10^4\ cm^2\,V^{-1}\,sec^{-1}$ for GaAs. The wavelength response spectrum of a photoconductor is determined by the absorption spectrum, which in turn is determined by the energy difference between occupied and empty energy levels. For an intrinsic photoconductor, this is the band structure of the material; the band gap defines the long wavelength cutoff, and the extremes of the valence and conduction bands determine the shortest wavelength. Extrinsic photoconductors are doped with impurities to produce energy states in the gap and thus enable photoconductivity at longer wavelengths. Because only one electron–hole pair can be produced per photon, the upper limit for η is unity and is usually less if some fraction of the charge pairs recombines or is trapped before the pairs are separated. However, for very energetic photons, for example, X-rays, multiple charge carriers can be produced as the hot excited charge carriers lose their energy by exciting additional charge.

PHOTODIODES AND PHOTOTRANSISTORS

The simple discrete, resistor-like photoconductor structure in Fig. 1 is used as a sensor, for example, in light meters or night detectors to switch on lighting. Typically these devices consist of a strip of cadmium sulfide (CdS) or selenide (CdSe) deposited on an insulating substrate that has electrical contacts. But the most widely used photoconducting detectors are photodiodes or phototransistors. They are almost exclusively planar sandwich structures that have transparent top electrodes to transmit the light. The advantages of planar structures are: first, that they can be produced in large quantities at low cost using photolithographic techniques and second, that they are amenable to large scale integration (LSI) techniques used by the semiconductor industry to produce imaging arrays that have a high density of discrete photodetectors for each pixel.

Materials

The semiconductor electronics industry uses single crystals of covalently bonded semiconductors of Si or germanium (Ge), III–V compounds such as gallium arsenide (GaAs) and indium antimony (InSb), or II–VI compounds such as CdS and mercury cadmium telluride ($Hg_xCd_{1-x}Te$) for photodetectors. The choice of material is governed primarily by wavelength requirements, which is determined by the band structure. Because Si LSI technology is highly advanced, imaging photodetectors that have the highest pixel density are based on Si. Because the band gap determines the long wavelength cutoff for intrinsic semiconductors, infrared photodetectors require low band-gap semiconductors such as InAs, PbSe, InSb, $Hg_xCd_{1-x}Te$ or $Pb_xSn_{1-x}Te$. (Varying the relative concentrations of Hg to Cd enables fine-tuning the cutoff wavelength from 4 to 30 μm and similarly for $Pb_xSn_{1-x}Te$.) These materials may be grown epitaxially on appropriate single crystal substrates. Large band-gap semiconductors such as Si can be made to photogenerate at longer wavelength by doping with appropriate impurities to introduce shallow energy states in the gap. For example, Si, which has a 1.2-μm wavelength cutoff, can be made sensitive to infrared light wavelengths between 2.5 and 7.5 μm by doping with In. By doping with different metals, the spectral response of Ge can cover the infrared range down to 100 μm. These long-wavelength intrinsic or extrinsic photodetectors must be operated at low temperatures because the dark conductivity would swamp any photocurrent. For details, see the sections on infrared detectors.

Photodiodes

Photodiodes rely on the electrical field produced by a p–n junction or Schottky barrier to separate photogenerated electron–hole pairs. The p–n photodiode consists of a p–n junction operated under reverse bias (Fig. 2). At the junction between the layers, the energy bands are bent to match the Fermi levels, resulting in a charge-depleted region and an electrical field across that region. Electron–hole pairs produced by photons absorbed in the depletion region are separated by the electrical field before they recombine and cause a current to flow in the circuit. For high-frequency operation, the thickness of the depletion layer must be kept small to reduce transit time. On the other hand, a thicker depletion layer is needed to allow a greater fraction of the light to be absorbed to increase the photogeneration yield. Thus there is a

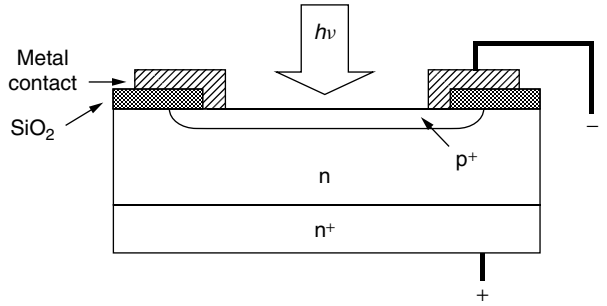


Figure 2. A typical reverse biased p–n photodiode. The p–n junction forms a charge depletion region in the n layer. Electron-hole pairs photogenerated in the depletion region are separated by the internal electrical field.

trade-off between speed and photogeneration efficiency. The heavily doped p region provides the front surface electrode, and the heavily doped n region at the bottom provides the other electrode.

The p–i–n photodiode (Fig. 3) is one of the most common photodetectors because the depletion region thickness, the intrinsic region, can be tailored to optimize the photogeneration efficiency and the frequency response. Usually, there is an antireflective overcoat on the surface such as SiO_2 to increase the photosensitivity. Otherwise, its operation is the same as the p–n diode.

Metal–semiconductor diodes (Fig. 4) rely on the Schottky effect to produce a region of charge depletion and electrical field in the semiconductor. The metal layer has to be thin, typically 10 nm, to be transparent, and an antireflective overcoat is required to minimize reflection losses. Because the depletion region is at the surface of the semiconductor, the metal–semiconductor is used particularly for ultraviolet and visible wavelengths, which are highly absorbed and do not penetrate far into the semiconductor.

In addition to operating in a photoconducting mode where the resistance is varied by the light as described before, photodiodes can also be operated in the photovoltaic mode, where the photodiode acts as a voltage source supplying an external current proportional to the photon flux.

Avalanche Photodiodes, Phototransistors, and CCDs

An avalanche photodiode is a special photodiode that has a sufficiently high reverse bias to cause charge-carrier multiplication by impact ionization. In addition to being very sensitive, these devices can operate at very high frequencies, up to microwave, because the transit times are very short. Another type of specialized photodiode is a heterojunction device formed by depositing a large band-gap semiconductor on a smaller band-gap semiconductor. The advantage of heterojunction devices lies in the fact that the large band-gap material is transparent to the light of interest, so there is little loss of light reaching the photogeneration region at the junction.

Bipolar-junction and field-effect transistors may also be used as light detectors. In a bipolar transistor, the supply voltage is applied between the collector and the emitter, and light creates charge carriers in the base region whose voltage is permitted to float to suit the current flow. Thus, the photon-produced current of microamperes controls the much larger collector-emitter current of milliamperes; the amplification is determined by the injection gain of the transistor. Figure 5 illustrates a typical n–p–n silicon phototransistor structure. Similarly, if an appropriately designed field-effect transistor has light focused on its gate-channel junction, the light will regulate the drain current, resulting in a photo-FET device.

Currently the most widely used photodetector imaging arrays are charge-coupled devices (CCD) (7). Although CCDs are discussed in detail in their own section, they deserve a comment here because they are distinct from the previous photodetectors and may be thought of as photocapacitors. These are MOS devices, metal–insulator–semiconductor, which function by using charge storage locations on the p-Si chip. By applying appropriate voltages to an array of electrodes separated by an insulating oxide layer from the p-Si, inversion regions or potential wells are created under the biased electrodes. In the simplest structure, every third electrodes forms a well that corresponds to a pixel. Photogenerated charges diffuse to the nearest potential well which collects and stores them on a timescale, the image frame time, that is long compared to the readout. The readout of the charge pattern is accomplished by changing the depth of the wells so that the charge spills to the adjacent (empty) well in a linear, bucket-brigade fashion that sends each pixel to the readout amplifier.

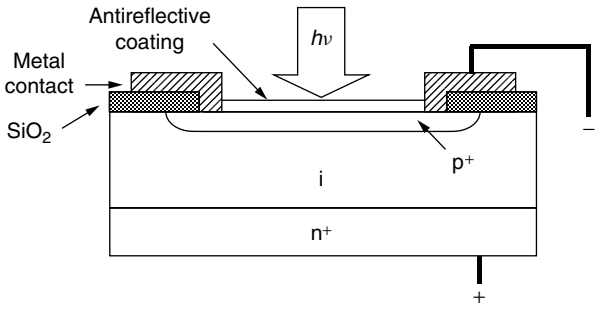


Figure 3. A p–i–n photodiode whose antireflective overcoat increases efficiency. The depletion region is the intrinsic region.

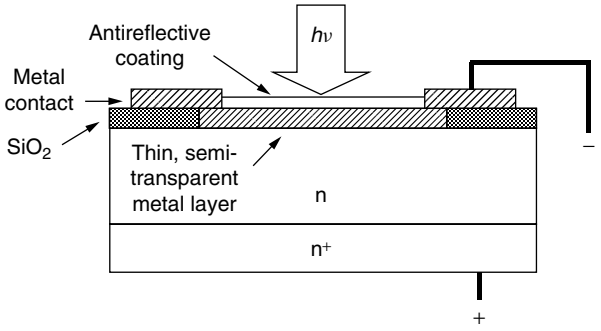


Figure 4. Structure of a metal–semiconductor photodiode.

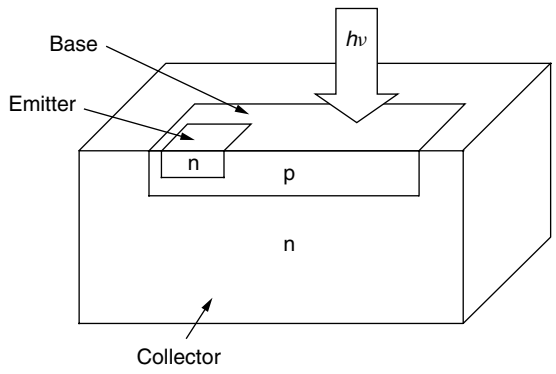


Figure 5. Schematic of an n–p–n phototransistor.

VIDEO CAMERA PHOTOCONDUCTORS

Although modern cameras use array imagers, primarily CCDs, since its beginning in the 1940s, the television industry relied on vidicon-type cameras throughout most of its history. Vidicon camera tubes (8) use electron beams to address and scan the pixels on a single photoconductor (Fig. 6). Typical vidicon tubes consist of a vacuum tube that has a transparent glass faceplate coated on the inside by a transparent conducting electrode of SnO_2 (NESA) and a photoconductor layer a few micrometers thick. A fixed positive voltage of 10 to 50 volts relative to the cathode of the electron beam gun is applied to the transparent electrode. The electron beam momentarily establishes a closed electrical circuit to successive pixels as the beam raster scans the photoconductor surface, driving the free surface voltage to zero by depositing electrons and charging the photoconductor surface. In the dark, the photoconducting layer is very insulating, so it acts as a capacitor to prevent any significant current flow on successive scans, once the surface voltage is driven to zero. Pixels exposed to light during each image frame, 1/30 second for television, discharge partially in proportion to the integrated light intensity over the frame time and neutralize some of the beam-deposited surface charge. The subsequent electron beam scan causes a current to flow in proportion to the charge required to recharge the surface back to zero volts. This small current signal is amplified to produce the video signal. Because the photodischarge occurs during a frame, charge-carrier transit time need not be very short. The original vidicon tube used vacuum-deposited amorphous selenium (a-Se) as the photoconductor (9). Because a-Se readily crystallizes and has sensitivity only to blue light, current vidicon tubes use antimony trisulfide (Sb_2S_3), which has a good spectral match to visible light. However, Sb_2S_3 has three drawbacks: high dark conductivity, image blooming, and image lag or ghosting. The high dark conductivity prevents its use in low light conditions because the signal-to-noise ratio becomes too small. Overexposure of pixels causes the excess charge to discharge neighboring pixels, resulting in rapidly growing white spots, called image blooming, that erase the surrounding image. Afterimage effects or ghosts are due to photoconductive lag. This "lag" arises because electrons are practically immobile in this p-type material and produce a space charge of trapped electrons. The increased electrical field resulting from the trapped charge enhances the photogeneration efficiency and dark current in subsequent frames. The Plumbicon tube improved on these problems by using PbO as the semiconductor with blocking electrodes (10). Basically, the photoconductor layer is a p–i–n structure where the n layer is next to the positive electrode and the p layer is at the free surface. Another improved design is the Saticon tube (11), which uses an arsenic (As)- and tellurium (Te)-doped a-Se photoconductor layer. a-Se is preferred because both electrons and holes are mobile and prevent electron trapping which leads to afterimage effects. Arsenic is added to prevent crystallization. Higher As concentration improves stability at the expense of electron lifetime, so a compromise value of 10% As is used. To extend the spectral response to red light, a thin layer adjacent to the faceplate is doped with Te. Finally, to trap the electrons from the e-beam at the surface, a thin layer of Sb_2S_3 is deposited on the free surface. Hydrogenated amorphous silicon a-Si(H) is another amorphous photoconductor that can be used in vidicon tubes. It matches the visible light spectrum very well and can be deposited as large area films needed in vidicon applications.

ELECTROPHOTOGRAPHIC PHOTORECEPTORS

The electrophotographic imaging industry, which produces of xerographic copiers and printers, has grown immensely since its invention in 1938 by Chester Carlson (12). In 2001, electrophotographic photoconductors will generate

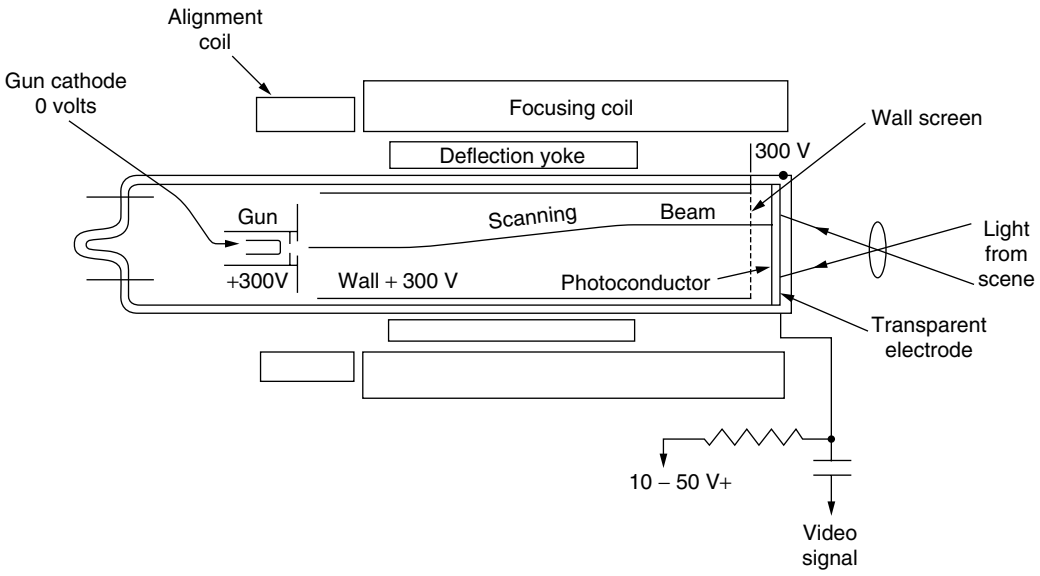


Figure 6. Vidicon tube camera structure. [After Weimer, Forgue, and Goodrich, *RCA Review* **12**, 309 (1951).]

revenue of about $9.1B. We are surrounded by documents that are either electrophotographic copies or prints produced by electrophotographic (laser) printers. Much as television (also invented in 1938) has become a ubiquitous medium of communication for the home, so electrophotographic documents have become an essential tool for informational communication in the office.

In its early history, the photoconductors used in electrophotographic imaging paralleled the materials and technical development of vidicon photoconductors. Like vidicon tubes they initially used films of amorphous chalcogenide glasses for photoconductors; one surface was connected to a conducting electrode, and the other surface was free and charged by a deposition of electrons or ions. Chalcogenide glasses are alloys or metal compounds of the chalcogen elements from column VI of the periodic table, that is, O, S, Se, Te, etc. Similarly, As was added for stability and Te for red light sensitivity, and layer structures that had thin Te-rich on Se layers were employed. However, the electrophotographic photoconductor, usually called the electrophotographic photoreceptor or simply photoreceptor, differs significantly in one key respect from all of the photoconductors discussed earlier. Whereas all the previously discussed photoconductor devices produce an electrical current, which results in a time-varying signal that is then electronically processed, the photoreceptor output is an electrostatic image that is developed. The photoreceptor is as much a display as a detector device, and in this respect it is similar to photographic film.

The photoreceptor converts an optical image or pattern into an electrostatic image, a full scale spatial pattern of charge or voltage corresponding to the desired final print. The electrostatic image is developed by using a fine powder of microscopic black or colored charged particles, called the toner. The resulting visible image on the photoreceptor is then transferred to paper or another substrate. Figure 7 is a schematic diagram of an electrophotographic copier or printer. The process consists of (1) charging with ions, (2a) or (2b) image light exposure, (3) development of the charge image by pigmented plastic powder, (4) electrostatic transfer to paper, (5) fusing of the powder image to paper, (6) cleaning, and (7) erasure of the photoreceptor. Light-lens copiers use reflected document light directly imaged on the photoreceptor, as shown in Fig. 7, step 2a. Modern printers and digital copiers employ digitally controlled light such as a scanning laser spot, 2b in Fig. 7, or an array of light-emitting diodes (LED). In this section, we are concerned only with the charging and light exposure processes that are essential to the operation of the photoreceptor. For details of the other processes, see the section on copiers and laser printers. The interested reader may also want to consult the very readable book, *The Anatomy of Xerography, Its Invention and Evolution*, by J. Mort (13).

In principle, very low dark conductivity, single-crystal semiconductors could be used as photoreceptors. But because the imaging is at a 1 : 1 scale and hence requires large surface areas (up to ten square feet) and continuous cylindrical or flexible belt geometries, the photoconducting materials used for photoreceptors have been amorphous or polycrystalline composites. The early commercialization of electrophotography was based entirely on inorganic photoconductors, primarily a-Se (14), but alloys of a-Se containing Te and/or As and arsenic triselenide (As_2Se_3) were also used. Other inorganic materials used were polymer dispersions of CdS and ZnO pigments (14), and beginning about 1980, amorphous (hydrogenated) silicon (a-Si : H) (15). The first commercial organic photoreceptors (16) did not appear until 1970, but by the mid-1980s, they surpassed inorganic photoreceptors in production volume. In the year 2000, more than 100 million photoreceptors were manufactured worldwide, of which more than 98% were of organic materials. Therefore, electrophotography is the largest commercial user of organic electro-optical devices. Before discussing organic materials, it is useful to review briefly photoreceptor structures and the physics underlying their operating functions.

Photoreceptor Structures

The electrophotographic photoreceptor consists of one or two layers of photoconductive material on a conductive substrate (Fig. 8). The single layer structures primarily used for inorganic photoconductors are charged positively, and the incident light is absorbed close to the surface.

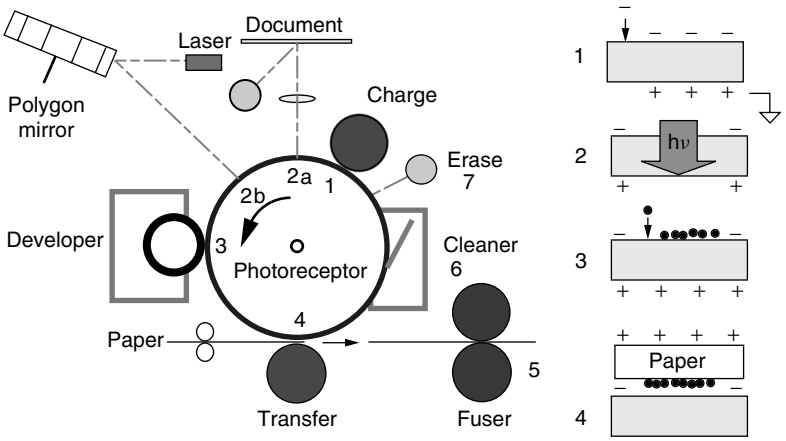


Figure 7. Schematic diagram of an electrophotographic copier or laser printer indicating the seven imaging steps. The copier uses a light lens imaging (2a), and the printer uses a laser-polygon mirror imaging system (2b). The diagram on the right schematically shows the photoreceptor during (1) charging, (2) photodischarge, (3) toner development, and (4) toner transfer to paper.

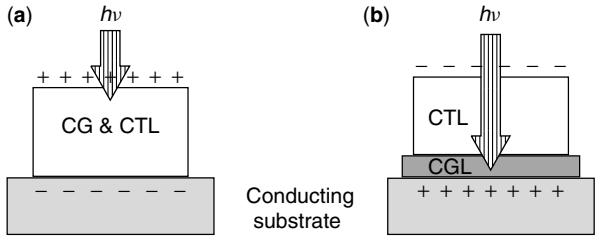


Figure 8. Photoreceptor structures of (**a**) single layer and (**b**) dual layer consisting of a transparent charge-transport layer (CTL) and a charge-generation layer (CGL). The plus and minus symbols indicate the polarity of the charge applied.

Organic photoreceptors are almost universally dual-layer structures consisting of a thin, 0.1–4 μm light absorbing layer, called the charge-generation layer (CGL), overcoated by a thicker, 10–40 μm transparent layer that transports the photogenerated charge carriers, called the charge-transport layer (CTL). These organic dual-layer structures are charged negatively. The polarity of charging for both single- and dual-layer structures is determined by the fact that they rely on the transport of holes, which originate at either the top or bottom surface. Additional layers or treatments may be used to provide adhesion to the substrate, to block charge injection from the metal substrate, to prevent interference fringes caused by coherent light reflection from the substrate, and to protect the free surface from wear. The substrates used are primarily rigid aluminum cylinders or flexible belts of a metal or metalized polymer. Aluminum cylinders have diameters that range from 16 mm to 280 mm and lengths of 240 mm to a meter, and belt diameters range from 60 mm to more than 3 meters.

Photoreceptor Properties

The operation of the photoreceptor involves two basic steps, charging and photodischarging. In turn, photodischarge consists of the two processes central to photoconductivity, photogeneration and charge transport. In the charging step, voltage is applied to the photoreceptor by externally depositing an ionic charge on the free surface. This is done by ionizing the air and depositing ions on the free surface by one of several methods. The simplest charging device, called a corotron, is a thin, 50–100 μm wire, surrounded on three sides by a conducting shield. A voltage of 4 to 8 kV applied to the wire ionizes the air, and a fraction of the ions flows to the photoreceptor surface, which is moving at a constant speed. The wire may be replaced by pins or a sawtooth air ionizer. In a device called a scorotron, a conducting screen, biased to the desired voltage level, is placed between the photoreceptor and the source of ions enabling more uniform voltage charging. Scorotrons are used in high-speed machines, but slower machines typically use a resistive charge roller that has applied dc and ac voltages. The air near the contact nip is ionized as the voltage difference between the roller and photoreceptor surface reaches air breakdown. The ac voltage provides polarity charging to ensure uniformity, and the dc level determines the final voltage on the photoreceptor (17). Conductive brushes are also used to deposit a charge. A surface charge σ on a photoreceptor of thickness L and dielectric permittivity ε results in a voltage

$$V = \frac{\sigma L}{\varepsilon}. \tag{7}$$

For example, an organic dielectric layer that has a thickness L of 25 micrometers and dielectric constant $\varepsilon/\varepsilon_0$ of 3 corresponds to a charging slope L/ε of 9.4 V nC^{-1} cm^2 or a capacitance ε/L of 107 picofarads cm^{-2}. A surface charge of 85 nC cm^{-2}, or about 0.53×10^{12} electrons/cm^2, produces a voltage of 800 volts and an internal electrical field of 32 V/μm on such a photoreceptor. Typical photoreceptor charge levels are 300 to 900 V and are determined by the requirements of the toner development process.

To maintain this charge in the dark requires that no charge be injected from the metal electrode or the surface and that the photoconducting layer(s) contains a negligible density of conducting charge carrier. The former requires that the free and metal surfaces be charge blocking. Additionally, to maintain a charge image, there can be no lateral conductivity on the surface requiring the surface to be charge trapping. In general, the photoreceptors can be charged to about 100 V/μm before breakdown occurs. No-bulk charge carriers imply that no significant charge is thermally generated on timescales of seconds to minutes. A density of N thermal emission sites, located an energy ΔE from the conducting states generates a charge,

$$n(t) = N(1 - e^{-Rt}), \tag{8}$$

where the thermal generation rate factor

$$R = \nu e^{-\Delta E/kT}, \tag{9}$$

where ν is the attempt to escape frequency, typically 10^{12}–10^{13} seconds^{-1}. Thermal emission states shallower than 0.5 eV empty in milliseconds, $Rt \gg 1$, and states deeper than 0.6 eV emit very slowly, $Rt \ll 1$. Thus if the photoconductor has N_S shallow states and N_D deep states, the thermally generated charge is

$$n(t) = N_S + N_D Rt \tag{10}$$

Because only holes are mobile in these materials, thermal generation results in a space charge of trapped electrons. For uniformly distributed thermal generation sites, the dark discharge is (18,19)

$$\Delta V = \frac{enL^2}{2\varepsilon}. \tag{11}$$

Thus the density of thermal emission sites is an important parameter and, for example, a dark voltage loss of less than 100 volts requires that there be fewer than 10^{14} per cubic centimeter shallow charge-generation sites, which corresponds to an atomic concentration of about 10 parts per billion.

Dark discharge can be enhanced by light overexposure or exposure to light when the photoreceptor is not charged,

producing extra charge that is not conducted to the surfaces. Because there is no electrical field to remove the generated charge, some fraction can become trapped, preventing recombination. When the photoreceptor is charged in the next image cycle, thermal emission of any remaining trapped charge carriers contributes to the dark discharge, a phenomenon called light fatigue or light-enhanced dark discharge. If this trapping occurs during the image exposure step, a ghost image is produced by the enhanced dark discharge pattern.

The discharge of the photoreceptor is accomplished by light exposure in a pattern of either a reflected image or a digitally generated pattern from a scanning laser spot or LED array. The exposure times range from a few milliseconds in copiers to nanoseconds in laser printers and therefore are very short compared to the time between light exposure and toner development. Typical hole mobilities for organic CTLs are 10^{-6}–10^{-5} $cm^2\,V^{-1}\,sec^{-1}$, so to discharge a 25-μm photoreceptor to 50 V requires a transit time, Eq. (4), of 12 to 100 ms. Thus a half second after exposure to light, the discharge is over, and the voltage for a photogeneration rate η is

$$V(X) = V_0 - SX \tag{12}$$

where V_0 is the dark voltage and X is the light exposure in energy units of erg cm^{-2} or mJ m^{-2}. The photosensitivity is

$$S = e\eta A \frac{L}{\varepsilon}\frac{\lambda}{hc} \tag{13}$$

usually expressed in units of V $erg^{-1}\,cm^2$ or V $mJ^{-1}\,m^2$. In the expression for photosensitivity, Eq. (13), e is the charge of an electron, A is the absorption efficiency at wavelength λ, and hc/λ is the energy per photon. The dielectric thickness L/ε is the capacitive amplification factor relating charge to voltage. Thus a 25-μm organic photoreceptor that has a photogeneration efficiency per incident photon ηA of 0.8 at a wavelength of 780 nm will have a sensitivity S of 470 V $mJ^{-1}\,m^2$ and require an exposure of 1.7 mJ m^{-2} to discharge 800 volts.

In practice, a linear photodischarge, as given by Eq. (12), dashed curve in Fig. 9, is not observed. The actual discharge tails off, Fig. 9 solid curve, because the photogeneration efficiency is field-dependent and decreases with field. Additionally, because some of the charge carriers do not transit the layer and are trapped, the discharge has a lower limit. Because nearly all of the photoconductors employ transport only holes, the light absorption and photogeneration take place near one or the other surface of the layer, and the holes must transit the layer L to discharge the photoreceptor. But because the electrical field moving the charge is due to the surface charge, the process of discharging collapses the field, slowing down the charge carriers. Although the leading edge of the moving charge carriers sees the full electrical field, each successive carrier moves more slowly as the field is reduced. Thus no matter how high the mobility or the initial field, the last charge will take an infinite time to transit (20). Eventually, the charge becomes trapped and for carrier lifetime

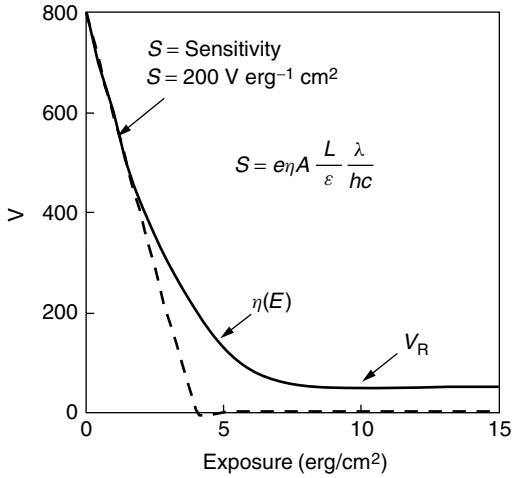


Figure 9. Photodischarge curves for an idealized constant-generation-efficiency η (dashed curve) and typical field-dependent η discharge (solid curve). V_R is a residual voltage due to trapped charge, and S is the sensitivity.

τ, a residual voltage V_R will result:

$$V_R = \frac{L^2}{2\mu\tau}. \tag{14}$$

The term $\mu\tau$ is called the charge-carrier range. Eq. (14) for residual voltage is applicable only to trapping that occurs uniformly in the CTL. In multilayer structures, trapping may occur at the interfaces and differently in each layer. Additionally, if photogeneration occurs at any distance from the positive surface or electrode, the motion of electrons becomes important, and the residual voltage may be due to their trapping. But in any case, the residual voltage due to a trapped charge sets a lower limit to the discharge (Fig. 9).

The factor that causes a nonlinear photodischarge is the field dependence of photogeneration. In any semiconductor, but especially in a molecular crystal or amorphous semiconductor where the mobility of the created charge is low, the photogenerated charge pair becomes self-trapped in their coulombic attraction, producing a bound charge pair called an exciton. The probability that these geminate charge pairs either recombine or separate depends on the ratio of their binding energy to thermal energy. The external electrical field enhances separation by supplying additional energy. Poole–Frenkel (21) and Onsager (22) models of this field-dependent photogeneration efficiency have been proposed (23,24), but the processes in typical layer photoreceptors are more complex. So empirical models are used for quantitative agreement, for example,

$$\eta = \frac{\eta_0}{1 + (E_C/E)^P}. \tag{15}$$

For most photogenerating materials, the power exponent P ranges between 0.5 and 2.5, and E_C is a constant indicating the electrical field where the generation efficiency is half its maximum value. At high electrical fields, $E > E_C$, the

generation efficiency saturates at η_0, that is, the discharge is linear at high voltage, and dV/dX is equal to S_0, Eq. (13) where $\eta = \eta_0$. At low field, $E < E_C$, the generation efficiency falls off as $(E/E_C)^P$, and dV/dX is proportional to $S_0(V/V_C)^P$ (the electrical field is equal to the voltage divided by the thickness L if there is no trapped space charge). Thus the tail of the discharge (Fig. 9) is due to field dependence, and E_C, along with P, is a measure of the "softness" of the photodischarge curve .

Photoreceptor Materials

The first commercial photoreceptors used a-Se as the photoconductor. The a-Se was vacuum-deposited to a thickness of 50 to 60 μm. a-Se has high sensitivity for blue light (Fig. 10) but is not a good match to the visible spectrum. Its key advantages are high hole and electron mobilities, 10^{-1} and 10^{-3} $cm^2\,V^{-1}\,sec^{-1}$, respectively, long lifetimes, and low thermal generation. Because a-Se crystallizes readily, about 0.5% of As is added to prevent crystallization. Te is added to enhance the red light response, but because a-Te crystallizes more readily than a-Se, the concentration of Te is kept to a few percent. To obtain more red light sensitivity, a higher concentration is deposited as a thin, typically a fraction of a micrometer, photogeneration layer, much like in the Saticon photoconductor (10). The other chalcogenide used extensively as a photoreceptor is a-As_2Se_3, which has the advantages of being very sensitive, a good match to visible light (Fig. 10), and is very durable. The main disadvantages of a-As_2Se_3 photoreceptors are the lack of electron transport, low hole mobility, 10^{-5} $cm^2\,V^{-1}\,sec^{-1}$, and significant thermal generation.

The discovery of hydrogenated amorphous silicon a-Si(H), formed by the chemical vapor deposition of decomposed silane (SiH_4) (25), led to its use as a photoreceptor in the 1980s. Normal a-Si has too many broken or dangling bonds that give rise to a high density of traps which result in poor charge transport and high thermal generation. By incorporating hydrogen to eliminate these dangling bonds, an electronic grade material is obtained. By doping with phosphorus or boron atoms, much like crystalline Si, p- or n-type material can be made that has hole mobilities of 10^{-1} $cm^2\,V^{-1}\,sec^{-1}$ or electron mobilities of about 10^{-3} $cm^2\,V^{-1}\,sec^{-1}$. In other respects, the a-Si(H) photoconductive properties are similar to those of a-As_2Se_3 (Fig. 10). The main advantage of a-Si(H) is that it is very hard and durable, enabling print lives of several million prints. Its main disadvantage is the slow vacuum-deposition process that makes the photoreceptor expensive. Due to the slow deposition rate, the typical layer thickness is kept down to about 20 μm. Other inorganic photoconducting materials that have been used are solvent-coated polymer dispersions of ZnO and CdS. The main difficulty with these material dispersions is that they have trap-dominated charge transport that makes them cyclically unstable due to buildup of residual voltage and enhanced dark discharge.

Because solvent coating of polymer films can be significantly cheaper than any vacuum-deposition process, organic photoreceptors that could be solvent coated were of considerable interest in the 1960s and 1970s. The earliest practical organic photoreceptor material to be used was poly-*N*-vinylcarbazole (PVK) (26) (Fig. 11), which exhibited hole transport. Although PVK has limited photosensitivity because it absorbs light only in the blue-wavelength spectrum, studies of charge transport in PVK and other organics identified the basic mechanism of hole transport as electron exchange between adjacent electron donor molecules (Fig. 12a), which are aryl amine units.

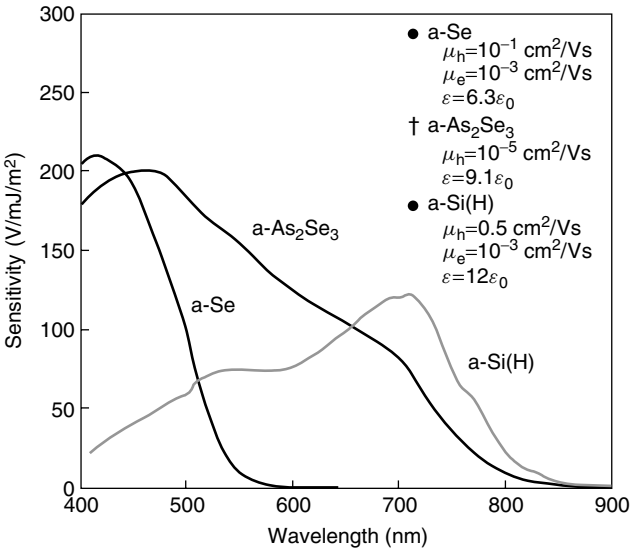


Figure 10. Spectral sensitivities of a-Se, a-As_2Se_3, and a-Si(H) photoreceptors. The insert indicates the corresponding hole and electron mobilities and dielectric constants.

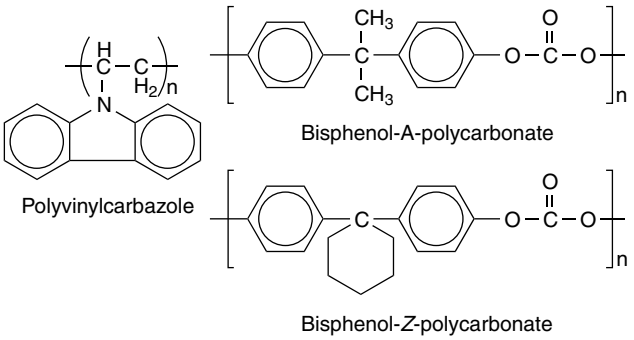


Figure 11. Monomeric units of polyvinylcarbazole (PVK) and two polycarbonate polymers used in organic photoreceptors.

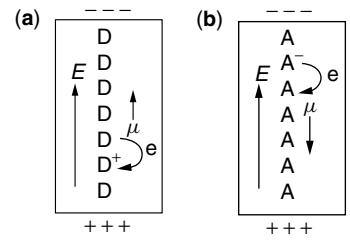


Figure 12. Schematic of (**a**) hole and (**b**) electron transport in an organic charge-transport layer. The charge moves by electron exchange between adjacent electron-donor (D) or electron-acceptor (A) molecules.

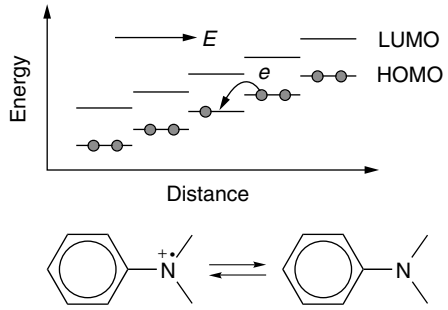


Figure 13. Energy diagram of electrical field-assisted hole transport and diagram of the electron transfer process. Electron-donor molecules are aromatic amines that can reversibly exchange an electron between the neutral molecule and its oxidized cation.

The process involves a reversible one-electron reduction-oxidation phenomenon, as depicted in Fig. 13. The hole charge carrier is an oxidized aryl amine (cationic radical), and the hole moves when an adjacent neutral aryl amine unit transfers an electron. The electrical field imposes a preferred direction by lowering the energy and increasing the exchange rate, as indicated schematically in Fig. 13. Thus, hole transport is a hopping process that involves electron exchange between electron-donor molecules. Organic charge-transport layers consist of solid-state solutions of 30–40% aryl amines in a polymer, called a doped polymer. Typical polymers are polycarbonates whose molecular weights range from 30 to 150 thousand units (Fig. 11). Charge transport in molecular doped polymers has been studied extensively, and the underlying physics is well understood and documented in an extensive literature. For details and references, the interested reader is referred to *Organic Photoreceptors in Xerography* by Borsenberger and Weiss (6).

Although an endless variety of aryl amines is possible (6), the two most widely used classes are hydrazones and triphenylamine derivatives (Fig. 14). Hydrazones are molecules characterized by the C=N–N structure. They are transparent to light wavelengths down to about 450 nm where they absorb strongly, giving them a yellow color. Because they absorb blue–UV light, they photodegrade when exposed to blue light. Triphenylamines, that have three aromatic groups on the nitrogen have the highest hole mobilities as a class and are transparent to wavelengths less than 400 nm, depending on the side groups. Because the latter are more stable and have higher mobility, hydrazones are being phased out in favor of triphenylamines. Analogous to hole transport by electron-donor molecules, electron-acceptor molecules are required for electron transport (Fig. 12b). Although a number of acceptor molecules that enable electron transport have been reported (6), to date none has met the mobility and range requirement of commercial CTLs.

Molecular polymer solutions make practical charge-transport layers, but photogeneration requires crystalline photoconductors. Solvent coating requires that these photoconductors be ground into fine particles, called pigment particles, and dispersed in a polymer binder. The process of grinding can introduce defects and impurities that act as traps or thermal generation centers, hence the amount of pigment must be minimized. Without electron transport, a dilute single layer dispersion that absorbs the light and photogenerates uniformly in the bulk is impractical. Instead, a separate CGL layer is used, thick enough to absorb the light but thin enough to permit electron removal. Because the pigment extinction coefficient is about $10^5\ cm^{-1}$, CGLs need be only a few tenths of a micrometer thick and have pigment concentrations of 40 to 80% in the polymer binder. The pigment particles must be less than 0.1 μm, typically of the order of 10 nm. The pigments themselves are usually electron transporters, so the high concentration permits particle contact, enabling electron transport. Some stable pigments such as perylenes or phthalocyanines can also be vacuum-deposited to form very uniform thin films without a binder. Because the thin CGLs are fragile, it is preferable to coat them on the metal surface and overcoat using the thicker CTL (Fig. 8). In this case, the photoreceptor must be charged negatively. If, instead, the CGL is coated on top of the CTL, the photoreceptor is charged positively, and a protective overcoat may be required. Considerable progress has been made in understanding charge transport in molecularly doped polymers, but fundamental understanding of photogeneration has been limited by the complexity of the practical CGLs. Because the generation layers are dispersions of very small polycrystalline pigments in a polymer, the generation efficiency depends on the initial generation of charge in the crystal and also on the transfer of the charge from the crystal to the charge-transporting medium, that is, the CTL for holes and through the CGL for electrons. On the one hand, the charge-transfer process that takes place on the surface of actual pigment particles is difficult to characterize. On the other hand, model single crystals of the pigment employed are too difficult to obtain. As a result, the design of CGLs is highly empirical, and the technology is treated as proprietary. So the literature on CGL design is limited.

A very large variety of organic pigments is used in photoreceptors, including azo pigments (27), dyes such as thiopyrilium (28) and squarilium (29), polyaromatic hydrocarbons and amines, and phthalocyanines (30). Figure 15 is an example of the spectral sensitivities of some typical organic photoreceptors. In general, the pigments are classified as either visible or "laser" pigments. The former have photogeneration efficiencies tailored to the visible spectrum, 400 to 700 nm. The latter are tailored for monochromatic LED (670, 680, 720 nm) or laser (630, 670, 780 nm) wavelength used in digital machines. The wavelength for most laser printers is 780 nm due to the wide use of diode lasers developed for the compact disk (CD) industry.

The two most widely used visible pigments are azos (27), Fig. 16, and perylenes (31,32), Fig. 17. Azo pigments, identified by the double nitrogen bond –N=N–, are widely used as color pigments and dyes for fabrics and inks and have a well-established chemical industry to supply them. Bisazos consist of a C–N=N–A–N=N–C structure, which offers unlimited possibilities through variations of possible central bridging units A and coupling

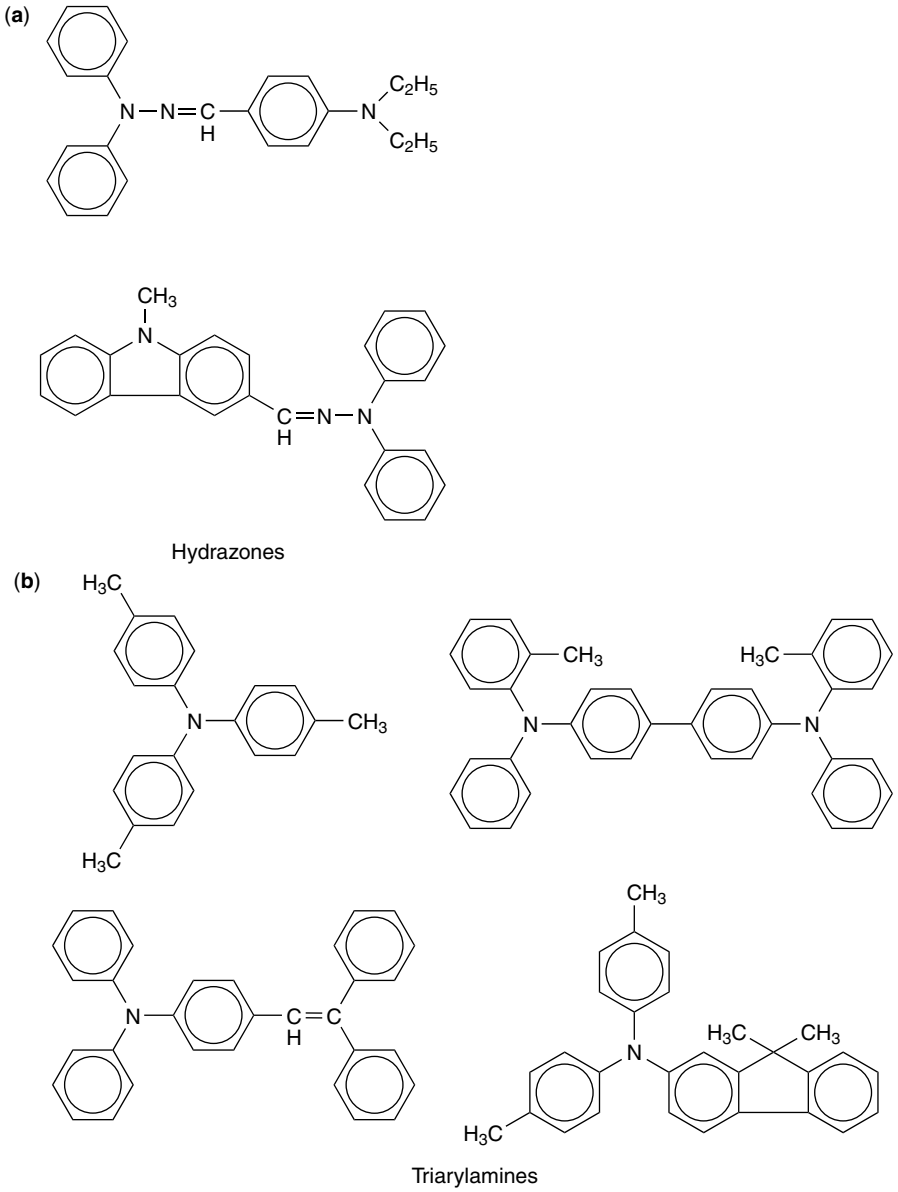


Figure 14. Examples of (**a**) hydrazone and (**b**) triarylamine electron-donor molecules employed in charge-transport layers.

units C (Fig. 16). The very fine particles readily obtained directly from synthesis make excellent dispersions. Bisazos are primarily visible absorbers, but near-IR pigments are also in use. Of the polyaromatic or polycyclic pigments, perylene imides are the best known because of their use in automotive paints and colorants for plastics. Their use in automobile paint attests to their stability to exposure to sunlight and the elements. The perylene pigments of interest as photogenerators fall into two classes, perylene diimides and perylene arylimidazoles (Fig. 17). Because of the stability of perylene pigments, thin CGLs can be prepared by direct vacuum deposition as well as by solvent coating of dispersions (31,32).

Phthalocyanines are the pigments most often used for laser and LED light sources. Although phthalocyanine photoconducting properties have been studied for some time and their use as an organic photoreceptor goes back to the 1960s (30), only after the appearance of diode lasers in the 1980s, were they introduced as pigments for laser printers. Phthalocyanines (Fig. 18) are metal complexes of porphyrin molecules that are related to natural photoactive pigments, such as chlorophyll. The molecules are flat, planar structures that are divalent metals, metal halides, or metal oxides at the center (33). They absorb light strongly in the visible red and near-IR, are extremely stable to decomposition, and are insoluble in common solvents. As for perylenes, their stability is exploited by the automotive paint industry for blue paints that will not degrade under full sun exposure.

Because of their flat pancake shape, phthalocyanine pigments exhibit many different crystal polymorphs, distinguishable by X-ray diffraction spectra, which influence their photogeneration efficiency (6,33). Figure 15 shows the sensitivity of two photoreceptors based on titanyl phthalocyanine (TiOPc) of different

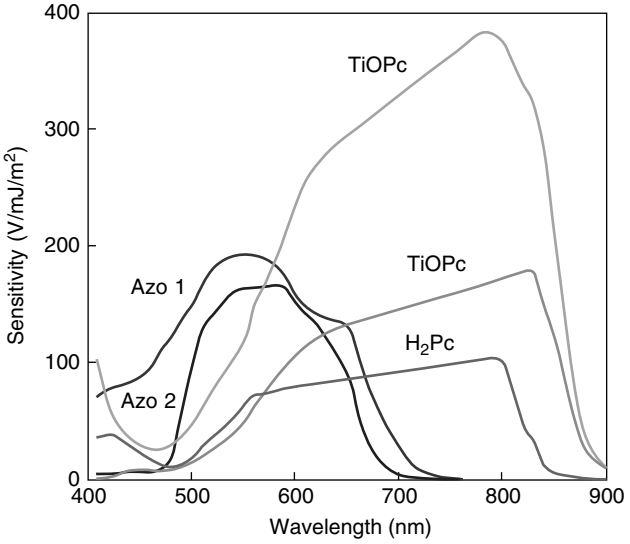


Figure 15. Spectral sensitivity of two visible (azo pigments) and three laser diode (phthalocyanine, Pc pigments) sensitive organic photoreceptors. The two TiOPc sensitivities are from a high- and medium-sensitivity polymorph.

crystal forms. The more stable polymorphs usually have lower photogeneration than the more sensitive polymorphs, which are obtained by techniques of milling, heating, solvent treating, or controled substrate temperature during vacuum deposition. Similarly to perylenes, phthalocyanine CGLs are made by either solvent coating from a binder dispersion or by vacuum deposition.

Resolution and Durability

Two critical parameters of an imaging device are its resolution and how many times it can be cycled before wearing out. The resolution of a photoreceptor is governed first by the surface resistivity and second by coulombic blooming caused by the collapsing electrical field as the photogenerated charge moves through the CTL. Normal resistivity is greater than 10^{16} ohm/cm^2, which is sufficient for resolution down to a few micrometers. Because the surface is exposed to ions and corona effluents such as ozone and NO_X during charging and to residues from toner development and air contaminants, film deposits of these contaminants on the surface can decrease the resistivity. For example, any salts that ionize in the presence of moisture will increase the surface conductivity. Maintaining high resolution requires aggressive cleaning that removes any film and the monolayer of the chemically oxidized surface. Because charge blooming varies with the thickness of the CTL, the limiting resolution is the thickness of the CTL for typical 70–90% discharge operating conditions.

Chemically, the best organic materials are very stable, and lifetimes of millions of cycles are possible. The occurrence of chemical degradation or instabilities is due to impurities introduced in the manufacturing process, so the remedy lies with improved processing. The primary photoreceptor failure modes are extrinsic damage, such as scratching, and ultimately, wear. Catastrophic damage failures are a function of machine reliability and design, but once those are addressed, wear is the limiting factor. The photoreceptor is subject to physical contact during

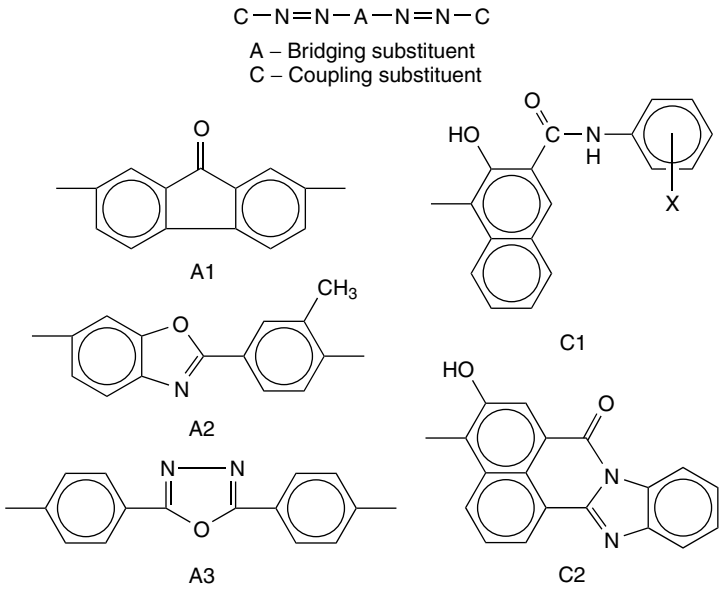


Figure 16. Examples of bisazo pigment molecular structures.

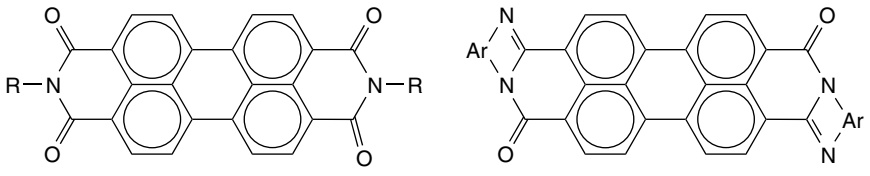
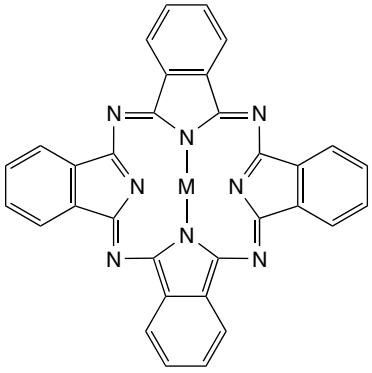


Figure 17. Perylene bisimide and bisarylimidazole pigment molecules.



Divalent M: H_2, Mg, Zn, Cu, Pb, Cd, Sn
Trivalent M: AlX, InX, GaX
Tetravalent M: SiX_2, GeX_2, VO, TiO
X = F, Cl, Br, OH

Figure 18. Phthalocyanine pigment molecules.

toner development, paper transfer, and cleaning (also charging if charge rollers are used), all of which contribute to abrasive wear. Wear rates of organic CTLs range from 10–60 nm per kilocycle depending on machine design, and the typical value is 20 nm per kilocycle, which results in a life of 300 to 600 kcycles. For comparison, the much harder a-Si(H), wears at 0.1–1 nm per kcycle and lasts about 3 to 10 million cycles. a-Si(H) is actually more sensitive to surface scratches and is overcoated by an oxygen- or carbon-doped thin layer which wears through. To increase photoreceptor wear life, CTL thickness has been increased from less than 20 μm to about 30 μm. Overcoated and reinforced photoreceptor designs are being introduced along with improved machine designs, and in 2000, organic photoreceptors that have a print life of more than one million cycles were introduced.

SUMMARY

Photoconductors used in imaging fall into two classes, those based on discrete units for each image pixel and large surface area imagers. The former use crystalline IV, III–V and II–VI semiconductors and integrated chip technologies, especially the well-established silicon VLSI technology, and are used in a wide spectrum of imaging applications, as well as in the communication and electronic industries. The large surface area devices use amorphous or molecular photoconductors that can be fabricated in large continuous areas, often measured in square feet. The primary application of large area continuous photoconductors continues to be the electrophotographic industry, which, almost exclusively, employs organic materials. But recent advances in organic thin film transistors and organic light-emitting diodes (OLEDs), essentially photodiodes operated in reverse, promise to expand the use of organic electro-optic materials considerably.

Photoconductivity involves the excitation of electrons from occupied energy states to empty states producing a mobile charge that is separated by an externally applied or internal electrical field. The key parameters are photogeneration and transport of charge carriers, measured by the photogeneration efficiency η and mobility μ, which determine the sensitivity and speed of the photodetector. The best materials are single crystals that have near unity photogeneration efficiencies and high mobilities. Choosing materials with different band gaps enables solid-state photodetectors that respond from 100 μm in the far infrared to X rays. Single-crystal photodetectors are fabricated, with few exceptions, in planar geometries in four basic structures: photoresistors, photodiodes, phototransistors, and CCDs. The basic photodiode and phototransistor structures come in many different varieties optimized for special applications, such as high sensitivity and low noise or high speed, up to microwave frequencies for optical communication.

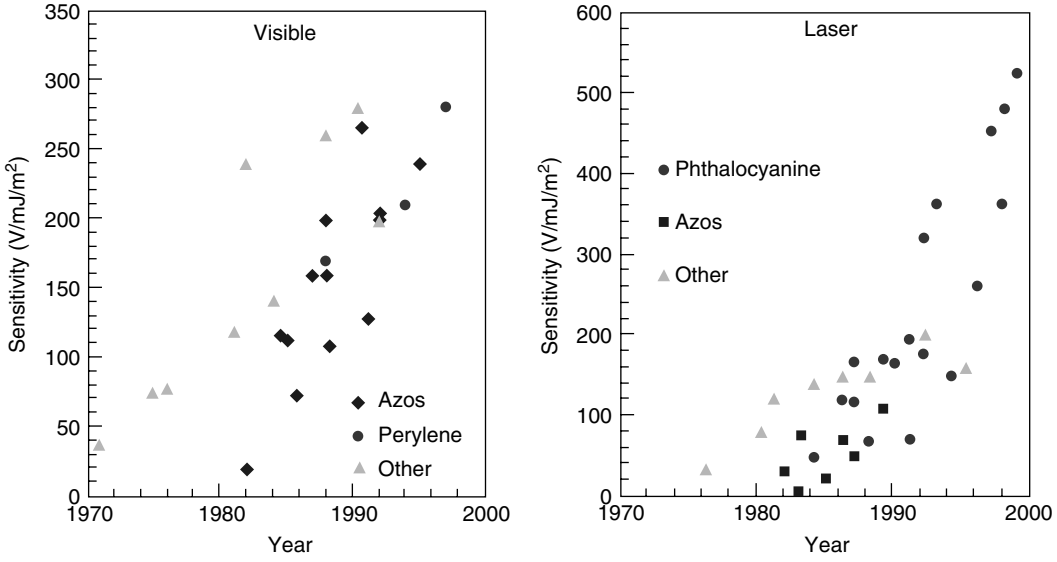


Figure 19. Sensitivity improvements from 1970 to 2000 of visible and laser-sensitive (primarily 780-nm) organic photoreceptors.

Single crystals are impractical for large area photoconductors such as photoreceptors, so they must rely on amorphous or disordered solids that have significantly lower mobilities. In modern organic photoreceptors, separate electronic material layers accomplish these functions, that is, photogeneration by organic crystalline particles and hole transport by aromatic-amine electron-donor molecules. A large variety of electron-donor molecules are in current use; the two main classes are hydrazone-based arylamines and triphenyl-based polyamines. At concentrations of 40%, typical mobilities of 10^{-5} $cm^2\,V^{-1}\,sec^{-1}$ are achieved, which is sufficient for most electrophotographic applications.

Considerable progress has been made in the last 30 years in the photosensitivity of organic photoreceptors (Fig. 19), and sensitivities approach the theoretical limit of unity photogeneration. The best materials are chemically very stable, and intrinsic lifetimes of millions of cycles are possible, but organic photoreceptors are soft, and wear limits the life to about half a million cycles, compared to a-Si(H) that has longer wear life by a factor of 10. Protective overcoats and reinforced polymers are being pursued to extend the life beyond a million cycles and approach that of a-Si. Typical two-layer designs in current use are limited in resolution to the thickness of the CTL, about 1,200 dpi (dots per inch). High quality color will require higher resolution and photoreceptors that have top surface photogeneration.

ABBREVIATIONS AND ACRONYMS

MOS	metal-oxide-semiconductor
CCD	charge coupled devices
CGL	charge generation layer
CTL	charge transport layer
LED	light emitting diode
CD	compact disk

BIBLIOGRAPHY

1. W. Smith, *Nature (London)* **7**, 303 (1873).
2. F. C. Nix, *Rev. Mod. Phys.* **4**, 723 (1932); A. L. Hughes, *Rev. Mod. Phys.* **8**, 294 (1936).
3. R. H. Bube, *Photoconductivity in Solids*, John Wiley & Sons, Inc., New York, NY, 1960.
4. A. Rose, *Photoconductivity and Related Processes*, Interscience, NY, 1963.
5. S. M. Ryvkin, *Photoelectric Effects in Semiconductors*, Consultants Bureau, NY, 1976.
6. P. M. Borsenberger and D. S. Weiss, *Organic Photoreceptors for Xerography*, Marcel Dekker, NY, 1998.
7. M. J. Howes and D. V. Morgan, eds., *Charge-Coupled Devices and Systems*, John Wiley & Sons, Inc., New York, NY, 1979.
8. P. K. Weimer, S. V. Forgue, and R. R. Goodrich, *RCA Rev.* **12**, 309 (1951).
9. P. K. Weimer and A. D. Cope, *RCA Rev.* **12**, 314 (1951).
10. E. F. deHaan, A. van der Drift, and P. R. M. Schampers, *Phillips Tech. Rev.* **25**, 133 (1963/4).
11. N. Goto, Y. Isozaki, K. Shidara, E. Maruyama, T. Hirai, and T. Fujita, *IEEE Trans. Electron Devices* **ED-21**, 662–666 (1974).
12. *Electrophotography*, U.S. Pat. 2297691, 1942, C. F. Carlson.
13. J. Mort, *The Anatomy of Xerography, Its Invention and Evolution*, McFarland, Jefferson, NC, 1989.
14. J. H. Dessauer and H. E. Clark, eds., *Electrophotography and Related Processes*, Focal Press, London and New York, 1965.
15. I. Shimizu, T. Komatsu, T. Saito, and E. Inoue, *J. Non-Cryst. Solids* **35**, 773 (1980).
16. W. D. Gill, *J. Appl. Phys.* **43**, 5,033 (1973).
17. H. Kawamoto and H. Saito, *J. Imaging Sci. Technol.* **38**, 383 (1994).
18. A. R. Melnyk, *J. Non-Cryst. Solids* **35**, 837 (1980).
19. M. Abkowitz, F. Jansen, and A. R. Melnyk, *Philos. Mag. B* **51**, 405 (1985).
20. I. P. Batra, K. K. Kanazawa, and H. Seki, *J. Appl. Phys.* **41**, 3,416 (1970).
21. J. Frenkel, *Phys. Rev.* **54**, 647 (1938).
22. L. Onsager, *J. Chem. Phys.* **B**, 599 (1934).
    L. Onsager, *Phys. Rev.* **54**, 554 (1938).
23. R. H. Batt, C. L. Braun, and J. F. Horning, *J. Chem. Phys.* **49**, 1,967 (1968).
24. D. M. Pai and S. W. Ing, *Phys. Rev.* **173**, 729 (1968).
25. W. E. Spear and P. G. LeComber, *Philos. Mag. B* **33**, 935 (1976).
26. H. Hoegl, *J. Phys. Chem. Solids* **25**, 1,221 (1964).
27. K. C. Nguyen and D. S. Weiss, *Electrophotography* **27**, 2 (1988).
28. W. D. Dulmage et al., *J. Appl. Phys.* **49**, 5,543 (1978); P. M. Borsenberger, A. Chowdry, D. C. Hosterey, and W. Mey, *J. Appl. Phys.* **49**, 5,555 (1978).
29. R. E. Wingard, *IEEE Ind. Appl.* 1,251 (1982).
30. J. Weigl et al., in W. Berg and K. Hauffe, eds., *Current Problems in Electrophotography*, Walter de Gruyter, Berlin, 1972, p. 287.
31. E. G. Schlosser, *J. Appl. Photogr. Eng.* **21**, 118 (1978).
32. J. Duff, A. -M. Hor, A. R. Melnyk, and D. Teney, *J. SPIE Proc.* **1,253**, 183 (1989).
33. R. O. Loutfy, A. -M. Hor, C. K. Hsiao, and A. R. Melnyk, *Proc. Fourth Int. Congress Non-lmpact Printing Technol.* SPSE, New Orleans, 1988.

PHOTODETECTORS

MARK WADSWORTH
Jet Propulsion Laboratory
Pasadena, CA

Photons that impinge upon matter interact in a manner determined by the nature of the chemical bonds in the material and the energy of the incident photons. Interacting photons may be reflected, refracted, diffracted, transmitted, or absorbed. Each of these phenomena can be used to measure some parameter of interest in chemical analysis. Photoassisted chemical analytical techniques require an accurate, sensitive method of photon detection and quantification. Photographic film proved to be the first useful method for such evaluations. Later, photomultiplier tubes provided the means for improving many photoanalytical techniques. As a result of the advent of semiconductor technology, numerous single-element, broad-spectrum

photodetecting devices were developed and applied to the task. Most recently, photodetector technology has entered the age of high-density integration. Large-area photon detectors are becoming commonplace in most analytical equipment. The advances afforded by these devices have allowed significant improvements in the performance of systems designed for photoanalytical chemical analysis in general, and specifically for spectroscopy. Therefore, a working knowledge of the operation and limitations of photodetectors is necessary for the modern chemist.

Photodetector devices convert electromagnetic radiation or photons to electric signals which can be processed to obtain the spectral, spatial, and temporal information inherent in the radiation. Photodetectors, as indicated in Table 1, may be operated in many modes. The more popular are photoconductors, photodiodes, charge-transfer devices, and pyroelectrics. Detectors may be used as single elements such as in street light controls, film camera exposure control, or motion detectors for security. Photodetectors also find application as linear arrays used in analytical spectrometers, night-vision equipment, in small, low cost spectrometers to control building ventilation and environmental pollution monitoring, or configured as large matrix arrays in video cameras.

PRINCIPLES

Photon Absorption and Thermal Emission

The basic detection process, (Fig. 1) is the generation of free electrons, holes, or both by the absorption of photon energy (1–4). The absorption may be intrinsic, that is, creating a free electron–hole pair, or extrinsic, creating a free hole or electron. The process can be indirect, whereby the absorbed photon energy raises the lattice temperature (thermal detection) and a phonon generates the free charged particle. The change in free charge is sensed in an amplifier circuit as a signal voltage or current. Random generation of free charge is the source of detector noise. The detector geometry, spectral response, electrical bias, and temperature are adjusted to optimize the signal-to-noise ratio. The absorption efficiency (photon to electron conversion) is a critical issue. The detector surface is typically coated to minimize reflections at a particular wavelength, and the internal efficiency is given by $1 - \exp(-\alpha x)$. Because the absorption coefficient α is typically 3,000 cm^{-1} for intrinsic detection and 3 cm^{-1} for extrinsic detection, detector thickness x has a wide range of values.

Important figures of merit (5) describe the performance of a photodetector: responsivity, noise, noise equivalent power, detectivity, and response time (2,6). However, there are several related parameters of measurement: temperature of operation, bias power, spectral response, background photon flux, noise spectra, impedance, and linearity. Operational concerns include detector element size, uniformity of response, array density, reliability, cooling time, radiation tolerance, vibration and shock resistance, shelf life, availability of arrays, and cost.

Photodetectors exhibit well-defined, cutoff-wavelength thresholds, whose positions are determined by the magnitudes of the band gap activation energy E_g or impurity-activation energy E_i. The cutoff wavelength λ_c, corresponds to a photochemical activation energy E, where

$$\lambda_c = hc/E. \tag{1}$$

The product of Planck's constant and the velocity of light $hc = 1.24$ when E is expressed in electron volts and λ_c in μm. Because the absorption coefficient is typically a weak function of wavelength above the threshold energy, the spectral response approximates that of the ideal quantum counter. The interest in semiconductor photodetectors with cutoff wavelengths longer than 3 μm results from their extreme sensitivity to variations in photon flux emitted from gray-body objects that are characterized by time and spatial variations in temperature and emissivity. Although some applications require detecting narrowband radiation, for example, atomic emission spectra and radiation from hot gases, most uses of infrared detectors involve detecting variations in photon flux from objects near ambient temperature. Typically, a lens system images the scene of interest on a single detector element, a linear array of detector elements, or an area array (7,8). The infrared detector is exposed to the signal photon flux through the optics. The ambient background photon flux ϕ_B, is often much greater than the signal flux and determines the ultimate system performance. The background portion ϕ_B, in photons/($cm^2 \cdot s$) may be calculated from Planck's radiation law.

Blackbody Emittance

Representative blackbody emittance (9,10), calculated as a power spectral density, is shown in Fig. 2. The wavelength λ of peak power density for a blackbody at temperature T is given by Wien's displacement law

$$\lambda_m T = 2{,}898\ \mu m\ K, \tag{2}$$

where λ_m is the wavelength of maximum power density and T is temperature in K. The peak spectral power density or emittance in W/($cm^2 \cdot \mu m$) is given by

$$M(T)_m = 1.288 \times 10^{-15}\ T^5\ \text{watts}/(cm^2 \mu m) \tag{3}$$

for each square centimeter of emitting surface and micrometer of wavelength at the peak value. The total power density W integrated over all wavelengths in W/cm^2 is given by the Stefan–Boltzmann law:

$$W = 5.67 \times 10^{-12}\ T^4\ \text{watts}/cm^2, \tag{4}$$

which expresses the total rate of emission in a 2π steradian solid angle from an element area on a blackbody surface. The very strong dependence of W on temperature and the sharp decline of emittance for wavelengths somewhat less than the peak wavelength results in very low ambient background flux density for ambient temperatures less than 320 K and detectors whose cutoff wavelengths are <3 μm. Short-wavelength detectors such as silicon charge-coupled devices (CCDs) do not work in the infrared for this reason but are excellent when hot sources such as light bulbs (1,500 K) and the sun (6,000 K) are available.

Table 1. Semiconductor Photodetectors in General Use

Bandwidth μm	Mode[a,b]	Temp.(K)	Material	Pixel Size (μm)	Focal Plane Size	Principal Use	Cutoff Wavelength (μm)	Response Time(μs)	Responsivity (V/W)	D* (cm $Hz^{1/2}$/W)
					Ultraviolet					
0.200 –0.400	CCD	300	Si	10–30	2,000 × 4,000	Astronomy	1.1		1×10^{11}	5×10^{13}
0.400–0.700[c]	CCD	300	Si	10–30	488 × 640–2,000 × 2,000	Video Camera	1.1		2×10^{10}	5×10^{13}
	PC	300	CdS	100–1,000	<100	Light meter	0.7	100,000	1×10^{6}	1×10^{13}
	PV	300	Si	30–200	Few	Optical switch	1.0	0.01–10	0.5[d]	5×10^{13}
	PV	300	GaAsP	30–200	Few	Optical switch	0.8	0.1–10	0.2[d]	1×10^{13}
					Short-wavelength infrared					
1 –3	PV	240	InGaAs	25–200	1 × 512	Spectroscopy	2.5	1	0.2[d]	2×10^{12}
	PV	200–300	InGaAs	25–50	512 × 512	Imaging	1.7	1	0.2[d]	2×10^{12}
	PV	200–300	HgCdTe	40–400	256 × 256–1,000 × 1,000	Astronomy; Mapping	2.9	10	2.0[d]	2×10^{12}
	PC	200–300	PbS	40–400	10 × 3,000	Mapping	2.7–3.2	300–1,000	5×10^{4}–1×10^{3}	5×10^{11}–1×10^{11}
					Mid-wavelength infrared					
3–5	PC	150–220	HgCdTe	40–60	1 × 100	Night scope; heat sensing	5.0	10	1×10^{5}	1×10^{11}
	PC	77–200	PbSe	50–80	1 × 100	Heat sensing	5.8[e]	150[e]	1×10^{6}[e]	2×10^{10}
	PC[f]	60–90	Ge : Au	80–200	Few	Fast response	7.8	0.1	1×10^{4}	1×10^{10}
	PV	60–90	InSb	40–60	512 × 512	Thermal imaging	5.25	1	3.4[d]	1×10^{11}
	PV	120	HgCdTe	40–60	512 × 512	Thermal imaging	5.0	5	3.4[d]	1×10^{11}
	SB[g]	60–80	Si : Pt	40–60	480 × 640	Thermal imaging	4.6	1	0.03[d]	3×10^{9}

Table 1. (*Continued*)

Bandwidth μm	Mode[a,b]	Temp.(K)	Material	Pixel Size (μm)	Focal Plane Size	Principal Use	Cutoff Wavelength (μm)	Response Time(μs)	Responsivity (V/W)	D* (cm $Hz^{1/2}$/W)
					Long-wavelength infrared					
8–12	PC	70–100	HgCdTe	40–60	1 × 512	Thermal imaging	12.2	1	1×10^5	3×10^{10}
	PV	60–90	HgCdTe	40–60	512 × 512	Thermal imaging	10.5	1	6.5[d]	5×10^{10}
	QWIP[h]	50–77	GaAs/AlGaAs	40–60	486 × 640	Thermal imaging	9.4[i]	1[i]	0.2[d,i]	1×10^{10}[i]
5–150	PC[f]	30–40	Ge : Hg	50–200	1 × 200	Thermal imaging	13	0.1	5×10^5	3×10^{10}
	PC[f]	20–30	Ge : Cd	50–200	1 × 100	Spectroscopy	28[j]	1[j]	5×10^5[j]	3×10^{10}[j]
	PC[f]	5–15	Ge : Zn	50–200	1 × 100	Spectroscopy	37[k]	1[k]	5×10^5[k]	2×10^{10}[k]
	PC[f]	4–10	Si : Ga	50–100	64 × 256	Threat detection	18	1	5×10^5	2×10^{10}
	PC[f]	4–10	Si : As	50–100	64 × 256	Threat detection	22	1	5×10^5	2×10^{10}
					Broadband					
0.200–200	Bol[l]	300	α-silicon	40–60	640 × 480	Thermal imaging	Full band	10,000	1×10^6	5×10^8
0.200–200	Bol[l]	300	VO	40–60	640 × 480	Thermal imaging	Full band	10,000	1×10^6	5×10^8

[a] Bol = bolometer; CCD = charge-coupled device; PC = photoconductor; PV = photovoltaic; QWIP = quantum well infrared photodetector; and SB = Schottky barrier.
[b] Detectors are all intrinsic unless otherwise noted. See Fig. 1.
[c] Visible bandwidth; near infrared = 0.700 –1.00 μm.
[d] Units are A/W.
[e] Values are for a temperature of 190 K.
[f] Detector is extrinsic.
[g] Detector is an internal photoemission (IPE) device.
[h] Detector is a superlattice (SL) device.
[i] Values are for a temperature of 77 K.
[j] Values are for a temperature of 15 K.
[k] Values are for a temperature of 6 K.
[l] Detector is a thermal device.

It is apparent from Fig. 2 that detection of ambient temperature objects (night vision) requires a detector whose cutoff wavelength is $>4\,\mu m$. Detectors whose response is beyond 8 μm are generally better for thermal imaging applications. However, system sensitivity also depends on several other factors: detector quantum efficiency, detector noise, electronics noise, number of detectors on the focal plane (ie., noise bandwidth), sampling rate, width of spectral band, optical aperture, and detector element size.

FIGURES OF MERIT

Responsivity

A detector responds to variations in photon-flux density ϕ_s by producing a signal voltage V_s, or current. The background photon flux ϕ_B, may result in an offset voltage which can be nulled by a differentiating circuit because ϕ_B is constant for a given measurement. For most detectors, the output voltage depends on the additive effect of ϕ_s over the detector area A. Each photon has energy $h\nu$. The integrated power density over a spectral interval is J_s. The ratio of signal voltage (amperes or electrons) to the signal photon rate is the responsivity R_V in units of V/W:

$$R_V = \frac{V_s}{AJ_S} \text{ volts/watt.} \tag{5}$$

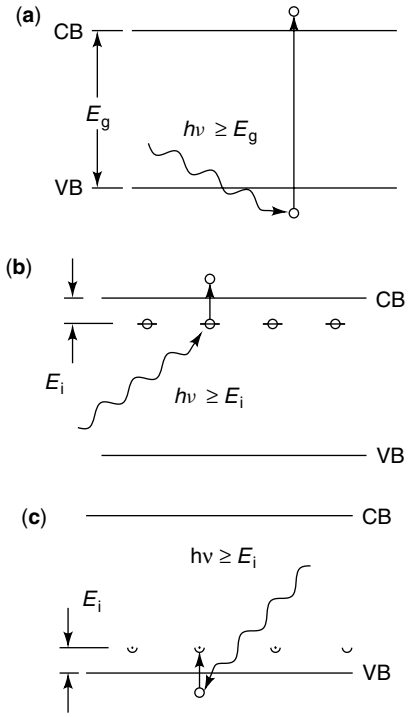


Figure 1. Photoexcitation modes in a semiconductor that has band gap energy, E_g, and impurity states, E_i. The photon energy $h\nu$ must be sufficient to release an electron (○) into the conduction band (CB) or a hole (⦿) into the valence band (VB): (**a**) an intrinsic detector; (**b**) and (**c**) extrinsic donor and acceptor devices, respectively.

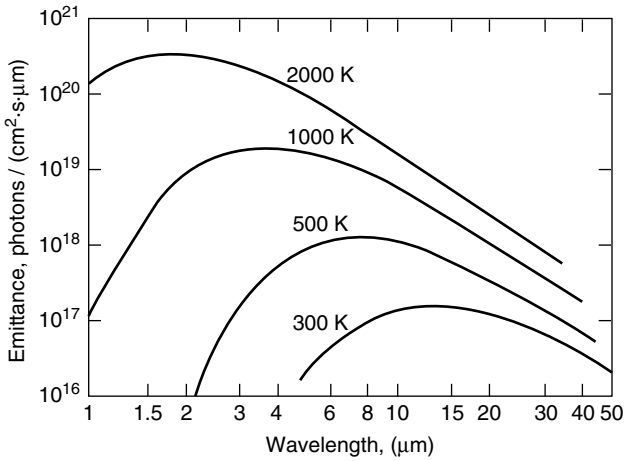


Figure 2. Blackbody radiated photon flux interval distribution from ambient temperature up to 2,000 K. Ambient objects do not have detectable flux for detectors of cutoff wavelengths <ca. 3 μm.

V_s usually is linear with J_s for low intensity but saturates at the higher signal-flux levels where the photon flux disrupts the thermal equilibrium of the semiconductor electronic states. When J_s refers only to those photons capable of generating free carriers, R_V is considered the spectral responsivity R_λ and when the flux includes all wavelengths, R_V is called the blackbody responsivity R_{bb}. The responsivity of photodiodes is often expressed in amperes per watt (A/W) because the diode is used as a current generator when connected to an integrating amplifier. The ideal responsivity for a photodiode is given by $\eta q/E_g$ where η is the quantum efficiency, q is the charge of an electron, and E_g is the band gap energy. Most commercial detectors achieve more than 50% of ideal responsivity except for silicide detectors which typically show less than 2% of ideal responsivity.

Noise

The performance of a system, whether it is an analytical spectrometer or thermal imager, improves as the detector noise is lowered until amplifier noise becomes the dominant noise source, provided that the responsivity is not reduced faster than the detector noise. A well-engineered approach to improving system sensitivity consists of (1) lowering amplifier noise consistent with system bandwidth, power, weight, size, and cost budgets; (2) perfecting the detector material and fabrication processes for least defects to achieve lowest noise; and (3) designing the detector mode of operation, geometry, response time, and bias power for the highest responsivity. The operating parameters of the detector are often adjusted to increase responsivity and noise to the point where the detector noise just exceeds the amplifier noise, if that is possible.

Although photodetector noise (11) originates as random fluctuations of photons or electrical carriers, it is measured using the Fourier transform process as a frequency function for purposes of analysis. The noise spectrum is characterized by frequency-dependent noise at low frequency (called $1/f$ noise), generation–recombination

(g–r) noise in a midrange, and system noise at high frequency. At low frequency, noise typically varies inversely as the square root of frequency and results from surface traps and/or fluctuating defects (12,13). This noise is often reduced by surface passivation to reduce the trap density or create a surface potential to keep the charge carriers away from the surface. Crystal growth resulting in fewer electronic defect states (typically $<1 \times 10^{15}$ cm^{-3}) and elimination of detector fabrication damage also reduce l/f noise. A well-designed detector is dominated by g–r noise associated with random carrier generation and recombination in the semiconductor. The minority carrier lifetime τ determines the extent of the g–r noise plateau. Trapping mechanisms, whether bulk or surface, exhibit behavior similar to the g–r process when the trapping time determines the detector response time. Careful analysis is required to distinguish between recombination and trapping. This identification helps in determining the nature of the defect such as the identity of the impurities. For example, copper atoms in HgCdTe reduce the lifetime and detector responsivity. Characterized by τ, the g–r or trapping noise declines as the inverse of frequency and the system or Johnson noise (JN) becomes dominant at the higher frequencies. In general, noise components V_i, are not correlated and add in quadrature to give the total noise as follows:

$$V_N^2 = V_f^2 + V_{g-r}^2 + V_{JN}^2 + V_{AMP}^2. \tag{6}$$

The noise is expressed as noise density in units of V/(Hz)$^{1/2}$ or is integrated over a frequency range and given as volts rms. Typically, photoconductors are characterized by a g–r noise plateau from 10^3 to 10^5 Hz. Photovoltaic detectors exhibit similar behavior, but the l/f knee may be less than 100 Hz, and the high frequency noise roll-off is determined by the p–n junction impedance–capacitance product or the amplifier (AMP) circuit when operated in a transimpedance mode. Bolometers exhibit an additional noise, V_{TC}, associated with thermal conductance.

Detectivity

Detector sensitivity (1,2) is expressed in terms of the minimum detectable signal power or noise equivalent power (NEP) given in units of watts or W/Hz$^{1/2}$. The reciprocal function when normalized for detector area A and noise bandwidth Δf, is defined as detectivity D^* in units of cm Hz$^{1/2}$/W. Thus,

$$NEP = \frac{V_N}{R_V} \quad \text{watts or watts}/\sqrt{\text{Hz}}, \tag{7}$$

$$D^* = \frac{R_V (A\ \Delta f)^{1/2}}{V_N} \quad \text{cm}\sqrt{\text{Hz/watt}}. \tag{8}$$

The rms noise is measured in a noise bandwidth Δf. The D^* is called D star lambda when the spectral band is limited to a given interval, and D^* blackbody when the total blackbody incident power density is used in the calculation.

The blackbody method of calibrating photodetectors is excellent for cutoff wavelengths larger than 2 μm, but for short wavelength detectors such as visible sensing charge-coupled devices, the unit of lux is the scattered illumination power density of 100 μW/cm^2 in the visible spectrum. Most CCD cameras have a sensitivity of 3–10 lx. Spatial resolution and image display is often a problem with arrays of discrete detectors because small objects are undersampled. The results are moiré patterns and loss of detail. Scanning arrays use overlapping detector geometries to improve the spatial quality of the image. Infrared detector element size is limited to larger than 10 μm by electrical contact methods or photolithographic minimum feature size (1 μm) for the detectors or readout electronics.

Ideal Performance and Cooling Requirements

Free carriers can be excited by thermal motion of the crystal lattice (phonons) as well as by photon absorption. These thermally excited carriers determine the magnitude of the dark current J_d and constitute a source of noise that defines the limit of the minimum radiation flux that can be detected. The dark carrier concentration is temperature-dependent and decreases exponentially with reciprocal temperature at a rate that is determined by the magnitude of E_g or E_i for intrinsic or extrinsic material, respectively. Therefore, it is usually necessary to operate infrared photon detectors at reduced temperatures to achieve high sensitivity. The smaller the value of E_g or E_i, the lower the temperature must be.

The operating temperature depends on the required sensitivity and the temperature dependence of the various thermal-noise components. In all cases, a practical limitation in sensitivity is imposed by an amplifier or recording device. Theoretical performance can be calculated, assuming that the amplifier noise is negligible compared with detector noise. Optimum or ideal detector performance is based on the assumption that the detector noise is determined only by the random fluctuations of the photon-absorption process. This ultimate limit on performance can be calculated easily and is used as a final measure of performance for infrared detectors (14). For example, consider a photodetector as a photon counter, such as a photodiode connected to an integrating amplifier. After a time interval Δt, the integrated charge density Q may be expressed by the following:

$$Q = (J_\phi + J_d)\Delta t. \tag{9}$$

The number of integrated carriers N is QA/q, where q is the electron charge. Because dark current, J_d, is a combination of thermal excitation processes, neglecting avalanche and tunneling, ideal performance occurs when the photon-induced current density J_ϕ is greater than J_d. Fluctuations of N are the dominant causes of noise. The cooling requirement may be calculated for each type of photon detector by analyzing the dark current noise mechanism in terms of thermal equilibrium processes and setting $J_\phi \geq J_d$, where $J_\phi = \eta\phi q$ A/cm^2 and ϕ is the sum of the signal and background photon flux. Because the charge density of a photon-counting detector is sampled at time intervals Δt, the number of charges sampled $N = \eta\phi A \Delta t$.

The generation of photons obeys Poisson statistics where the variance is N and the deviation or noise is $N^{1/2}$. The noise spectral density N_S is obtained by a Fourier transform of the deviation yielding the following at sampling frequency f:

$$N_S = \left(\frac{2N\Delta f}{f}\right)^{1/2} = \left(\frac{2\eta\phi A\Delta t\Delta f}{f}\right)^{1/2}. \quad (10)$$

The signal S, in electrons, is that portion of N which is generated by a signal flux ϕ_S in the same time interval Δt. Therefore,

$$S = \eta\phi_s A\Delta t. \quad (11)$$

The ideal detectivity may be obtained by combining Eqs. (5), (8), (10), and (11). When $\Delta t = f^{-1}$, the ideal spectral detectivity (2) in cm $Hz^{1/2}$/W for a photon limited process becomes

$$D^*_\lambda = \frac{\lambda}{hc}\left(\frac{\eta}{2\phi}\right)^{1/2} \text{cm}\sqrt{\text{Hz/watt}}. \quad (12)$$

When Eq. (12) is valid, the detector is said to be a background-limited infrared photodetector (BLIP). When this is the case, attempts often are made to improve D^* by cold shielding which reduces ϕ. The ideal D^*_λ is shown in Fig. 3 as a function of wavelength, where background photon flux is a parameter. The line of termination in the lower left corner represents D^*_λ values for a $180°(2\pi)$ detector field of view, 300 K ambient background temperature, and background emissivity $\in = 1.0$. As long as the photon-generated current exceeds the dark currents, D^*_λ may be increased by reducing the background flux. For very low values of photon flux, the ideal detection condition of Eq. (12) can be restored by cooling the detector to decrease the dark currents.

The cooling needed to achieve a given detectivity for intrinsic-type detectors may be readily calculated on the basis of the temperature dependence of the minority carrier lifetime and mobility. The calculated

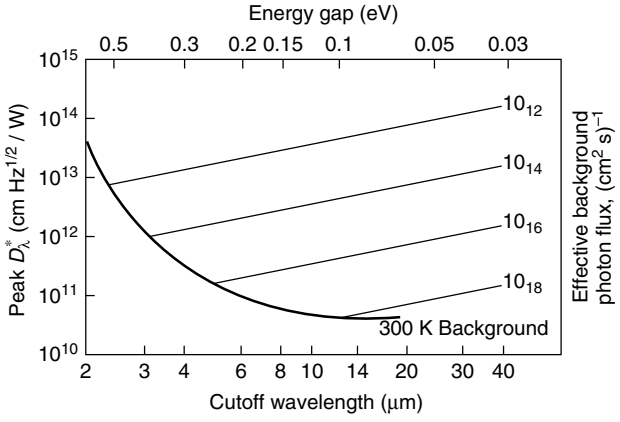


Figure 3. Ideal photon detector sensitivity as a function of cutoff wavelength. Lower background flux generates less photon-induced noise giving higher sensitivity. The sensitivity limit for the condition of 300 K background temperature and hemispherical (2π) field of view is shown.

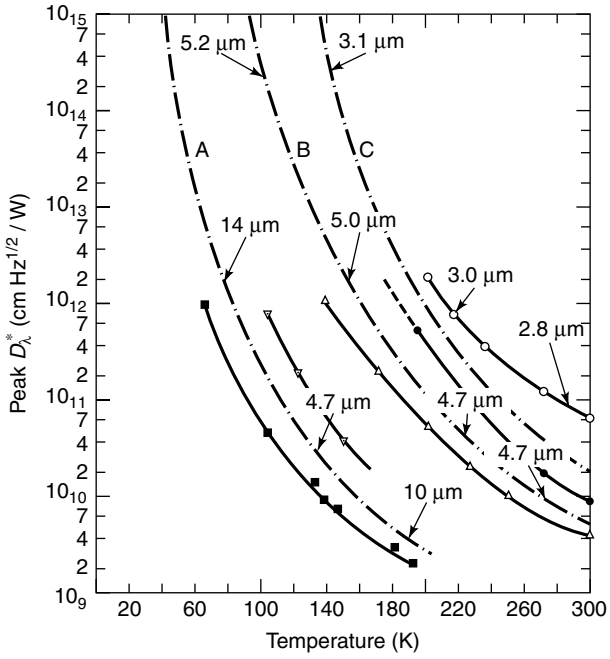


Figure 4. Sensitivity as a function of detector temperature showing (−) experimental results for HgCdTe; (■ and △) $Hg_{1-x}CdTe$, $x = 0.185$ and $x = 0.280$, respectively; (▽) InSb; (•) InAs; and (○) PbS. Also shown are theoretical values (− • −) based on thermal equilibrium statistics, A and B, $Hg_{1-x}CdTe$ where $x = 0.195$ and $x = 0.29$, respectively; and C, InAs.

values for indium arsenide (InAs) and two compositions of HgCdTe, are given in Fig. 4. Experimental data for lead sulfide (PbS) and three HgCdTe detectors of different compositions are also shown. The importance of cooling is obvious.

Detectors are often cooled by providing good thermal conductivity to a suitable cryogen (2). The most readily available coolants are solid carbon dioxide (CO_2) at 195 K, liquid nitrogen (N_2) at 77 K, and liquid helium (He) at 4.2 K. If the desired temperature is not about 200 K, a thermoelectric cooler may be used. Joule–Thompson and expansion engine coolers are available for operation to 4.2 K and below. A cooled detector must be contained in a vacuum vessel that thermally isolates the detector element from other than the coolant by low thermal conduction materials and vacuum. If low absorbing extrinsic detectors are used, it is customary to enclose the detector element in a cooled chamber that has reflecting walls, that is, in an integrating chamber, to obtain multiple passes of the radiation. The latter can also be accomplished by shaping the element so as to ensure multiple, total, internal reflections. Thin detectors can be sized to produce optical resonance in the detector element, making the detector itself an optical resonant cavity.

A detector for a specific application should be chosen to minimize the cooling requirements and the magnitude of the background radiation noise; therefore, in detector selection, the cutoff wavelength should be only slightly greater than that required by the application. If the signal radiation has a substantial path length through the atmosphere, then atmospheric

transmission characteristics must be considered. Because of the presence of CO_2 and H_2O vapor, many absorption bands appear in the atmospheric transmission spectrum. Windows of high transmittance occur between the bands. The principal windows are at 1.0–2.5, 2.9–4.2, 4.4–5.3, 7.5–14.0, and 16–23 μm (2).

PHOTODETECTOR MODES OF OPERATION

The need for high performance and low cost detectors has resulted in fewer than two dozen types of detectors that are commercially available. These are made from only 10 basic semiconductor elements or compounds. Popular detectors, their modes of operation, cutoff wavelength, temperatures of operation, response times, responsivities, and detectivities are listed in Table 1 (15). The values chosen are for near-optimum performance at a given temperature. Silicon-based photon detectors such as photovoltaic and charge-coupled detectors operate near room temperature. Cadmium sulfide and germanium detectors also require no cooling. Wide acceptance of the ternary compound semiconductor mercury–cadmium–telluride (MCT), HgCdTe, as a photoconductor began in 1972. Requirements for large focal planes has since led to significant developments of the MCT photodiode focal plane (16). The controllable (by composition) band gap of MCT in each case results in a cutoff wavelength that is tailored for a specific application (Fig. 4). Indium gallium arsenide, lead sulfide, indium arsenide, and MCT (40% at. wt CdTe) detectors can be operated at 300 K but perform much better at lower temperatures (2,14).

Four principal modes of semiconductor-based photodetectors are the photoconductor, photodiode, charge-mode device, and bolometer. The Schottky barrier mode using internal photoemission has been well developed using platinum silicide, but low quantum efficiency restricts the range of uses. The quantum well infrared photodetector (QWIP) is included here because artificially structured materials such as this superlattice detector represent a new metallurgy. The semiconductor materials for most applications are Si, Ge, GaAsP, InGaAs, InSb, PbS, CdS, and HgCdTe. These materials grown as single crystals and thin polycrystalline or amorphous films are doped with various elements such as boron, phosphorus, indium, and gold, to control the polarity of the conductivity and for selective optical absorption by the introduction of impurity states in the forbidden energy gap. Photodetection using these materials extends from the ultraviolet (0.3 μm) to long-wavelength infrared (200 μm). Higher sensitivity, especially in the infrared, can be achieved by cooling the detector and its immediate surroundings to low temperature, typically 77 K. Interfacing with signal processing electronics can be difficult when the photon signal is weak and imaging systems require sophisticated signal processing to handle the large number of pixels found on a typical focal plane.

Photoconductors

A photoconductor (17) is a semiconductor resistor whose resistance lowers when photons generate excess carriers. The two general classes of photoconductor are extrinsic, which is characterized by impurity states that are emptied by photons, and intrinsic, where the photons excite electrons directly from the valence band to the conduction band (Fig. 1). The photoconductor is connected in series to a constant load resistor and current source. The ambient light or background radiation sets up a steady-state resistance in the detector. Changes in illumination produce changes in resistance and thereby changes in the voltage across the load resistor. When the load resistance is much larger than the detector resistance, the signal voltage is given by Eq. (13):

$$V_s = \frac{\eta\ \Phi_s \tau\ V}{n\ t}, \tag{13}$$

where the signal flux ϕ_s, is caused by the changing background. At low bias, the signal is linear with the bias voltage V, but at high bias, the minority carrier sweep time becomes less than the lifetime τ, and the signal saturates. The majority carrier density n is typically 1×10^{15} to $1 \times 10^{16}/\text{cm}^3$. The thickness t is typically one to three absorption lengths (8–20 μm).

Photodiodes

Photovoltaic detectors (17) generate a voltage (open circuit) or a current (closed or integrating circuit) when they receive radiation and therefore do not require an external bias for operation. The charge integrated on a capacitor of the integrating circuit is proportional to the intensity of radiation. The current I_s, generated by the photovoltaic cell is given by the following:

$$I_s = \eta\ \Phi_s\ q\ A. \tag{14}$$

The quantum efficiency can be determined by using the measured current and a calibrated photon source, for example, a blackbody. The detector area A should be measured by a spot-scanning apparatus because the minority carrier diffusion length effectively increases the diode photon-active area. A photovoltage can arise from the presence of internal fields which are formed at the interface between a semiconductor and a metal (Schottky) barrier, at the interface between different dopings in a semiconductor (homojunction), and at the interface between two different semiconductors (heterojunction). Light releases mobile electric carriers within these barrier layers. These carriers are separated by the field and produce a photovoltage between them, as indicated in Fig. 5. Photovoltages can be produced in the bulk of a semiconductor because of the difference in the diffusion of electrons and holes or if there is a gradient in the impurity concentration within the material, but these mechanisms are not widely used.

The impedance area product ZA of a diode may be obtained from differentiating the diode current–voltage relationship and may be expressed as follows, at zero applied basis:

$$ZA = \frac{k\ T}{Bq\ J_d}. \tag{15}$$

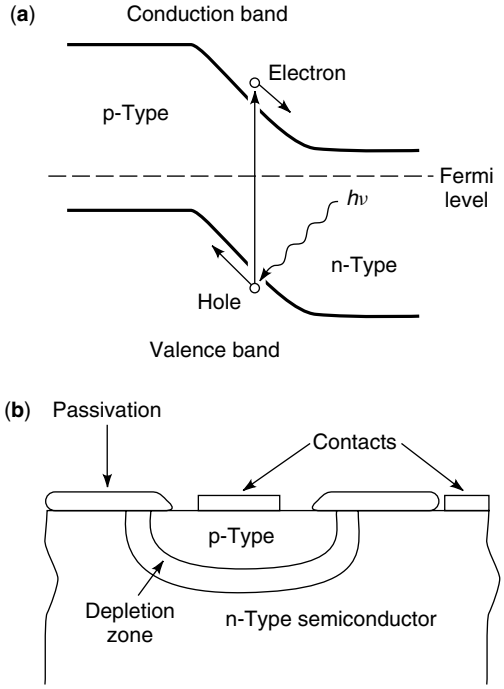


Figure 5. The photodiode detector: (**a**) band model where the photon generates electron–hole pairs that are separated by the built-in potential that sets up a photocurrent; (**b**) physical model for a planar diode. The passivation is typically SiO_2 for Si diodes, an In oxide for InSb diodes, and CdTe for HgCdTe diodes.

Under little or no illumination, J_d must be minimized for optimum performance. The factor B is 1.0 for pure diffusion current and approaches 2.0 as depletion and surface-mode currents become important. Generally, high crystal quality for long minority carrier lifetime and low surface-state density reduce the dark current density which is the sum of the diffusion, depletion, tunneling, and surface currents. The ZA product is typically measured at zero bias and is expressed as R_0A. The ideal photodiode noise current can be expressed as

$$i_N^2 = 2qA(2J_d + J_\phi)\Delta f, \tag{16}$$

where J_ϕ is the photon current density. Equations (15) and (16) can be used to determine the cooling requirements for photodiode detector performance because diffusion current depends exponentially on temperature (18). Because the diffusion current represents a valence band-to-conduction band excitation, the thermal limit, when $J_d > J_\phi$, is essentially the photoconductive limit (Fig. 4). Combining Eqs. (5), (8), (14), (23), and (24) gives the following:

$$D^* = q\eta/h\upsilon(4\,kT/ZA + 2qJ_\phi)^{1/2}. \tag{17}$$

Thus, ZA must be increased to achieve BLIP D^* such that $ZA > 2kT/qJ_\phi$. When $J_\phi > J_d$, Eq. (17) reduces to the ideal photon limit of Eq. (12). The response time is equal to the ZC product, where C is the diode capacitance, or when using an integrating amplifier, the response time is determined by the closed-loop gain.

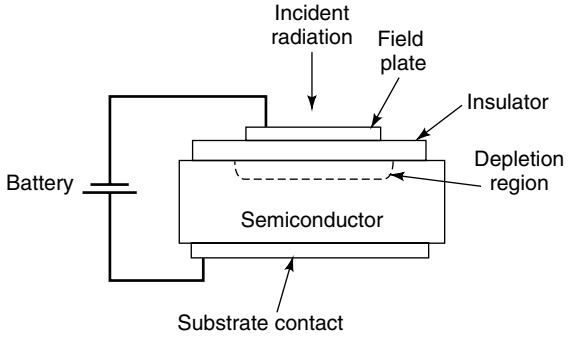


Figure 6. Device model for an MIS photodiode, the basic building block of the CCD photodetector. The depletion region is generated by the battery potential. See text.

Charge Mode Detector

Charge mode devices (19,20) are a group of detectors that use the metal–insulator–semiconductor (MIS) capacitor as their basic constituent. An MIS capacitor is a device that consists of a nondegenerately doped semiconductor substrate, a thin dielectric layer, and a metal gate, as shown in Fig. 6. The application of a potential difference across the metal gate and the semiconductor produces a surface potential barrier and an associated depletion region in the semiconductor. Periodic pulsing of the gate-to-substrate bias creates a nonequilibrium condition suitable for storing optically generated free carriers.

The model for an MIS capacitor, shown in Fig. 7, includes a metal gate-to-semiconductor bias of V_g volts that produces an insulator bias V_I, and surface potential ϕ_s. The negative for n-type material gate charge Q_s, produces an ionic space charge Q_d, and inversion layer charge Q_p. From the basic relationships (21) that follow,

$$V_g = V_I + \Psi_s, \tag{18}$$

$$Q_s = Q_d + Q_p, \tag{19}$$

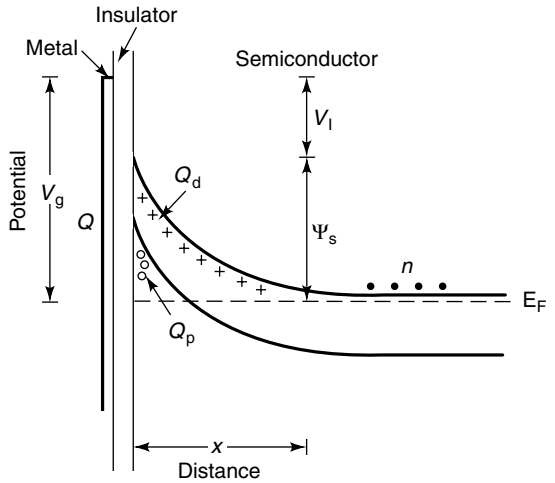


Figure 7. Band model for the charge mode detector biased to deep depletion. The charge Q_p, integrates in the potential well defined by the insulator and valence band. See text.

where $Q_s = C_I V_I$ and $Q_d = C_d \Psi_s$,then

$$\Psi_s = V_g/(1 + C_d/C_I), \tag{20}$$

where C_d is the capacitance associated with the space–charge layer geometry of charge n at distance x. From semiconductor theory,

$$\Psi_s = n.qx^2/2\varepsilon, \tag{21}$$

where ε is the dielectric permittivity of the semiconductor and, from geometry,

$$C_d = \varepsilon A/x. \tag{22}$$

The basic MIS condition is obtained by combining Eqs. (18)–(22):

$$\Psi_s + (\varepsilon qn/2)^{1/2} A(\Psi_s)^{1/2}/C_I = VQ_p/C_I. \tag{23}$$

At initial turn-on, the inversion charge $Q_p = 0$, and when n is small and C_I is large, $\Psi_s =$ approximately V, that is, a potential well of nearly V is formed. Thus, the charge-handling capacity Q_{FW} is

$$Q_{FW} = \text{approximately } C_I V, \tag{24}$$

and it can be seen that it depends on both the physical dimensions of the capacitor and the applied voltage. Typical full well capacities range from 10,000 carriers to more than 5,000,000 carriers, depending on the specific charge mode device.

As time progresses, charge Q_p is generated by photon currents, semiconductor dark currents, diffusion, depletion, surface, avalanche, and tunneling. The time required to fill the well by dark currents is the storage time τ_{st}, given by the following:

$$\tau_{st} = C_I V/J_{dark}. \tag{25}$$

The storage time is the maximum time the potential well can be used for collecting a photon-generated charge. Because J_d depends strongly on band gap energy and temperature of operation, τ_{st}, in practice, extends from hours for silicon at 300 K to a few microseconds for HgCdTe, where the cutoff wavelength $\lambda = 12$ μm at 77 K. Indium antimonide operates in a few tenths of a second at 77 K (22).

As previously indicated, an MIS capacitor can be utilized as a photon detector. A negative pulse on an n-type semiconductor generates a deep depletion region at the surface which, in turn, acts as a storage site for photogenerated holes. This type of detection is true charge integration and is the simplest charge mode detection concept. Other MIS devices exist that contain arrays of overlapping, that is, metal gates in a given detector element. These MIS-based detectors are called charge-transfer devices (CTD). The two most common forms of CTDs are charge-injection devices (CID) and charge-coupled devices (CCD).

The CID is a dual gate version of a CTD. In the CID, two gate electrodes form two distinct but physically adjacent depletion capacitances, as shown in Fig. 8a. Under normal

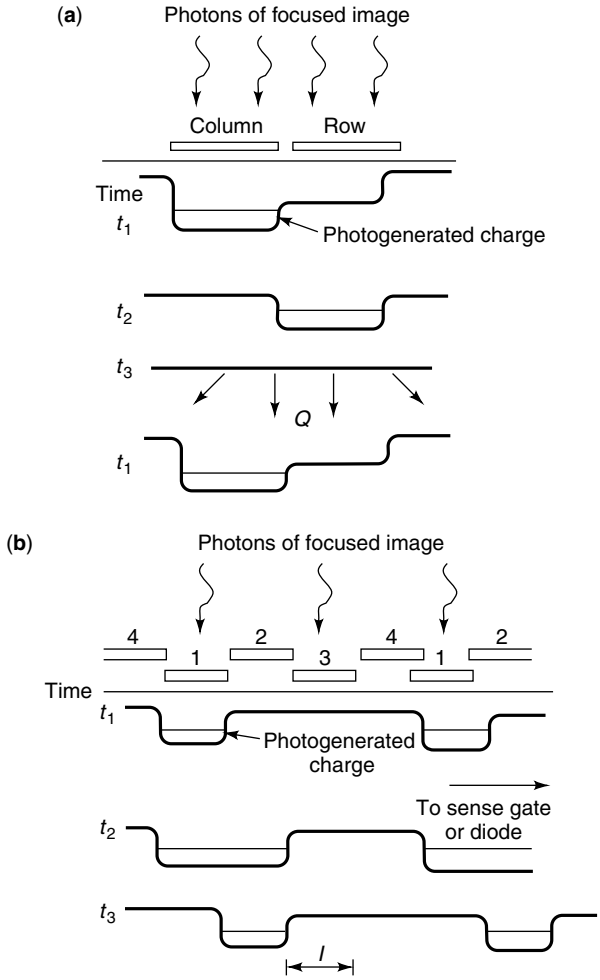


Figure 8. Conceptual cross-sectional views of (**a**) a CID imager showing charge integration (t_1), charge measurement (t_2), and charge removal (t_3); and (**b**) a CCD shift register showing charge transfer from a phase one well to a phase two well.

operating conditions, each of the two gate electrodes is biased to form a depletion region in the underlying semiconductor. However, the surface potential minimum under one of the electrodes is much greater than that of the other electrode. As a result, the photogenerated charge produced in the detector collects in the deeper potential minimum. To measure the number of stored carriers, the potential difference between the collection electrode and the semiconductor substrate is reduced, thereby collapsing the storage well and forcing the collected charge to transfer to the region controlled by the second gate electrode. The collected charge can, in principle, be quantified by monitoring the change in surface potential of this second electrode during the transfer operation. In practice, this is accomplished by monitoring the instantaneous current through the electrode at the instant of charge transfer. The detector is reset to the empty state by simultaneously collapsing the depletion regions of both capacitors. This action forces the free carriers into the semiconductor substrate where they are subject to recombination mechanisms. Thus, a key feature of a CID is that the photogenerated charge is both collected and

sensed inside a single detector element. The signal charge is transferred from that element simply to eliminate it from the detector before the next detection cycle.

The charge-coupled device (CCD) is a multigate CTD. As for the CID, the CCD gates are physically adjacent and are periodically pulsed, inducing the formation of deep depletion regions in the underlying semiconductor. However, unlike the CID, the CCD physically transfers the collected photogenerated charge from a given detector site to a separate charge measurement structure. Figure 8b shows a conceptual cross-sectional view of a CCD shift register. As the photons of an image that is focused on the device generate a charge, the MIS gates of the shift register are biased in sequence to produce, continually, a shifting potential well pattern. Thus, at time t_1, phase one is on, and a charge is integrating in the potential well beneath the MIS phase one gates. At time t_2, phases one and two are turned on, and the charge distributes beneath phase one and phase two gates. At time t_3, phase one is turned off, and the charge is only beneath phase two. Pulsing phases three and four in the proper sequence with phase one and two moves the charge packet through the shift register. For an area array CCD, the imaging portion of the array transfers charge, once per line interval, into a readout shift register in a manner equivalent to that described for the imaging shift register. Each charge packet in the imaging register is then shifted discretely to the charge-detection structure that resides at the end of the register. This sense electrode or diode is in turn typically connected to a low-noise, charge-sensitive amplifier that converts the signal charge to a voltage. After all charge packets in the shift register are read out, the next line of information from the imaging portion of the array is transferred, and the process is repeated.

Bolometers

The bolometer (23) has made a comeback as a popular detector due to recent advances in micromachining technology. When applied to silicon, it is feasible to fabricate large arrays of bolometers and detecting elements that have very small thermal mass. The imaged infrared radiant power slightly heats the bolometer film a few millidegrees K reducing the electrical resistance. The resulting change of the bias current is the signal. The noise components are l/*f*, Johnson, thermal conductance, and photon. The photon noise is actually the radiative part of the thermal conductance noise.

From bolometer theory (1), the change in film temperature is proportional to the absorbed power and thermal resistance (inverse of the thermal conductance) and is given by the following:

$$\Delta T_\mathrm{b} = \Delta J A \varepsilon_\mathrm{a} R_\mathrm{th} \frac{\left(1 - \exp \frac{-\tau_\mathrm{s}}{R_\mathrm{th} C_\mathrm{h}}\right)}{\left(1 + \exp \frac{-\tau_\mathrm{s}}{R_\mathrm{th} C_\mathrm{h}}\right)} = \Delta J A \varepsilon_\mathrm{a} R_\mathrm{th} \tanh\left(\frac{-\tau_\mathrm{s}}{2R_\mathrm{th} C_\mathrm{h}}\right), \quad (26)$$

where ΔJ is the differential power density at the detector caused by a change in target temperature in W/cm^2; A is the active pixel area, 1.2×10^5cm^2; ε_a is the IR absorption efficiency, 90%; R_th is the thermal resistance and is approximately 1.5×10^7K/W; τ_s is the half period of chopped radiation ($1/2f_\mathrm{c}$), typically 16.7 ms (30-Hz modulation); and C_h is the heat capacity of the pixel element and is approximately 0.7×10^{-9} J/K.

The power density J_s may be calculated by using Plank's radiation law. For a 300 K scene temperature and spectral region from 8 to 12 μm,

$$J_\mathrm{s} = 2.0E - 4\left(\varepsilon_0 \Delta T_\mathrm{s} \frac{\Omega}{\phi}\right), \quad (27)$$

where ε_0, the optical efficiency, is typically 80%, ΔT_s is the differential target temperature, and Ω is the collection solid angle of the optical system [$= \pi \sin^2(\tan^{-1} 1/2f_\mathrm{n})$].

For the bolometer detector that has the characteristics above and operates with *f*/l optics ($f_\mathrm{n} = 1$), the peak to peak temperature change of the α-Si film is approximately 3 mK per degree K change in scene temperature.

Signal. The peak to peak signal voltage V_s, is proportional to the bias power and can be expressed as follows:

$$V_\mathrm{s} = \alpha (P R_\mathrm{b})^{1/2} \Delta T_\mathrm{b}, \quad (28)$$

where α is the temperature coefficient of resistance and is 2.8% K for α-Si, P is the bias power for the bolometer element (typically 0.3 μW), and R_b is the bolometer resistance (typically $3 \times 10^7\Omega$ for α-Si and $1 \times 10^4\Omega$ for VO).

The voltage responsivity is simply given by Eq. (29):

$$R_\mathrm{V} = \frac{V_\mathrm{s}}{\Delta J A} = \alpha (P R_\mathrm{b})^{1/2} \varepsilon_\mathrm{a} R_\mathrm{th} \tanh\left(\frac{-\tau_\mathrm{s}}{2R_\mathrm{th} C_\mathrm{h}}\right). \quad (29)$$

Responsivity values are typically in the 5×10^5 V/W range for nominal bias.

Noise. The noise components may be expressed as voltage. Excess (l/*f*) noise:

$$V_\mathrm{EX} = \int_{\Delta f} \left(\frac{\gamma}{\sqrt{f}}\right)^2 df = K_\mathrm{V} (P R_\mathrm{b})^{1/2}, \quad (30)$$

where K_V is the l/*f* noise coefficient that is typically in the range from 6 to 12 μV per volt bias when integrated from 1 to 100 Hz. Johnson noise V_JN is

$$V_\mathrm{JN} = (4kTR_\mathrm{b}\Delta f)^{1/2}, \quad (31)$$

where k is the Boltzmann constant, T is the detector temperature (300 K), and Δf is the noise bandwidth (100 Hz). Thermal conductance noise V_TC is

$$V_\mathrm{TC} = \alpha (P R_\mathrm{b})^{1/2} \left(\frac{kT^2}{C_\mathrm{h}}\right)^{1/2}, \quad (32)$$

when the thermal response time $R_\mathrm{th} C_\mathrm{h}$, is less than the frame time.

NEP. The noise components may be combined with the responsivity to give the noise equivalent power (NEP) equation:

$$\mathrm{NEP} = \frac{(V_{\mathrm{JN}}^2 + V_{\mathrm{EX}}^2 + V_{\mathrm{TC}}^2)^{1/2}}{R_{\mathrm{V}}} \tag{33}$$

Measured values are typically less than 100 pW for a 30-Hz scene modulation.

Detector Fabrication and Performance

Detector performance and focal plane array development have improved significantly since the early 1980s (24,25). The spectral sensitivities of several photodetectors given in Fig. 9 cover the spectrum from ultraviolet to far infrared. The typical operating temperature is given, and the theoretical limit is that calculated for a detector exposed to hemispherical radiation from a 300 K background.

Charge-Coupled Devices and Imaging Arrays

The single most popular type of visible photon detector is the charge-coupled device (CCD). Developed in 1970 (26,27) as a dynamic memory storage device, the CCD was quickly adapted for use as an imaging device. The CCD has been a resounding commercial success and has become the detector of choice in the video camera and electronic still photography markets. Furthermore, since its inception, the CCD has also been an invaluable tool in a wide range of scientific applications involving the detection of near-infrared, visible, ultraviolet, and X-ray photons, as well as in the detecting ionizing radiation and charged particles.

CCDs are sensitive photon detectors. The physical mechanism governing photon detection in silicon depends on the energy of the incident photons. For the near-infrared to visible spectral band, silicon performs as an indirect band gap semiconductor. Incident photons are converted to electron–hole pairs through interactions with the silicon lattice, producing one electron–hole pair for each absorbed photon. For higher energy photons such as those of the extreme UV and X-ray spectral bands, direct band gap detection dominates. In this case, one electron–hole pair is generated for every 3.65 eV of energy in the absorbed photon. If the scene of interest produces only a small number of high energy photons that impinge on the detector array during each frame interval, the number of signal electrons in each pixel is a measure of the energy of the incident photon. In such applications, CCDs can be used to obtain spectral as well as spatial information (28).

Utilization. CCDs are often grouped into two broad categories: commercial grade and scientific grade. The format and required performance of the CCD for both categories are driven by the application for which the array is designed. In general, commercial-grade CCDs are designed to optimize image resolution and color fidelity. On the other hand, scientific-grade CCDs are designed to maximize photon quantum efficiency and minimize device noise. The vast majority of available CCDs have been designed for commercial applications. Because of the relatively small number of devices supplied yearly for scientific applications and the more stringent performance requirements, scientific-grade CCDs are typically more expensive

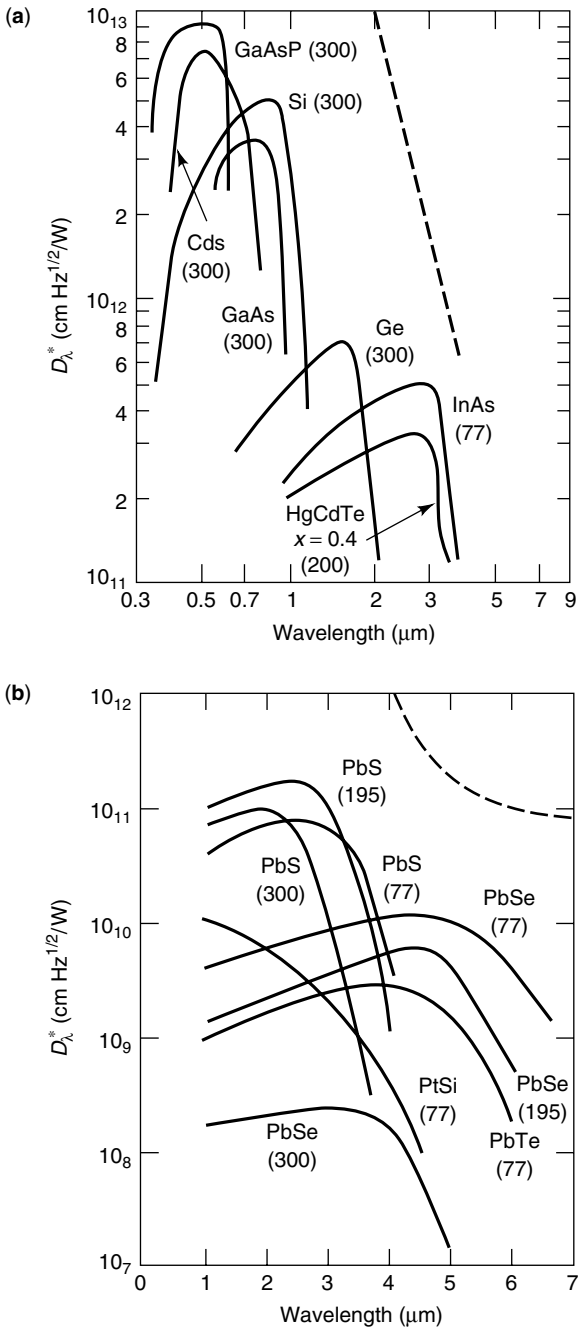


Figure 9. Spectral sensitivity of detectors where the detector temperatures in K are in parentheses and the dashed line represents the theoretical limit at 300 K for a 180° field of view. (**a**) detectors from near UV to short-wavelength infrared; (**b**) lead salt family of detectors and platinum silicide; (**c**) detectors used for detection in the mid- and long-wavelength infrared. The $Hg_{1-x}CdTe$, InSb, and PbSnTe operate intrinsically, the doped silicon is photoconductive, and the GaAs/AlGaAs is a structured superlattice; and (**d**) extrinsic germanium detectors showing the six most popular dopants.

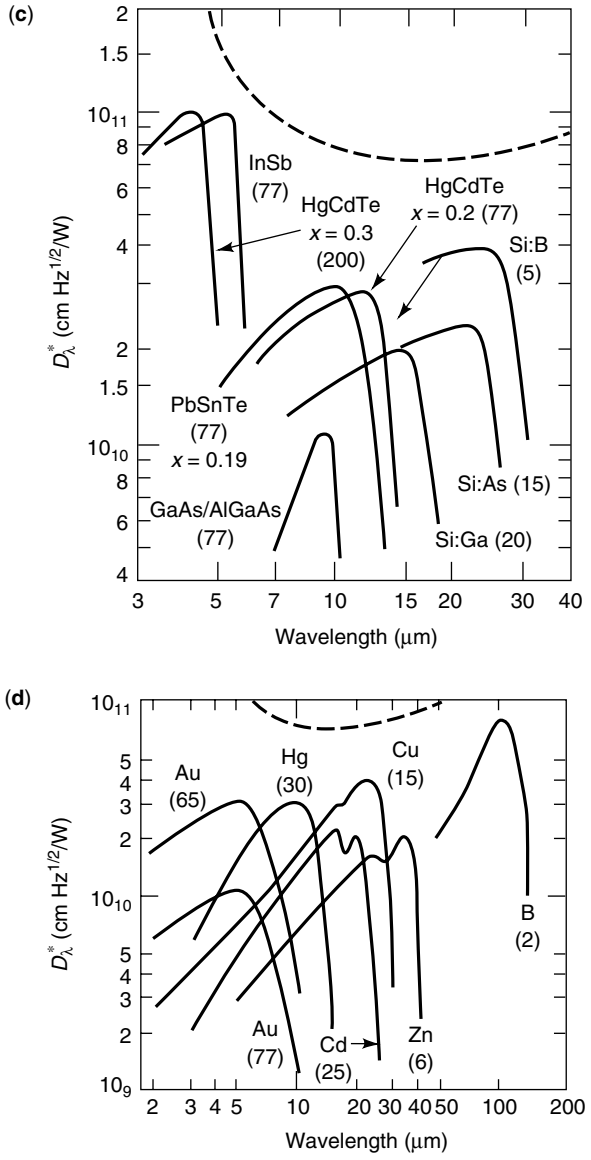


Figure 9. (*Continued*)

than their commercial counterparts. Commercial-grade CCDs can be adapted for use in scientific applications where a moderate photon level is present, such as some forms of visible spectroscopy.

Most CCDs are specifically designed for video camera applications, which detect photons in the visible portion of the electromagnetic spectra. Video camera CCDs have specific camera formats compatible with standard video display systems. In the United States, video displays utilize the RS-170 standard, commonly known as the National Television System Committee (NTSC) 525-line television standard. The RS-170 standard places many constraints on imaging devices. These devices must have a frame time of 1/30 second, and two interlaced image fields of 1/60 second each are required during each frame. The standard luminance electrical bandwidth of 4.2 MHz coupled with the 4 : 3 picture aspect ratio require the pixel count and the pixel size of the commercial CCD array to fit within narrowly defined regimes. Specifically, to meet the electrical bandwidth requirements of the RS-170 standard, monochrome imagers must have a minimum of 240 pixels in the vertical dimension and a minimum of 450 pixels in the horizontal dimension. The distance along the diagonal of the active imaging portion of the CCD must be one of several specific values in the 5.46–25 mm range. The specific value depends on the particular camera format, for example, 8 mm, VHS, for which it is designed. These format and size requirements dictate a rectangular rather than square CCD pixel design. Data rates from the imaging devices must be chosen so that the image scene readout time matches the RS-170 interval. The minimum data rate for monochrome imagers is 6 MHz. The actual rate spans from 6 to 10 MHz, depending on the specific device.

Color applications require more complex image detection schemes due to the need for spectral differentiation. Signal samples from at least three distinct spectral bands are required to reproduce a color image accurately. Typically, a monolithic color filter array that consists of alternating windows of appropriately colored filter media is attached to the CCD for color differentiation. A common approach to obtaining the scene information required for faithful color image reproduction is to increase the horizontal pixel count of the CCD by a factor of 3 (29). In this method, a filter that has alternating red, blue, and green stripes is then attached so that each pixel column aligns with a filter stripe. This color CCD uses three adjacent pixels, one red, one green, and one blue, to represent a single spatial sampling of the scene. In effect, the device can be thought of as three integrated monochrome CCDs. The color CCD has three on-chip output channels, one for each spectral component. As each pixel is read, the simultaneously generated signal information from each channel is remixed so that it correctly simulates a color display. More ambitious methods exist for achieving color image reproduction (30,31). These methods use the horizontal pixel count found in monochrome CCDs but require intricate clock control methods to intermix the spectral components appropriately for display during the actual integration interval. Intermixing is produced by using a repetitive mosaic pattern of colored filters, typically cyan, yellow, and green. The imaging results from these devices rival those obtained from the three-channel method.

Video camera applications obligate the CCD imager to have an internal, electronic, image shuttering capability to differentiate the data in one image frame clearly from that of the next. The two most common forms of internal shuttering are frame transfer and interline transfer (32). In the frame transfer technique, the image gathered from one data frame is rapidly transferred from the contiguous image collection region to a secondary, optically shielded storage site that resides between the image collection and shift register regions. The image held in the shielded region is then read out during the time interval of the next image formation in the optically active area. In the interline transfer technique, each pixel in the imaging portion of the array is segmented into an optically active portion and a shift register portion. The image gathered from one data frame is quickly transferred

from the optically active to the shift register portion of the pixel and is read out during the next image collection interval. Unlike the frame transfer method, no additional image storage section is required for the interline transfer method. However, the frame transfer device is more efficient at photon detection because there are no optically shielded features in the image collection region.

Most CCDs for consumer applications are illuminated from the front surface of the device. Photons that impinge on the front surface of the detector must pass unimpeded through the gate electrode layers to be collected in the photon-sensitive silicon beneath. Typically, some loss does occur for photons of wavelengths shorter than about 550 nm due to absorption in the gate electrodes. For standard polysilicon gate CCDs, photons of wavelength less than 400 nm are absorbed by the gate material. This absorption process affects photons that span the UV to the low-energy X-ray spectrum. Photons of wavelengths greater than 650 nm pass easily through the gate electrodes, but due to the lower value of the absorption cross section of silicon, a several micrometer path length is required to provide a high probability of absorption. A deeply generated signal charge can produce image smearing due to charge diffusion before collection in the potential well of the CCD pixel. Both the photon collection efficiency and image quality are strong functions of the particular CCD design. Typical consumer-quality CCDs provide sensitivities that span the 50–500 mA/W range and have minimum detectable signal levels from 40–400 photons at the standard 16-ms image integration interval.

CCDs are widely used in electronic still photography. The late 1990s saw the development of a myriad of commercial digital electronic camera systems. The quality of these systems spans the range of very low-cost, low-resolution cameras for Internet imagery to very expensive, 35-mm film quality imagery for professional photography. Unlike film-based cameras, images are stored on electronic media such as floppy or optical disks. The image can be read into a computer-based system, edited, or enhanced as desired, and stored again on the digital medium. Hard copies are obtained by devices ranging from low-cost photographic quality color printers for personal computers to separate image transfer systems specifically designed for photographic paper. Still photography places more stringent specifications on a CCD than those required of RS-170 video cameras in both performance and resolution. Matching the image quality of 35-mm film requires a CCD whose format is larger than $4{,}000 \times 4{,}000$ pixels and can faithfully reproduce an image at low light levels. The development of digital photography has raised the performance level of some commercial CCDs to that of custom scientific quality devices.

Whereas some consumer CCDs can occasionally be adapted for scientific endeavors, in general these devices have format and operability constraints that negatively impact their performance in many applications. Scientific CCDs are somewhat more difficult to categorize because of the wide range of capabilities available in such devices. The performance features of a specific scientific CCD are determined by the device design. A few of the more significant performance features which may be found in a given scientific CCD include the following: (1) spectral response that extends over a wide range of photon and charged particle energies; (2) noise levels as low as one electron per pixel, which result in very high signal-to-noise ratios; (3) very high charge-transfer efficiencies to achieve no measurable degradation in spatial or spectral information upon image readout; (4) dark current generation rates that permit several minute integration times at room temperature and multiple hour integration times at cryogenic temperatures; (5) high photon detection efficiency; and (6) high frame rate operability. In addition, scientific CCDs are typically operated in the full-frame readout mode (32), in which an external mechanical shutter shields the imager during scene readout.

Scientific CCDs have been used in a wide range of applications. In the visible and UV portions of the spectrum, scientific CCDs have proven to be effective tools for low photon flux applications. Specifically, CCDs have been used as image detectors in ground-based telescopes (33,34), as well as in space-borne systems such as the wide-field planetary camera of the earth-orbiting space telescope (35,36), the Giotto mission to Halley's comet (37), the Galileo mission to Jupiter (38,39), and the Cassini mission to Saturn (40). Figure 10 is a photomicrograph of a 6" diameter silicon wafer that contains more than a dozen different varieties of scientific CCDs. The wafer contains a $4{,}096 \times 4{,}096$ pixel device for astronomy, several $1{,}024 \times 2{,}048$ pixel imagers for slow scan frame store operation, a $1{,}024 \times 1{,}024$ element CCD for use in very high speed image acquisition, and numerous other imager types. The photomicrograph was

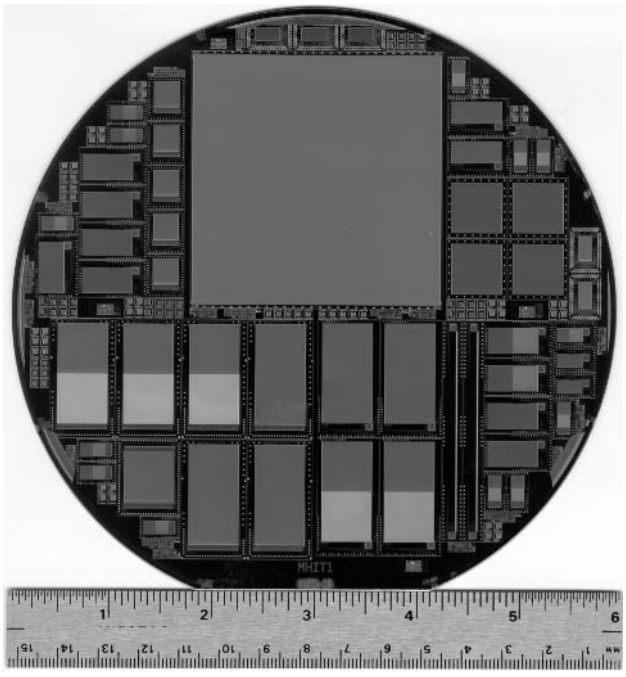


Figure 10. Figure 10 is a photomicrograph of a 6" diameter silicon wafer containing over a dozen different formats of scientific CCDs. Included in the wafer is a $4{,}096 \times 4{,}096$ pixel device for astronomy, several $1{,}024 \times 2{,}048$ pixel imagers for slow scan frame store operation, and a $1{,}024 \times 1{,}024$ element CCD for use in very high speed image acquisition.

taken from a complete Si wafer after device processing but before device packaging.

The same features that make scientific CCDs excellent devices for astronomy, that is, high photon collection efficient and low readout noise, also make CCDs excellent tools for chemical analysis (41). CCDs can be used in many forms of spectroscopy (42), including absorption (43,44), fluorescence (45–47), luminescence (48,49), emission (50–52), and Raman (53–55), across a spectral range of 0.1 to 1,100 nm. The high dynamic range (typically >100 dB) of the CCD is an invaluable property for identifying weak spectral lines. Furthermore, the inherent integration of many highly sensitive photodetector elements in a small area has, in some instances, allowed revolutionary experimental designs (56). Several visible light spectrometer systems that use CCDs as the imaging device are commercially available.

In the X-ray portion of the spectrum, scientific CCDs (57,58) have been used as imaging spectrometers for astronomical mapping of the sun (59), galactic diffuse X-ray background (60), and other X-ray sources. Additionally, scientific CCDs designed for X-ray detection are also used in X-ray diffraction, materials analysis, medicine, and dentistry. CCD focal planes designed for infrared photon detection have also been demonstrated in InSb (61) and HgCdTe (62) but are not available commercially.

Fabrication. Although CCDs have been fabricated in many semiconducting materials such as Ge (63), InP (64), and HgCdTe (65), by far the most readily available devices are those that use Si as the semiconductor. There are several common types of silicon CCDs. All share certain processing steps, and most utilize p-type silicon as the semiconducting material. Single silicon crystals are grown (66) in conventional Czochralski vertical pullers using a single-crystal silicon seed dipped and rotated in a silicon melt. The large boules of silicon are sawed into wafers. The following fabrication discussion describes a process for creating a generic four-phase CCD from p-type silicon (67).

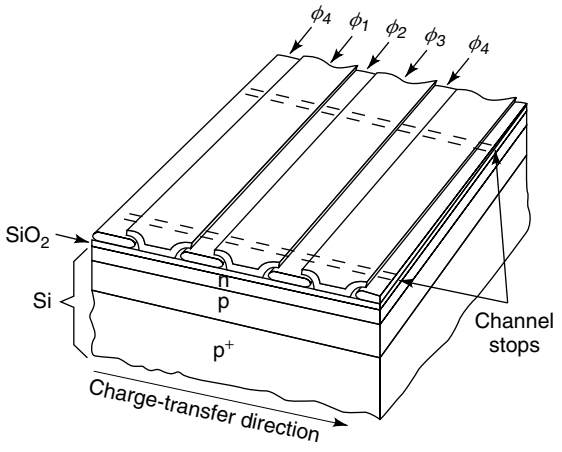


Figure 11. Cutaway view of a CCD shift register where the ϕ_i represent gate electrodes. Voltage pulses applied to the phase gates move a photogenerated charge in the charge-transfer direction. The channel stops confine the charge during integration and transfer. See text.

A cross-sectional view of a four-phase Si-based CCD is shown in Fig. 11. The starting wafer is a thin p-type silicon layer grown epitaxially on a degenerately doped p+ silicon substrate roughly 500 μm thick. The epitaxial layer is typically 8–15 μm thick and is doped with boron to a nominal resistivity of 10 ohm cm. The initial step of the fabrication process consists of a cleaning procedure designed to ensure a nearly defect-free surface. Next, a thin layer of protective SiO_2 is grown on the silicon surface by using elevated temperatures in a steam and oxygen ambient. Channel stops, thin stripes that laterally contain the stored charge, are created by high energy implantation of p-type ions into appropriate regions of the silicon. Most CCDs are designed with a thin layer of n-type dopant at the silicon surface which holds the stored charge physically away from the silicon surface. This dopant layer, called the buried channel, is produced by implanting n-type dopant ions such as P or As into the silicon. After the implantation of the channel stops and buried channel, the original SiO_2 layer is removed, and a fresh, undamaged layer is SiO_2 is regrown. Next a layer of heavily impurity-doped polysilicon is deposited and patterned to form the ϕ_2 and ϕ_4 gate electrodes. An additional layer of SiO_2 is then grown to provide an effective insulating layer atop the first polysilicon layer. The ϕ_1 and ϕ_3 gate electrodes are formed from a second layer of heavily impurity-doped polysilicon. A low resistance material such as aluminum is then used to form electrical connections between appropriate parts of the CCD. Typically this same metal layer is used to form bond pads that connect the CCD to external control signals from off-chip electronics. However, some process sequences require including a second metal layer for this purpose. Finally, the CCD is covered with an overcoat of protective material such as SiO_2 or borophosphosilicate glass. This 500 to 1,000-Å layer is a barrier between the environment and the contamination-sensitive CCD.

Device Type Versus Application. The application for which the CCD is designed dictates the variants in the process that are used to provide the desired performance enhancements. Consider as an example the effect of front-side illumination of the CCD on photon collection efficiency. Applications that require very high photon collection efficiency in the visible blue, extreme UV, or X-ray spectral bands are not well served by front-side illumination of the CCD. A useful CCD variant for such applications is the back-side illuminated CCD. Back-side illuminated CCDs undergo additional processing to remove the underlying p+ substrate from the p-type epitaxial layer. As the name implies, photons impinge upon the device from the back side, thereby avoiding the absorption layers of gate electrodes present in front-side illuminated devices. In this manner, the photon collection efficiency of the device can be improved in the blue, UV, and low-energy X-ray photon regimes. Substrate is removed by using an etchant whose etch rate is highly dependent on the concentration of boron dopant in the silicon. The process of substrate removal is well understood and highly repeatable. However, thinned (to ca. 25 μm)

back-side illuminated CCDs are typically more costly than thick front-side illuminated devices. The additional cost is associated with the reduced mechanical rigidity produced by the substrate removal.

Other CCDs have special processing steps that lower the rate at which surface dark current is generated during the interval in which signal charge is collected. One such device is known as the virtual phase CCD (VPCCD) (68). The VPCCD uses a series of implanted dopant layers near the surface of the silicon to eliminate one of the polysilicon deposition steps required for electrode formation. The implanted layers serve as virtual electrodes which force the surface potential of the silicon in these regions to a constant bias during operation. These virtual electrodes replace two of the four gate phases described in the generic CCD process. The two remaining gate control phases are incorporated into a single physical gate electrode, again by ion-implanted layers in the silicon. Thus, the top side of the VPCCD pixel is only partially covered by the polysilicon gate electrode. The ion-implanted layers provide storage capability and unidirectional charge transfer, as the single polysilicon electrode phase is clocked. The virtual electrodes also greatly reduce the rate of surface-related dark current generated in the CCD. At 25 °C, the nominal dark current value of a VPCCD is 0.1 nA/cm^2, which is nearly five times lower than the dark current in a typical buried channel CCD.

A second device that has additional specialized ion-implanted layers is the multipinned phase (MPP) CCD (69). The MPP device merges the best performance features of multiphase and virtual phase CCD technologies and is becoming the CCD of choice for scientific applications. As in the generic multiphase CCD, the entire charge storage region is covered by polysilicon gates. However, before the electrode deposition, additional dopant layers are implanted into the silicon. These implanted dopant layers enable the MPP CCD to integrate the charge with the applied gate bias set, so as to attract opposite polarity charges to the Si–SiO_2 interface. This method of operation produces the lowest possible dark currents that rival or exceed the performance of the virtual phase CCD. Operation of the CCD in the MPP mode does, however, reduce the total charge storage capacity to typically 20% of that of an equivalent, non-MPP device. Although front-side illuminated devices are available, the MPP CCD is typically back-side illuminated to achieve state-of-the-art photon collection performance. Even with this stipulation, the MPP CCD is still a popular detector due to its availability in various array sizes and formats specifically designed for scientific applications.

Silicon Photodiodes

The popularity of silicon photodiodes is directly related to their ability to detect photons over a spectral range that spans the near-infrared to low-energy X-ray regimes. Typical spectral response characteristics are given in Fig. 9a. The fast response time of less than 1 μs is attractive compared to the response times of photoconductive or bolometer-based devices. The silicon photodiode has a responsivity of about 0.4 A/W at the peak response wavelength. Some manufacturers list NEP values of less than 2×10^{-15} W/Hz$^{1/2}$, D^* above 1×10^{14} cm Hz$^{1/2}$/W, and optical areas of 1–10 mm^2. The photodiode has proven to be a useful tool for photon-counting and imaging applications across the entire range of spectral sensitivity and has been used for power generation in the visible portion of the spectrum (70,71).

Silicon photodiodes are available in both discrete and array formats. The simplicity of the discrete photodiode makes this device one of the least expensive photon detectors available. Discrete photodiodes are merely p–n homojunctions and are available from numerous manufacturers. Typical, commercially available devices have one to four discrete diode detectors per package, and frequently, device manufacturers include operational amplifier-based readout electronics inside the packaged device for ease of use. Discrete photodiodes can be used in a myriad of applications, including high-speed optical switching, intensity determination for automatic exposure control circuitry in film cameras, and photon counting for spectroscopic analyses (72).

Photodiode arrays are more complex than their discrete counterparts due to the difficulty of directing the signal information from each diode to off-chip electronics. The more common linear arrays contain internal multiplexing circuitry located on the periphery of the imaging area. This circuitry amplifies and buffers the signal from each diode and presents the information from each pixel through a single-output amplifier in a controlled time sequence. The performance characteristics of linear photodiode arrays typically rival those obtained from discrete diodes. Arrays ranging in size from 1×64 pixels up to $1 \times 4{,}096$ pixels and larger are available from commercial sources. Linear photodiode arrays are commonly found in high-resolution image scanning applications such as photocopiers, facsimile (FAX) machines, and handheld scanners (73).

As in linear photodiode arrays, two-dimensional photodiode arrays require internally integrated circuitry to mediate the signal information from each pixel. However, except for the outermost rows and columns, the pixels in two-dimensional arrays are surrounded on all sides by other pixels. Thus the required circuitry cannot reside solely on the periphery of the array but must be integrated into the actual pixel site. In a two-dimensional array, each pixel consists of both a photodiode and electronic circuitry designed to attach or detach that diode from the readout electronics located at the chip periphery. The voltage pulses that control the time-sequenced readout are generated from two multiplexing circuits, one each for the x and y chip dimensions, which are also located at the chip periphery. In its simplest form, the two-dimensional photodiode array uses a large capacitor that is common to all pixels in a given row to convert the signal charge sequentially from each pixel to a voltage. The resulting signal voltage is very small and results in a signal-to-noise ratio roughly one-half that of an equivalent format CCD array. Generally device performance varies inversely with the number of pixels in the array. Prior to about 1984, two-dimensional photodiode arrays were heavily used in commercial imaging applications such as video cameras. Recently,

CCDs and metal oxide semiconductor (MOS) arrays, two-dimensional imaging arrays that use p–n diodes as photosites, and complimentary metal oxide semiconductor (CMOS)-based components for readout circuitry have emerged as strong competitors in this arena. Nevertheless, some two-dimensional standard video format photodiode arrays are still manufactured. These devices are most useful in situations unsuited for CCD and CMOS-based imagers, such as ionizing radiation environments.

In addition to use as a popular image detector, the silicon homojunction photodiode and avalanche photodiode (74) can also be used for power generation by operating the device in the photovoltaic mode. In this mode, incident photons produce a voltage drop across the device that is proportional to the number of absorbed photons. When a finite load resistance is placed across the diode leads, a current is produced. Thus, silicon homojunction diodes can be used to convert optical energy into electrical energy. The energy conversion efficiency for common silicon photocells is in the 3–15% range, depending on the specifics of the device design. High levels of illumination are required to produce useful output power. For typical photosensitive cells an illumination of 10 lux can produce an open-circuit output voltage of about 0.5 volts. However, any desired voltage or current can be generated by appropriate series or parallel interconnection of multiple individual elements. Although silicon photodiodes are not as efficient in generating power as more exotic Group 2–16 (II–VI) and Group 13–15 (III–V) materials, these devices are commonly used as power sources for both terrestrial and space-borne applications.

Fabrication. Photodiodes can be made by manufacturing processes resembling those for constructing CMOS as well as bipolar integrated circuits. For a typical p- on n-diode formation process, the starting wafer is an n-type silicon substrate roughly 500 μm thick. The silicon has been doped with either P, As, or Sb during the wafer formation process to a nominal resistivity of 10 $\Omega \cdot$ cm. The initial step of the fabrication process consists of growing a thick layer of SiO_2 on the top surface of the silicon wafer by using elevated temperatures in a steam and oxygen ambient. Next, circular windows are etched in the SiO_2 layer. The wafer is then placed in a high-temperature diffusion furnace that introduces boron dopant into the silicon through the open circular windows. Ion implantation is a common alternative to high temperature diffusion for boron doping. The result of the doping procedure is the formation of a p–n junction, where one diode is formed for every window in the SiO_2. Next, a low resistance material such as aluminum is used to form ohmic contacts to the p-type silicon regions. The aluminum layer is also used to form distinct electrodes, one per diode, to which external connections can be made. Finally a single common ohmic contact is formed on the back surface of the n-type silicon wafer by an additional metal layer.

The method of diode formation as well as the density and profile of the impurity ions determines the specific optical and electrical performance parameters of the photodiode. When photon absorption occurs in the depletion region of the diode, the resulting carriers are quickly swept from the diode and measured by the readout circuitry. The same behavior occurs for an optically induced charge formed within a diffusion length of the depletion region. A charge generated at a distance greater than one diffusion length recombines in the undepleted silicon and consequently cannot be detected as a signal charge. A similar fate befalls photogenerated carriers produced in heavily doped or heavily lattice-damaged regions. Heavy surface doping and lattice damage are common by-products of the homojunction formation process. Therefore, the diode fabrication process balances the reduction in the rate of surface dark current generation against the charge collection loss produced by heavily doping the silicon surface. Similarly, the improvement in photon collection efficiency obtained by increasing the thickness of the depletion region must be weighed against the associated increase in bulk-generated dark current.

CMOS Image Sensors

A relatively new entry in the field of visible photon detection is the complementary metal oxide semiconductor (CMOS) image sensor (75). Sometimes referred to as active pixel sensors (APS), these devices are variants of readout electronics commonly used with infrared photodiode arrays (76,77) that have been modified to allow detecting visible photons. Unlike CCD or photodiode imaging arrays, CMOS image sensors have active transistor elements that reside in each pixel. These elements are configured into amplification circuitry that can convert an integrated photogenerated charge to a signal voltage inside each pixel. Before conversion to a voltage, a signal charge is collected and stored in either a diode or MIS capacitor region. Amplifier and multiplexer circuitry residing at the periphery of the imaging area poll, each pixel in an image line in parallel, and transmit the signal voltage measurements to an on-chip analog-to-digital converter (ADC). This readout process continues until every line in the image has been polled and output.

In theory, the use of CMOS fabrication technology affords a number of advantages over CCD technology. CMOS provides a very high degree of circuit integrability, thereby allowing complete camera systems to be constructed on a single silicon chip (78). The combination of signal charge detection in the pixel of generation, without the need for transfer through the array, and the use of complementary n- and p-channel field effect transistors (FET) provide very low power operation. Finally, because CMOS fabrication is currently the most popular and most readily available semiconductor process for silicon, imagers can be manufactured at very low cost.

At this time (2000), CMOS image sensors are still in their infancy. Compared with CCDs, CMOS imagers suffer from several performance problems, including poor quantum efficiency, fixed pattern noise, and image cross talk (79). Quantum efficiency is hampered in two ways. First, including active circuitry decreases the photonsensitive area of the pixel and thereby diminishes the total collection area for photons. Second, compared to a CCD, standard CMOS processes reduce the depth

in silicon from which the signal charge is collected, thereby retarding the collection efficiency of about 600-nm and longer wavelength photons. Fixed pattern noise is introduced into the image through variations in the threshold voltages of interpixel amplifiers across the imaging array, as well as through variations in dark current generation on a pixel-by-pixel basis. Image cross talk occurs through the diffusion of photon charge generated beneath the depletion region of the pixel array. Cross talk also results from changes in pixel operating voltages across the array, which are produced by current flow through the resistance of critical bias lines that feed the pixel circuitry. Much improvement in performance is anticipated during the next few years and will be required to strengthen the prospects for CMOS imagers in scientific usage.

Cadmium Sulfide Photoconductor

CdS photoconductive films are prepared by both evaporating bulk CdS and settling fine CdS powder from aqueous or organic suspension followed by sintering (80,81). Evaporated CdS is deposited to a thickness from 100 to 600 nm on ceramic substates. The evaporated films are polycrystalline and are heated to 250 °C in oxygen at low pressure to increase photosensitivity. Copper or silver may be diffused into the films to lower the resistivity and reduce contact rectification and noise. The copper acceptor energy level is within 0.1 eV of the valence band edge. Sulfide vacancies produce donor levels, and cadmium vacancies produce deep acceptor levels.

Settling can be accomplished from an ink which contains 1-μm crystallites of CdS and selected concentrations of $CdCl_2$. The coating is fired in a restricted volume of air at 500–700 °C. During sintering, the $CdCl_2$ acts as a flux and forms a solution with the surface of the grains of CdS. A few ppm of chloride and copper (from impurities) enter the crystal lattice and act as activator centers for trapping. Excess chloride evaporates and leaves a continuous polycrystalline photoconductive film. The layers are from 5–30 μm thick and have a linear I–V relationship.

The films have an area resistance of 100 000–300 000 Ω/square. Most applications, such as switching on outdoor lights at twilight, require a detector resistance near 1,000 ohms to operate the switching circuit without the need for impedance matching electronics. This is accomplished by depositing the contacts in an interdigitated geometry as shown in Fig. 12. A protective film is deposited over the detector and contacts to provide long-term stability, or the detector structure is mounted in a hermetic package as shown. The spectral sensitivity of a thin-film CdS detector is shown in Fig. 9a. The response shape is similar to that of the human eye (see Table 1).

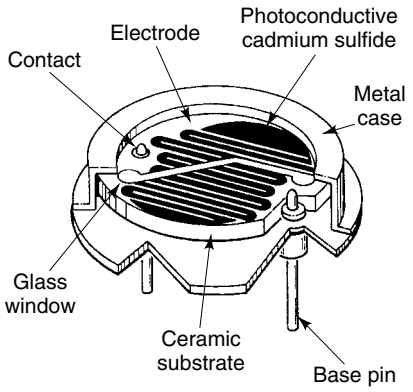


Figure 12. CdS film detector in a package showing interdigitation to reduce resistance. See text.

GaAs, GaAsP, and InGaAs Photodiodes

Gallium–arsenic and gallium–arsenic–phosphorus diodes are fabricated as photodiodes as well as light emitters. Fabrication is typically by mesa etch technology of the films of GaAsP or InGaAs grown (82) by the vapor-phase epitaxial process using metal organic chemical vapor deposition (MOCVD). This growth technique results in impurity densities less than 1×10^{14} atoms/cm^3. The spectral cutoff range extends from 500–900 nm, depending on the phosphorus content. These detectors can be used for color discrimination and do not require expensive interference filters. Emitter–diode pairs are used for very high impedance signal coupling in high-speed integrated circuits. The spectral sensitivity is shown in Fig. 9a for typical GaAs and GaAsP photodiodes. Surface leakage, caused by the lack of a native surface-passivation technology such as SiO_2 for Si detectors, is the most significant performance limitation for diodes made from this material system.

The indium–gallium–arsenide (InGaAs) photodiode is becoming a very popular detector for very near-infrared and short-wavelength infrared radiation. InGaAs devices are made by improved epitaxial growth techniques, and their performance rivals or surpasses HgCdTe. The available cutoff wavelength of commercially available discrete InGaAs photodiodes extends to 2.5 μm. Linear InGaAs arrays that have cutoff wavelengths as high as 2.3 μm (see Table 1) are available, enabling research efforts in room temperature atmospheric spectroscopy (83) and environmental monitoring (84). Uncooled InAsSbP/InGaAs photodiodes that have good performance characteristics to a 3.4 μm cutoff wavelength have also been reported (85).

PbS and PbSe Photoconductors

Lead chalcogenides, PbS, PbSe, and PbTe, were among the first infrared-detector materials investigated. Although photovoltaic effects are observed from p–n junctions in single-crystal material, the response is quite poor and not reproducible. However, very sensitive photoconductors are prepared as polycrystalline thin films about 1 μm thick, which are deposited on glass or quartz substrates between gold or graphite electrodes.

Detector elements are prepared either by sublimation in the presence of a small partial pressure of O_2 or by chemical deposition from an alkaline solution that contains a lead salt and thiourea or selenourea (86). Lead sulfide and lead selenide deposit from solutions as mirror-like coatings made up of cubic crystallites 0.2–1 μm on a side. The reaction may be represented nominally by the following:

$$Pb^{2+} + SC(NH_2)_2 + 2OH^- \rightarrow PbS + C(=NH)_2 + 2H_2O$$

The actual reaction probably is more complex. The photoconductive behavior depends on the pH of the solution from which deposition occurs. It is likely that oxygen-containing compounds are present in the deposited films. For either method of preparation, the effect of oxygen, which is introduced during preparation or by subsequent heat treatment in air or oxygen, is critical for the development of optimum sensitivity. Maximum sensitivity is obtained near the point at which the film conductivity type changes from n to p. The long response times are suggestive of trapping. It is likely that deep trapping states are located at the oxidized surface of the micrograins (87,88). The evaporation technique produces the best results, especially for PbSe and the more obsolete PbTe. The spectral sensitivity is shown in Fig. 9b for different operating temperatures (see Table 1).

Platinum Silicide Schottky Barrier Arrays

The Pt:Si detector (89–91) is essentially a metal semiconductor barrier in which the platinum silicide is a quasi-metal that generates a small energy barrier to electrons. The effective photons are absorbed in a very thin region of the silicide next to the barrier and generate free electrons that flow over the barrier and tunnel through it into the n-type silicon. The efficiency of this process is only a few percent, even at high energies (short wavelengths), because of the low electron diffusion coefficient in the silicide. At wavelengths beyond 1 μm, the efficiency drops off dramatically because of decreasing tunneling probability for the lower energy electrons. The effective quantum efficiency is about 0.1% at a wavelength of 4.8 μm.

Techniques of platinum deposition vary, but sputtering and annealing of a very thin layer of platinum produces a uniform platinum silicide film that has less than 0.3% variation in responsivity in an area of 2×2 cm. Details of deposition and annealing processes are considered trade secrets by the manufacturers. Deposition is typically onto silicon integrated circuit (IC) chips that have a readout structure. Small regions of the IC chip at each pixel are dedicated for the silicide detector element. The readout is typically an in-line charge-coupled device. The spectral sensitivity is shown in Figure 9b, and array information is given in Table 1.

Although the photon efficiency is quite low, focal plane performance for infrared imaging of ambient temperature scenes is acceptable because of long television display frame time, reasonably low detector noise, and the excellent uniformity of responsivity that allows for high on-chip input gain without offset correction. The responsivity is about 10 millivolts per degree delta scene temperature using $f/2$ optics, and scene sensitivity with large arrays is about 0.1 °C. The Pt:Si detector finds typical use in security and defense-related applications but is also useful in spectroscopy (92) and other applications where device cooling is not a significant systems issue.

InSb Photodiode Detectors and Arrays

Sensitive photodiodes (93,94) have been fabricated from single-crystal InSb using cadmium or zinc to form a p-type region in bulk n-type material. High-quality InSb crystals can be grown by the infinite-melt process (95,96) where an InSb film is grown epitaxially (from the liquid phase) on a slice of InSb which was prepared in a conventional Czochralski vertical puller. The diode formation process typically is a closed-tube diffusion. Cleaned and etched samples of InSb are placed in a quartz ampule with a limited amount of zinc or cadmium. After evacuation and sealing, the ampule is heated to about 50 °C below the crystal melting point. The metal vaporizes partially or completely, depending on the amount, volume of ampule, and temperature, and diffuses into the crystal after a few hours. The impurity–diffusion profile approximates the error–function law and the p–n junction is 1–5 μm below the surface. Diode arrays (16) are formed by etching mesas about 50×50 μm. Typical performance of InSb detectors is given in Table 1, and the spectral sensitivity is shown in Fig. 9c. Up to 480×640 matrix arrays of InSb photodiodes in a mesa configuration have been demonstrated. Commercial units 256×256 are available. The mesa detector array is mated to a silicon chip that has an array of amplifiers and multiplex circuitry. Each diode is connected to an amplifier input. The hybridization process consists of forming indium bumps on each diode mesa and on each amplifier input, using a photolithographic process and In evaporation and pressing the detector array chip to the silicon integrated circuit chip. The infrared radiation must pass through the InSb chip to reach the photodiode junction. To improve quantum efficiency, the InSb is grown as a thin layer on GaAs or GaAsSb substrates. Quantum efficiency is greater than 50% for wavelengths greater than 2 μm and less than the cutoff of InSb, 5.3 μm. A protective coating (passivation) of the InSb photodiode for stable operation during several hours without frequent signal normalizations has not been found. However, infrared imaging cameras using hybrid InSb focal planes are a commercial reality. In a real-time imaging configuration, the scene sensitivity is about 0.04 °C using an InSb infrared camera. As a result of their commercial availability, InSb arrays are finding increased use in infrared astronomy (97) and other scientific applications.

Mercury Cadmium Telluride

HgCdTe has proven to be an excellent infrared detector material (2) the CdTe content can be readily adjusted to obtain cutoff wavelengths from 2–20 μm. The benefit is high spectral sensitivity of the photon detector, low defect density, and high cooling efficiency. The dependence of energy gap on mole fraction is linear. For $Hg_{1-x}Cd_xTe$, the x values of most interest lie between 0.17 and 0.50. The need for large focal planes up to 2×2 cm has dramatically changed the direction of single-crystal HgCdTe growth technology.

Crystal Growth. The method of solid-state recrystallization (quench and anneal) was marginally adequate during the 1970s and 1980s for linear photoconductor arrays used in military night vision systems, but quenching kinetics restricted sample size to 6 mm × 2 cm. The crystals had many low-angle grain boundaries, defect densities were

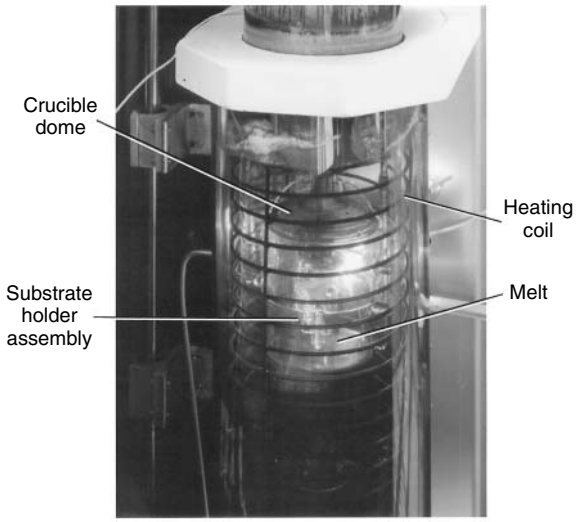


Figure 13. Furnace reactor for growing HgCdTe films on CdZnTe substrates using the liquid-phase epitaxial process. The melt is tellurium in a quartz crucible. A mercury reservoir in a cooler zone maintains the required Hg overpressure.

high, and there were large nonuniformities in composition and carrier concentration. Therefore, epitaxial growth technologies were developed. Crystal growers were hampered, however, by the lack of suitable substrate material until methods were developed to grow the HgCdTe films onto large-area silicon and GaAs single crystals. The problem with misfit dislocations has not yet been satisfactorily solved except for detector cutoff wavelengths of less than 5 μm. Thus, CdTe single crystals are used to obtain large, low defect density substrates for growth from the liquid phase. Lattice matching is achieved by adding 3% Zn at. wt.

Large-area CdZnTe substrates form the basis for liquid-phase epitaxy (LPE) growth of mid- and long wavelength IR HgCdTe detector material. Substrates are obtained from 3.5-kg CdZnTe ingots grown in graphite boats in sealed quartz ampules (98). The horizontal Bridgman growth process yields $41 \times 6.4 \times 5.0$ cm semicylindrical ingots that have large single-crystal portions which are then sectioned, sawed into slabs, and diced into required dimensions. The substrate surfaces are diamond turned and etched to assure flat, damage-free surfaces. Standard substrate sizes are 3.6×2.0 cm and 3.6×1.5 cm but the ability to obtain single-crystal ⟨111⟩ oriented CdTe substrates as large as 5.1×7.6 cm has already been demonstrated. The material typically exhibits dislocation densities of $1 \times 10^5/\text{cm}^2$ and high purity as judged by Hall measurement made on evaluation samples of n-type LPE films grown on these substrates.

Liquid-phase epitaxial films (98,99) are grown in production prototype dipping reactors, as shown in Fig. 13, using CdZnTe substrates. Film growth is from both tellurium and mercury solutions. Phase diagrams for the Te and Hg corners are shown in Fig. 14. Growth is from lightly (ca $5 \times 10^{14}/\text{cm}^3$) indium-doped 4,000-g tellurium solutions. The mercury vapor pressure is maintained by a mercury reservoir positioned in an independently controlled furnace zone. Up to 54 cm^2 of material can be grown in a single run from the largest reactors. Multiple furnace zones in the vicinity of the melt crucible permit adjusting the melt temperature profile. The substrates are positioned horizontally during growth and rotated to promote uniformity of composition. The holder design allows reorienting substrates vertically for withdrawal to facilitate melt drainage. LPE films are also grown in mercury solutions of several kilograms in which small amounts of tellurium and cadmium have been added. In both cases, the cutoff wavelength varies less than 0.1 μm across the entire film.

The films grown in tellurium are annealed in Hg vapor. Mercury atoms diffuse into the films to reduce the density of Hg vacancies that act as acceptor sites. The carrier concentration is shown in Fig. 15 as a function of the annealing temperature. For photoconductive detectors, the films are annealed so that the indium donor density is dominant, this results in n-type material that has a $1 \times 10^{15}/\text{cm}^3$ excess electron density. Minority carrier lifetime is typically 1–5 μs at 77 K. For photodiodes, the films are annealed to an acceptor (via Hg vacancies)

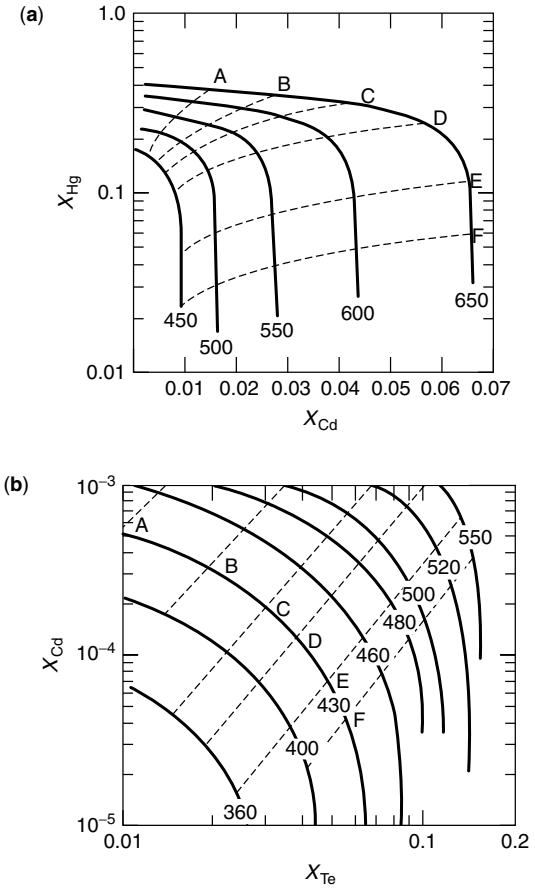


Figure 14. Phase diagrams of HgCdTe used to define the liquid-phase epitaxial growth process where composition is in mole fraction X, and the numbers represent temperatures in °C: (**a**) Te-rich corner where the dotted lines A–F correspond to values of X_{Te} of 0.1, 0.2, 0.3, 0.5, 0.8, and 0.9, respectively, and (**b**) Hg-rich corner where A–F correspond to values of X_{Hg} of 0.9, 0.8, 0.6, 0.4, 0.2, and 0.1, respectively.

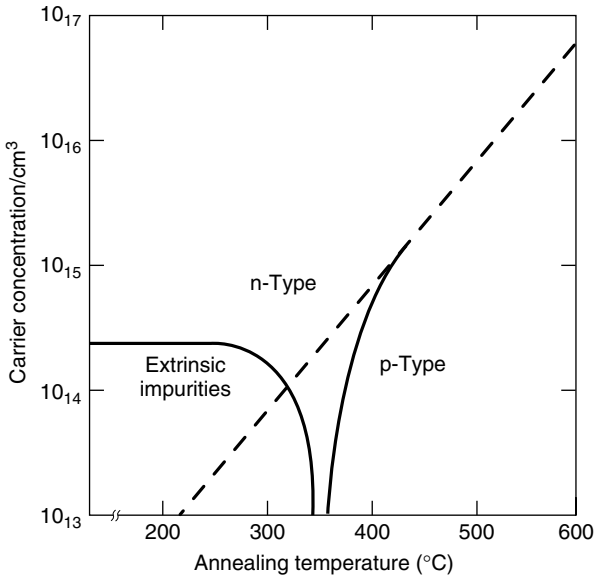


Figure 15. Excess carrier concentration in HgCdTe in a saturated Hg vapor as a function of temperature, where the dashed line represents Hg vacancies. The extrinsic impurity concentration can be adjusted in the growth process from low 10^{14} up to mid-10^{17}. Low temperature annealing reduces Hg vacancy concentration and acceptor density.

density of $3 \times 10^{16}/\text{cm}^3$. Carrier lifetimes are less than 50 ns at 77 K. Extrinsic acceptor doping is being developed to replace vacancy doping to achieve longer minority carrier lifetime and lower dark current density. Film thickness is 40–80 μm.

Films grown in Hg (99,100) are usually structured to make heterojunction photodiode arrays. The first or base layer is narrower band gap HgCdTe, grown on CdZnTe substrates, doped with indium for excess electrons (n-type) in the $3 \times 10^{16}/\text{cm}^3$ range and is 10 μm thick. The second or cap layer is wider band gap HgCdTe, doped with arsenic for excess holes (p-type) in the $5 \times 10^{15}/\text{cm}^3$ range and is 4 μm thick. The composite HgCdTe film is photolithographically etched part way into the base layer to form an array of mesas; each one is a photodiode detector element. The p–n junction is close to or coincident with the metallurgical heterojunction. For infrared detection in the 8–12 μm atmospheric spectral window, the base layer CdTe content is about 20%, and the cap layer CdTe content is about 30%.

Photoconductive Detector Arrays. Since 1972, more than 50,000 HgCdTe linear arrays (101,102) have been produced in the United States for Department of Defense infrared systems ranging from night vision for M1 tanks to targeting sights for smart weapons. These common module detector arrays consist of 180 elements photoetched on a 50-μm pitch. Virtually all of the material used for these arrays was prepared by solid-state recrystallization. Small slabs of material were epoxied to sapphire or ceramic and were thinned to 8 μm. The newer epitaxial HgCdTe is also epoxied to ceramic where the film side is down. The material is thinned by diamond turning, and the CdZnTe substrate is removed. In each case, an 8-μm thick layer of HgCdTe remains. The last few micrometers of HgCdTe removed must be carefully etched to avoid generating defects. The arrays are defined by photoetching and passivated using a ZnS film; indium electrical contacts are applied. The contacts define the optically active area of about $2 \times 10^{-5}\text{cm}^2$. The arrays are mounted in a vacuum dewar for cryogenic cooling. Each element is connected to a series resistance bias circuit and ac coupled to an amplifier. Detector resistance is nominally 100 ohms, and bias current is about 2 mA. The signal voltage is given by Eq. (13) and the responsivity by Eq. (34).

$$R_V = \eta\tau V/h\nu nv. \tag{34}$$

The responsivity becomes independent of the bias voltage V when the electric field-induced sweep time of the holes equals the hole lifetime.

g–r noise is dominant in a well-designed, well-made HgCdTe photoconductor detector (103,104) and may be expressed in terms of a minority carrier density p and majority carrier density n.

Semiconductor noise analysis for the HgCdTe photoconductor yields

$$V_{g-r}^2 = \frac{4p\tau V^2 \Delta f}{n(n+p)v}. \tag{35}$$

The responsivity and g–r noise may be analyzed to obtain background photon flux and temperature dependence of responsivity, noise, and detectivity. Typically, $n > p$, and both are determined by shallow impurity levels. The minority carrier density is the sum of thermal and optical contributions,

$$p = \frac{n_i^2}{n} + \frac{\eta\Phi_B\tau}{t}, \tag{36}$$

where, typically, the background flux ϕ_B is much greater than the signal flux ϕ_s. The lifetime τ is about 1 μs but for narrow band gap, defect-free detectors, it becomes the Auger lifetime and can be calculated readily from basic semiconductor properties (101). The cooling requirement is determined according to Eq. (36) because n_i is an exponential function of band gap energy and temperature. Combining Eqs. (35) and (36), detectivity can be expressed as follows:

$$D_\lambda^* = \frac{\eta}{2h\nu}\left(\frac{\tau}{pt}\right)^{1/2}. \tag{37}$$

At low background flux, this gives the temperature dependence of the D_λ^* shown in Fig. 4. At high flux, the D_λ^* equation (Eq. 37) reduces to Eq. (12) except for a factor of $(2)^{1/2}$ which results from the random recombination process not present in diodes. The scene sensitivity of a scanning photoconductor array infrared camera is about 0.15 °C.

The long minority carrier lifetime in n-type HgCdTe has been exploited to perform signal processing in the element (SPRITE) (105). The bias field is adjusted to sweep the photogenerated holes along the HgCdTe detector element synchronously with the scanned image, creating a time delay and integration (TDI) enhancement of the signal-to-noise ratio.

Photovoltaic Detectors, Arrays, and Focal Planes. The two popular types of photodiode arrays in HgCdTe are based on homojunction (106) and heterojunction (107,108) technologies. Homojunction diode arrays are fabricated from p-type epitaxial HgCdTe. The surface of the grown HgCdTe film is flattened by diamond milling and chemical lapping using a weak solution of bromine in methanol. The epitaxial film is then epoxied to a silicon chip that has an array of amplifiers (typically on 50-μm centers) and readout multiplexer (mux) circuitry. The CdZnTe substrate is milled away, and the HgCdTe is thinned to about 10 μm. An array of small (10–20 μm) holes is etched in the HgCdTe film so that each hole is located over an amplifier input pad of the silicon IC chip. The diodes are made by implanting boron (150 kV, $1 \times 14/cm^2$) through a 500-nm ZnS layer. Planar diodes are generated by using a patterned photoresist to define the implanted regions located between the holes. The implant process disrupts the lattice, creating Hg ion interstitials which cause the formation of very shallow donor states. The damage layer extends about 150 nm from the surface and the n–p junction varies from 2–8 μm deep, depending on the implant and annealing conditions. The fabrication is completed by applying a CdTe passivation layer (about 1 μm thick) directly to the HgCdTe surface, forming a ZnS layer for antireflection and electrical isolation, and formation of a metal film lead to connect each implanted n region to each amplifier input. Infrared illumination is directed onto the HgCdTe film, and quantum efficiency typically exceeds 75%. Because infrared focal planes are typically cooled to less than 100 K, the differential coefficient of thermal expansion causes the shrinking silicon to put tensile stress on the HgCdTe film. However, the thinness of the film and the presence of the holes allow the HgCdTe to strain and retain the integrity of the epoxy film and the electrical contacts. Array dimensions up to 2.5 cm have proven reliable.

Heterojunction diode arrays utilize the grown p–n junction and mesa etch technology. The mesa arrays are passivated by a thin layer (about 500 nm) of CdTe using an evaporation or molecular beam epitaxial process. Annealing at 250 °C in a Hg atmosphere creates a short (about 100 nm) graded layer from the CdTe to the HgCdTe. The benefit of CdTe passivation is efficient isolation of the p–n junction from the surface, low dark current noise, and immunity to ionization radiation. As in InSb hybrid arrays, indium bumps are formed using evaporation at the center of each mesa and on each amplifier contact pad of the silicon integrated circuit chip. The two chips are pressed together in a bump bonding process called hybridization (109). Sometimes the space between the chips is filled with epoxy. Infrared illumination is through the CdZnTe substrate which has been coated with ZnS for antireflection. Quantum efficiency typically exceeds 75%.

HgCdTe photodiode performance depends, for the most part, on high quantum efficiency and low dark current density (110,111), as expressed by Eqs. (23) and (25). Typical values of R_0A at 77 K are shown as a function of cutoff wavelength in Fig. 16 (95). HgCdTe diodes sensitive out to a wavelength of 10.5 μm have shown ideal diffusion current limitation down to 50 K. Values of R_0A have exceeded $1 \times 10^6\ \Omega \cdot cm^2$. Spectral sensitivities for three compositions of HgCdTe detectors are shown in Fig. 9a,c. More information is listed in Table 1. The scene sensitivity of a HgCdTe diode area array cooled to 80 K is about 0.02 °C for ambient temperature scenes.

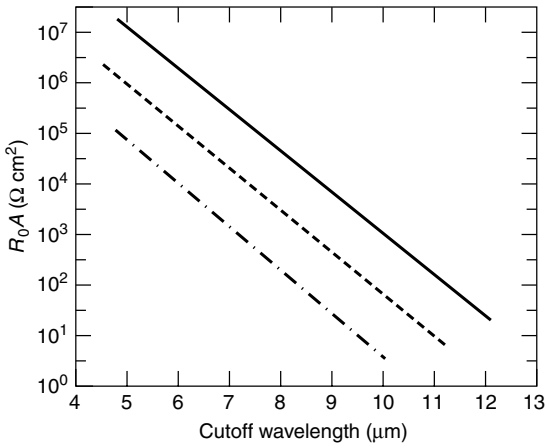


Figure 16. Resistance area (R_0A) product for HgCdTe photodiodes cooled to 77 K. The solid line represents the theoretical limit, the dashed lines (– –) and (– • –) high and low performance, respectively. Dark current caused by defects lowers R_0A and detector sensitivity. In the high-performance range, dark current is an exponential function of cutoff wavelength and temperature.

Doped Germanium and Silicon Photoconductors

Extrinsic photoconductors are typically single-crystal germanium doped with zinc, cadmium, mercury, boron, and gold (112,113) and silicon doped with indium, gallium, or arsenic (114). The doping density ranges from 1×10^{15} to 1×10^{17} impurity atoms/cm^3 leading, respectively, from a low (1 cm^{-1}) to a higher absorption coefficient (50 cm^{-1}).

Information on ionization energies, solubilities, diffusion coefficients, and solid–liquid distribution coefficients is available for many impurities from nearly all columns of the periodic table (113). Extrinsic Ge and Si have been used almost exclusively for infrared detector applications. Of the impurities, Cu, Au, Zn, Cd, Hg, and some of the elements of Groups 13 (III) and 15 (V) have been used in detectors. Germanium and silicon, which are used in the preparation of detectors, must be of high purity before they are doped with the desired activator impurity to avoid unwanted compensation by impurities whose ionization energies are smaller than the activator. The required purity can be achieved by zone refining in which a short molten zone is repeatedly passed from end to end of an ingot of impure Ge or Si. Impurities that have distribution coefficients larger than unity collect near the seed. The concentration of electrically active residual impurities in the center portion of the ingot can be reduced to 10^{12}–10^{13}/cm^3. Single crystals can be grown by using the Czochralski method, in which an oriented seed crystal is brought into contact with the melt and then is withdrawn slowly while being rotated, or by using a horizontal zone melting method, in which a seed crystal is melted onto one end of a polycrystalline ingot. A molten zone is produced at the junction of the ingot and seed and is moved slowly along

the ingot, leaving behind a single crystal. All of these operations must be carried out in an inert or reducing atmosphere to prevent oxidation of the germanium or silicon.

In most cases, the activator impurity must be incorporated during crystal growth. An appropriate amount of impurity element is dissolved in the molten Ge and, as crystal growth proceeds, enters the crystal at a concentration that depends on the magnitude of the distribution coefficient. For volatile impurities, for example, Zn, Cd, and Hg, special precautions must be taken to maintain a constant impurity concentration in the melt. Growth occurs in a sealed tube to prevent escape of the impurity vapor or in a flow system in which loss caused by vaporization from the melt is replenished from an upstream reservoir.

Some impurities, for example, Cu, Ag, Ni, Co, and Fe in Ge, and In, Ga, and As in Si, have diffusion coefficients that are large enough to permit doping by solid-state diffusion well below the melting point of the host crystal. A thin layer of the diffusant, which is deposited on the surface of the crystal by electroplating, vacuum evaporation, or electrochemical replacement, serves as the source for diffusion. After homogeneity is achieved, the sample is quenched. The alloyed surface layer must be removed by lapping and etching before the electrical contacts are applied. The impurity concentration should be as large as possible, within limits, to maximize the absorption coefficient. In some cases, the concentration is limited by the impurity solubility, which may be small depending on the element. For impurities of high solubility, the upper limit is set by the onset of impurity banding. When the average separation of impurity atoms in the lattice is small enough, conduction by direct transfer of carriers from atom to atom can occur. Impurity banding limits the extent by which cooling can reduce the dark current and, therefore, the noise, in Ge and Si. This is significant above impurity concentrations of about $10^{16}/cm^3$.

Impurities other than those from Groups 13 (III) and 15 (V) generally exhibit two or more impurity levels in Ge. If an activator is used which is not of the lowest lying level, the lower lying levels must be compensated for by adding an impurity of the opposite conductivity type. For example, if the second Zn acceptor level for 0.095 eV is to be used, the lower lying level at 0.035 eV must be compensated for by adding a donor impurity, for example, As in a concentration slightly greater than that of Zn. Electrons from the As donors fall into the low-lying Zn level and render it inactive. A special case is the lowest lying Au level; Au acts as a donor at 0.045 eV above the valence band. If a shallow acceptor, for example, Ga, is added in a concentration that is slightly less than that of Au, electrons from the donor level fall into the Ga level. At low temperatures, holes that are bound to the compensated Au centers can be photoexcited to yield photoconductivity that has a long wavelength threshold at about 25 μm. Single-detector elements and arrays are formed by dicing and etching and attaching electrical contacts. Linear arrays of several hundred elements have been made. The spectral sensitivities of doped Si and Ge detectors are shown in Fig. 9c,d. See Table 1 for other information. Monolithic and hybrid extrinsic focal planes have also been demonstrated (115,116). Such arrays play an important role in areas that require detecting very long wavelength infrared radiation and can observe at the sensitivity limits imposed by natural sky background radiation (117).

GaAs–AlGaAs Quantum Well Arrays

The quantum well infrared detector (118–120) is a technology based on the artificial structure called a superlattice. Infrared detection out to a 15-μm wavelength and beyond has been achieved using engineered material that has a controlled energy gap. The technique utilizes multiple stacked layers of semiconducting material; each has a tailored band gap energy, such that a series of potential wells exists in the direction normal to the layers. Typically, very thin layers of GaAs and AlGaAs are grown by molecular beam epitaxy (MBE). The MBE process is conducted in stainless tubing and consists of simply controlled evaporation of the elements from side chambers into a main chamber that contains a wafer of the substrate material, typically, GaAs. The electron wave function is confined in the potential wells when the photon-generated electron wavelength is nominally the layer thickness (about 10 nm). The electrons are freed from the wells by photons, and a signal is detected by applying an external bias voltage.

AlGaAs quantum well infrared photodetector (QWIP) focal planes have achieved sufficient sensitivity out to a 10-μm wavelength to result in scene temperature sensitivity of about 0.04 °C when the focal plane is cooled to 70 K (121). Spectral sensitivity is shown in Fig. 9c, and array information is given in Table 1. To date, QWIP devices have been used in a number of scientific applications, including geology (122), medicine (120), and astronomy (123), as well as in public service applications such as fire fighting.

Semiconductor Bolometer Arrays

The use of bolometers to sense infrared radiation is not new. What is different is that, rather than using a single large cell, a matrix of miniature cells or microbolometers is created. Rather than a single amplifier connected externally, a custom integrated circuit is built under each cell to form a totally integrated focal plane array, as indicated in Fig. 17. Key elements of the structure (124,125) are the thin (100-nm) amorphous silicon (α:Si) or vanadium oxide (VO) thermally sensitive membrane, the thermally insulating support arms, and the integrated circuit underlying this structure. Infrared energy focused on the individual pixels heats the membrane, as given by Eq. (24). The support arms act as electrical contacts but are sufficiently thin and narrow to prevent significant conduction of thermal energy from the membrane to the surroundings. The temperature-dependent electrical resistance of the semiconductor film is monitored by the circuitry and is converted into an electrical signal proportional to the incident radiance. This signal is displayed on a standard television monitor, thus forming the image of the scene.

Two technological advances have occurred that make such a structure feasible to build. The first is the development of microetching techniques that can be used

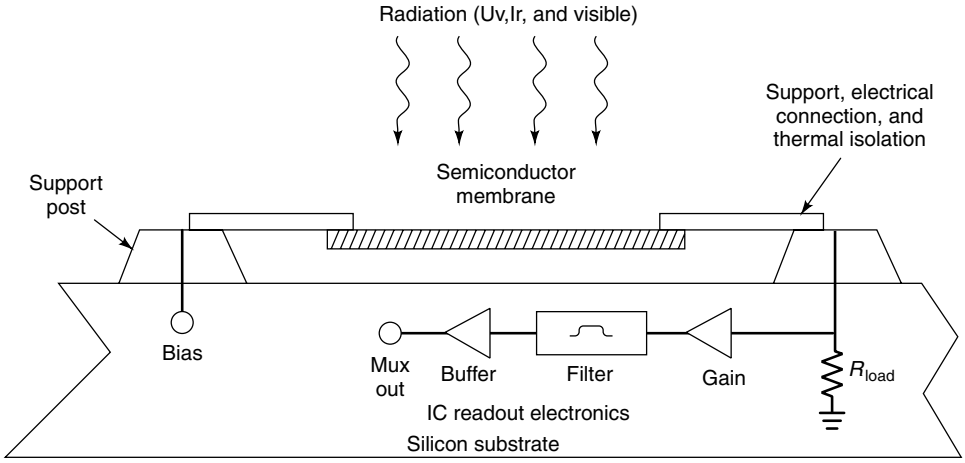


Figure 17. Cross-sectional schematic of a microbolometer photodetector. Micromachining is used to construct small, very low mass detectors. The dimension of a detector–amplifier cell is 50 μm. Mux = multiplexer, IC = integrated circuit.

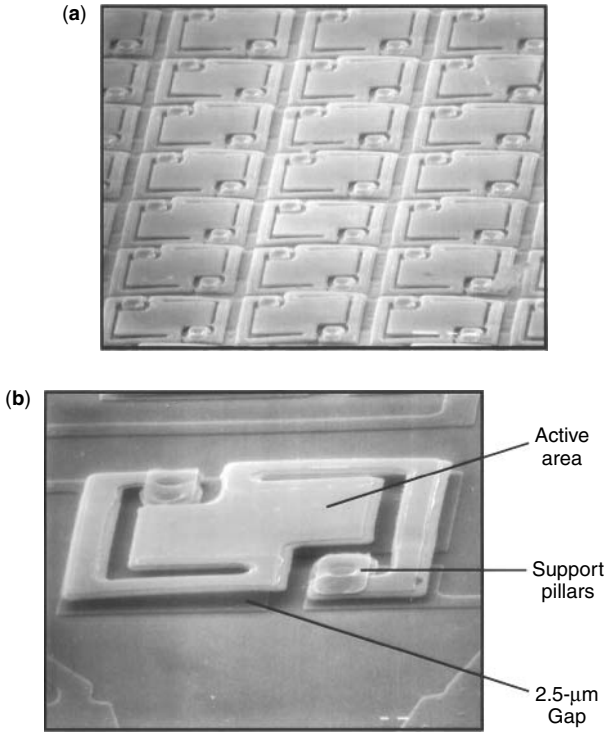


Figure 18. Microbolometer (**a**) array portion showing pixels on a 50-μm pitch. Each pixel is connected to a readout amplifier in the supporting silicon IC chip. (**b**) Detector has a 35×40 μm active area. The serpentine arms give excellent thermal isolation, and the low mass results in a 10-ms response time, ideal for thermal imaging.

to form microscopic structures in silicon and its coatings. This advance makes forming the membrane and support arms possible, and thickness control is in the 10-nm range. Cell dimensions of less than 50 μm have been demonstrated. Scanning electron microscope photos of an array and a test pixel are shown in Fig. 18. The second critical factor is the increased circuit density in silicon integrated circuits that makes it possible to build a small circuit for each cell of the array. By doing this, the noise bandwidth of each cell can be minimized, thereby maximizing performance. This device is a thermal detector in the infrared and therefore is independent of wavelength. Measured responsivity is about 5×10^5 V/W, and the NEP is about 50 pW. The focal plane operates at ambient temperature. Imaging arrays that have scene sensitivity better than 0.1 °C have been demonstrated (see Table 1).

HEALTH AND SAFETY FACTORS

The completed photodetector is usually packaged hermetically in inert glass or plastic or is enclosed in an evacuated metal or glass container. Although most detector materials are toxic, the means taken to passivate and isolate these materials are often adequate to protect the user. However there are exceptions. The preparation of detector materials and detector fabrication can present considerable hazards. Some crystal preparation techniques require using a toxic substance, for example, Hg at vapor pressures above 1 MPa (10 atm). Ampule explosions do occur. The electrical circuitry required to operate photodetectors almost always couples to detector devices at low voltages, for example, <10 V; therefore, electrical hazards are minimal. Detectors that are operated at low temperatures are mounted in evacuated containers or dewars that are partially or nearly completely made of glass. These containers are subject to implosion by improper handling.

BIBLIOGRAPHY

1. R. A. Smith, F. E. Jones, and R. P. Chasmar, *The Detection and Measurement of Infrared Radiation*, Oxford University Press, London, 1968.
2. W. L. Wolfe and G. J. Zissis, eds., *The Infrared Handbook*, rev., ed., Environmental Research Institute of Michigan, Ann Arbor, 1985.
3. P. W. Kruse, L. D. McGlauchlin, and R. B. McQuistan, *Elements of Infrared Technology*, John Wiley & Sons, Inc., New York, NY, 1962.

4. S. M. Ryvkin, *Photoelectric Effects in Semiconductors*, Consultants Bureau, NY, 1964.
5. P. W. Kruse, in R. J. Keyes, ed., *Topics in Applied Physics, Optical and Infrared Detectors*, Springer-Verlag, NY, 1980.
6. S. M. Sze, *Physics of Semiconductors*, John Wiley & Sons, Inc., New York, NY, 1969.
7. R. D. Hudson Jr., *Infrared System Engineering*, John Wiley & Sons, Inc., New York, NY, 1969.
8. J. A. Jamieson et al., *Infrared Physics and Engineering*, McGraw-Hill, NY, 1963.
9. H. L. Hackforth, *Infrared Radiation*, McGraw-Hill, NY, 1960.
10. C. S. Williams and O. A. Becklund, *Optics, A Short Course for Engineers and Scientists*, John Wiley & Sons, Inc., New York, NY, 1972.
11. M. J. Buckingham, *Noise in Electronic Devices and Systems*, Halsted Press, NY, 1983.
12. S. Borrello and Z. Celik-Butler, *Solid State Electron.* **36**, 407 (1993).
13. R. S. List, *J. Electron.* **22**(8), 1,017 (1992).
14. S. R. Borrello, C. G. Roberts, B. H. Breazeale, and G. R. Pruett, *Infrared Phys.* **11**, 225 (1971).
15. C. Hilsum, in T. S. Moss, ed., *Handbook on Semiconductors*, vol. 4, North-Holland, NY, 1981.
16. J. H. Longo et al., *IEEE ED* **25**(2), 213–232 (1978).
17. D. Long, in Ref. 5.
18. M. A. Kinch and S. R. Borrello, *Infrared Phys.* **15**, 111 (1975).
19. L. J. M. Lesser and F. L. J. Sangster, in Ref. 15.
20. F. Milton, in Ref. 5.
21. E. H. Nicollian and J. R. Brews, *MOS Physics and Technology*, John Wiley & Sons, Inc., New York, NY, 1982, Chap. 2.
22. C. Y. Wei et al., *IEEE Trans. Electron Dev.* **ED-27**(1), 170 (1970).
23. E. H. Putley, in Ref. 5.
24. A. Rogalski, in J. Thompson, ed., *SPIE Milestone Series*, SPIE Optical Engineering Press, Bellingham, WA, 1992.
25. W. D. Rogatto, in J. S. Accetta and D. L. Shumaker, eds., *The Infrared & Electro-Optical Systems Handbook*, vol. 3, SPIE Optical Engineering Press, Bellingham, WA, 1993.
26. W. S. Boyle and G. E. Smith, *BSTJ Briefs* **49**(4), 587 (1970).
27. G. F. Amelio, M. F. Tompsett, and G. E. Smith, *BSTJ Briefs* **49**(4), 593 (1970).
28. J. R. Janesick et al., *Proc. SPIE* **501**, 2 (1984).
29. J. Hynecek, *IEEE Trans. Electron. Devices* **ED-32**(8), 1,421 (1985).
30. K. A. Parulski, *IEEE Trans. Electron. Devices* **ED-32**(8), 1,381 (1985).
31. Y. Takemura et al., *IEEE Trans. Electron. Devices* **ED-32**(8), 1,402 (1985).
32. D. F. Barbe, *Proc. IEEE* **63**, 38 (1975).
33. D. E. Groom et al., *Nucl. Instrum. Methods. Phys. Res. A. (Accel.)* **442**, 216 (2000).
34. D. Morrison, *Electron. Design* **48**, 30 (2000).
35. M. M. Blouke, J. R. Janesick, and J. E. Hall, *Opt. Eng.* **22**(5), 607 (1983).
36. M. D. Gregg and D. Minniti, *Publ. Astron. Soc. Pacific* **109**, 1,062 (1997).
37. H. U. Keller et al., *J. Phys. E* **20**(6), 807 (1987).
38. K. P. Klaasen, M. C. Clary, and J. R. Janesick, *SPIE J.* **331**, 376 (1982).
39. K. P. Klassen et al., *Opt. Eng.* **38**, 1,178 (1999).
40. C. D. Murray, *J. Brit. Interplanetary Soc.* **45**(9), 359 (1992).
41. J. V. Sweedler, K. L. Ratzlaff, and M. B. Denton, eds., *Charge-Transfer Devices in Spectroscopy*, VCH, NY, 1994.
42. C. Lengacher et al., *Amer. J. Phys.* **66**, 1,025 (1998).
43. T. W. Barnard et al., *Anal. Chem.* **65**, 1,231 (1993).
44. D. M. Rowley, M. H. Harwood, R. A. Freshwater, and R. L. Jones, *J. Phys. Chem.* **100**, 3,020 (1996).
45. J. C. Sutherland et al., *J. Anal. Biochem.* **163**, 446 (1987).
46. E. Borroto et al., *Gastroenterol. Endoscopy* **50**, 684 (1999).
47. Z. Szigeti, P. Richter, and H. K. Lichtenhaler, *J. Plant Phys.* **148**, 574 (1996).
48. P. M. Epperson, R. D. Jalkain, and M. B. Denton, *Anal. Chem.* **61**, 282 (1989).
49. G. A. T. Duller, L. Botter Jensen, and B. G. Markey, *Radiat. Meas.* **27**, 91 (1997).
50. T. W. Barnard et al., *1992 Pittsburgh Conference*, New Orleans, paper no. 990.
51. T. W. Barnard et al., 1992 Pittsburgh Conference, New Orleans, paper no. 991.
52. I. B. Brenner, S. Vats, and A. T. Zander, *J. Anal. Atom. Spectrosc.* **14**, 1,231 (1999).
53. R. L. McCreery, in Ref. 37, pp. 227–279.
54. A. Wang, L. A. Haskin, and E. Cortez, *Appl. Spectrosc.* **52**(4), 477 (1998).
55. A. Wang, B. L. Jollif, and L. A. Haskin, *J. Geophys. Res.* **104**(E11), 27 067 (1999).
56. J. Hynecek, *IEEE Trans. Electron. Dev.* **ED-38**(5), 1,011 (1991).
57. J. A. Gregory, B. E. Burke, B. B. Kosicki, and R. K. Reich, *Nucl. Instrum. Methods. A.* **436**, 1 (1999).
58. P. E. Pool, R. Holtom, D. J. Burt, and A. D. Holland, *Nucl. Instrum. Methods. A.* **436**, 9 (1999).
59. S. Tsuneta et al., *Solar Phys.* **136**(1), 37 (1991).
60. D. N. Burrows et al., in O. H. W. Siegmund, ed., *EUV, X-Ray, and Gamma-Ray Instrumentation for Astronomy III, Proc. SPIE* **1,743**, 72 (1992).
61. R. D. Thom, T. L. Koch, J. D. Langan, and W. J. Parrish, *IEEE Trans. Electron. Devices* **ED-27**(1), 160 (1980).
62. M. A. Kinch et al., *Infrared Phys.* **20**, 1 (1980).
63. D. K. Schroder, *Appl. Phys. Lett.* **25**(12), 747 (1974).
64. K. Y. Han et al., *Electron. Lett.* **28**(19), 1,795 (1992).
65. M. Wadsworth et al., *IEEE Trans. Electron. Dev.* **ED 42**(2), 244 (1995).
66. W. Runyon, in *Encyclopedia of Semiconductor Technology*, M. Grayson, ed., 1984, John Wiley & Sons, Inc., New York, NY, p. 809.
67. D. J. Burt, in M. J. Howes and D. V. Morgan, eds., *Charge Coupled Devices and Systems*, John Wiley & Sons, Inc., New York, NY, 1979.
68. J. Hynecek, *IEEE Trans. Electron. Dev.* **ED-28**(5), 483 (1981).
69. J. R. Janesick et al., in *Optical Sensors and Electronic Photography*, in M. M. Blouke and D. Pophal, eds., *Proc. SPIE* **1,071**, 153 (1989).
70. R. H. Bube, in Ref. 15.
71. R. H. Bube, *Photoconductivity of Solids*, John Wiley & Sons, Inc., New York, NY, 1960.

72. A. Campion, J. Brown, and W. H. Grizzle, *Surf. Sci.* **115**, L153–(1982).
73. I. Fujieda et al., *Opt. Eng.* **38**(12), 2,093 (1999).
74. S. M. Sze, *Physics of Semiconductor Devices*, 2nd ed., John Wiley & Sons, Inc., New York, NY, 1981.
75. M. Schanz et al., *IEEE. Trans. Electron. Dev.* **ED-44**(10), 1,699 (1997).
76. N. Bluzer and A. S. Jensen, *Opt. Eng.* **26**(3), 241 (1987).
77. G. C. Bailey, *Proc. SPIE* **311**, 32 (1981).
78. B. H. Olson et al., *Proc. IEEE CCD Workshop* (1997).
79. A. J. Blanksby and M. J. Loinaz, *IEEE Trans. Electron. Dev.* **ED47**(1), 55 (2000).
80. A. B. Dreeben and R. H. Bube, *J. Electrochem. Soc.* **110**, 456 (1963).
81. J. A. Beun, R. Nitsche, and H. U. Bolsterli, *Physica* **28**, 194 (1962).
82. G. B. Stringfellow, *J. Crystal Growth* **137**, 212 (1994).
83. V. V. Doushkina et al., *Appl. Opt.* **35**, 6,115 (1996).
84. B. Matveev et al., *Sensor Actuat. B-Chem.* **51**, 233 (1998).
85. A.Krier and Y. Mao, *Infrared Phys. Technol.* **38**(7), 397 (1997).
86. E. H. Putley, in C. A. Hogarth, ed., *Materials Used in Semiconductor Devices*, John Wiley & Sons, Inc., New York, NY, 1965.
87. A. Carbone and P. Mazzetti, *Phys. Rev. B-Cond. Matt.* **57**(4), 2,454 (1998).
88. G. Giroux, *Can. J. Phys.* **41**, 1,840 (1963).
89. W. F. Kosonocky, *Proc. SPIE* **1,308**, 2 (1990).
90. B. Capone et al., *Opt. Eng.* **18**(5), 535–541 (1979).
91. M. Kimata et al., *IEEE Solid State Circuits* **SC22**(6), 1,124–1,129 (1987).
92. C. A. Gresham, D. A. Gilmore, and M. B. Denton, *Appl. Spectrosc.* **53**(10), 1,177 (1999).
93. C. Hilsum and A. C. Rose-Innes, *Semiconducting III-V Compounds*, Pergamon, NY, 1961.
94. J. T. Wimmers, R. M. Davis, C. A. Niblack, and D. S. Smith, *Proc. SPIE* **930**, 125–138 (1988).
95. S. R. Jost, V. F. Meikleham, and T. H. Myers, in *Symposia Proceedings of the Material Research Society*, R. F. C. Farrow, J. F. Schetzina, and J. T. Cheung, eds., Pittsburgh, PA, 1987, p. 429.
96. P. Mohan et al., *J.Crystal Growth* **200**(1–2), 96 (1999).
97. A. Phillips et al., *Astrophys. J.* **527**(2), 1,009 (1999).
98. C. A. Castro and J. H. Tregilgas, *J. Crystal Growth* **86**, 138–145 (1988).
99. T. Tang, M. H. Kalisher, A. P. Stevens, and P. E. Herning, in Ref. 70, p. 321.
100. G. N. Pultz et al., *J. Vacuum Sci. Technol. B* **9**(3), 1,724–1,730 (1991).
101. M. A. Kinch, S. R. Borrello, and A. Simmons, *Infrared Phys.* **17**, 127 (1977).
102. M. Itoh, H. Takigawa, and R. Ueda, *IEEE ED* **27**(1), 150–154 (1980).
103. C. T. Elliott, Infrared Detectors, in Ref. 15.
104. A. Józwikowska, K. Józwikowski, and A. Rolgalski, *Infrared Phys.* **31**(6), 543–554 (1991).
105. C. T. Elliott, D. Day, and D. J. Wilson, *Infrared Phys.* **22**, 31–42 (1982).
106. L. O. Bubulac, *J. Crystal Growth* **86**, 723–734 (1988).
107. J. P. Rosebeck, R. E. Starr, S. L. Price, and K. J. Riley, *J. Appl. Phys.* **53**(9), 6,430–6,440 (1982).
108. R. B. Bailey et al., *IEEE ED* **38**(5), 1,104–1,109 (1991).
109. J. P. Rode, *Infrared Phys.* **24**(5), 443–453 (1984).
110. R. E. DeWames et al., *J. Crystal Growth* **86**, 849–858 (1988).
111. Y. Nemirovsky, D. Rosenfeld, R. Adar, and A. Kornfeld, *J. Vac. Sci. Technol.* **7**(2), 528 (1989).
112. P. R. Norton, *Opt. Eng.* **30**(11), 1,649–1,663 (1991).
113. P. R. Bratt, in R. K. Willardson and A. C. Beer, eds., *Semiconductors and Semimetals*, vol. 12, Academic Press, NY, 1977.
114. R. D. Nelson, *Opt. Eng.* **16**(3), 275–283 (1977).
115. D. H. Pommerrenig, *Proc. SPIE* **443**, 144–150 (1984).
116. S. B. Stetson, D. B. Reynolds, M. G. Staplebroek, and R. L. Stermer, *Proc. SPIE* **686**, 48–65 (1986).
117. J. Wolf, *Opt. Eng.* **33**(5), 1,492 (1994).
118. S. D. Gunapala et al., *IEEE Trans. Electron. Dev* **ED-44**(1), 45 (1997).
119. S. D. Gunapala, B. F. Levine, L. Pfeiffer, and K. West, *J. Appl. Phys.* **69**, 6,517 (1991).
120. S. D. Gunapala et al., *IEEE Trans. Electron. Dev.* **ED-45**(1), 1,890 (1998).
121. S. D. Gunapala and S. V. Bandara, in H. C. Liu and F. Capasso, eds., *Semiconductors and Semimetals*, vol. 62, Academic Press, NY, 1999.
122. V. J. Realmuto, A. J. Sutton, and T. Elias, *J. Geo. Phys. Res* **102**, 1,5057 (1997).
123. S. D. Gunapala et al., *Proc. SPIE* **3,379**, 382 (1998).
124. Method for forming uncooled detector, Feb. 22 1994. U.S. Pat. 5288649, W. F. Keenan (to Texas Instruments Inc.).
125. R. A. Wood, in R. K. Willardson and E. R. Weber, eds., *Semiconductors and Semimetals*, vol. 47, Academic Press, NY, 1997.

PHOTOGRAPHIC COLOR DISPLAY TECHNOLOGY

BRUCE E. KAHN
Imaging and Photographic Technology
Rochester Institute of Technology
Rochester, NY

SCOPE

This article reviews color photographic output technology based on silver halide materials. The largest product market in this category is reflection print material, such as that used in color photographic paper. Color paper is used primarily for making prints from photographic negatives and recently, as a high quality output medium for digital images. A number of other types of products employ very similar emulsion technology. In addition to reflection display materials, several types of transmissive color photographic display materials use this technology. The film that is used for projecting motion pictures (motion picture "print" film) is based on this technology. Large color photographic display materials, used primarily for

advertising, are available on either clear or translucent plastic supports. These materials also use nearly identical emulsion technology to that used in photographic paper and motion picture print film. Because photographic emulsion technology is the basis for, and common among all of these different types of color photographic display products, photographic emulsion technology for display materials is the focus of this review.

There is relatively little information published in journal articles or product information on the technologies described in this report. For this reason, this article relies heavily on information published in the patent literature. Although the patent literature gives a good indication of the type of technologies available and areas of research and intellectual property interest, it does not specifically show which technologies are actually used, or in which products. As such, this article focuses on the technologies available in color photographic display materials and should not be taken to indicate or imply which technologies, compounds, or materials are used in any particular product.

INTRODUCTION

Silver halide materials have been the standard for color photographic output since they were first produced in the late 1960s. These materials provide exceptional image quality and durability at a very reasonable cost. A number of developments have been incorporated during the last 40 years to ensure that these materials remain the standard for high quality photographic output for some time to come. Silver halide color output materials continue to be used in a variety of new ways, including serving as output media for digital printing.

STRUCTURE

There are a quite a few different types of silver halide color photographic output materials. Most of these materials employ a reflective support. Motion picture print film uses very similar technology, except that the layers are coated on a transparent support for projection. There are also some special purpose display materials, which use clear or translucent supports.

Color photographic output materials are composed of a number of layers coated on a suitable support. A variety of sequences of emulsion layers can be used (1,2) but the following is the most typical.

Overcoat	1–5 μm
Red sensitive (cyan dye forming) layer	2 μm
Interlayer	1 μm
Green sensitive (magenta dye forming) layer	2 μm
Inter layer	1 μm
Blue sensitive (yellow dye forming) layer	2 μm
Polyethylene	30 μm
Paper support	160 μm
Polyethylene	30 μm
Antistatic layer	0–1 μm

The interlayers contain scavengers for oxidized developing agents. These scavengers prevent the formation of dye in one layer caused by development in an adjacent layer and thus prevent unwanted color contamination from the carryover of oxidized developing agents between layers. These scavengers are composed of ballasted hydroquinones or aminophenols. Compounds that absorb ultraviolet light are also commonly coated over the emulsion layer units to prevent unwanted dye formation and dye fade.

PAPER BASE

The most obvious difference between photographic paper and film is the support on which the silver halide emulsions are coated. There is a great deal of technology involved in making paper supports for photographic output. Although often neglected, many of the improvements in overall photographic performance have come from new support technologies.

History

Paper has been used as a photographic support since around 1800 when Thomas Wedgewood (3) experimented with paper impregnated with silver nitrate. Around 1840, William Henry Fox Talbot used paper as a support in a negative/positive system (4). The first reflective silver halide material for viewing color images was "Type I Kodacolor Paper" produced in 1942 by Eastman Kodak (5). In spite of the historical uses of paper, this initial support was actually a pigmented acetyl cellulose film (6), rather than a more conventional paper base.

Since those early days, many of the requirements for an optimum reflective photographic support have been realized. Some of the most important technical requirements of reflective photographic supports are chemical inertness (to avoid adversely affecting the photographic process), low absorption of processing chemistry, opacity, reflectivity, roughness, wet strength, emulsion adhesion, and uniformity. Physical properties, including flexibility, wet strength, dimensional stability, and freedom from curl, are also important.

Baryta

One of the earliest technologies used for color photographic supports was baryta paper. Baryta paper is of paper that is coated with barium sulfate (baryta). Barium sulfate is a commonly used white pigment. The baryta coating provided a high surface reflectance by coating a reflective pigment on the surface of the paper. Not only is barium sulfate highly reflective, it also maintains this high reflectivity fairly uniformly across the visible spectrum. This gives baryta coatings a white appearance without otherwise affecting the color or tone of the image. Barium sulfate could also be made in the correct particle size and shape to give smooth coatings and could be produced sufficiently pure so that it did not affect the photographic process (7).

Baryta coated paper had a number of drawbacks, however (6). Because the paper was highly absorbent,

it absorbed a considerable amount of photographic processing chemicals, all of which eventually needed to be washed out of the paper. This necessitated relatively long processing and drying times. To produce a glossy finish, the rough paper surface had to be pressed onto a heated polished drying drum (known as ferrotyping). Physical limitations, of the wet baryta paper, such as wet strength, dimensional stability, and curl, also limited its usefulness.

Resin Coating (RC)

The demand for color paper increased in the 1960s to such an extent that automatic printing and processing became a necessity (7). This need for automation and increased throughput spurred the development of an improved support for color paper. The most important requirement of a new support was that it would be waterproof. A number of methods were tried, including coating or impregnating paper with latex or silicone, chemical modification of the paper fibers such as partial acetylation, using synthetic fibers, and coating the paper support with various polymers (8).

These efforts to develop an improved support led to the creation of paper supports that were extrusion coated with polyethylene. Polyethylene was chosen for a number of reasons, including price and availability, chemical inertness, water impermeability, flexibility, thermal stability, and low viscosity change with temperature. Additionally, polyethylene was chosen because of the ease of extrusion coating.

One significant problem that had to be overcome before polyethylene coated papers could be used for photography was the adhesion of emulsion layers to the polyethylene. Aqueous gelatin emulsion solutions do not adhere well to the very hydrophobic polyethylene surface. Fortunately this problem was fairly easily solved by technology that was developed for the printing and packaging industries (6). The printing industry solved the problem of applying hydrophilic dyes to a hydrophobic (plastic) surface by irradiating the plastic surface with a high-voltage electrical discharge. This corona discharge treatment (CDT) works very well to promote the adhesion of gelatin emulsion layers to polyethylene surfaces. CDT oxidizes the surface of the polyethylene, affording better adhesion of the hydrophilic solution.

These extrusion-coated polyethylene paper supports are called resin coated (RC) papers. The first RC paper was introduced by Eastman Kodak in 1967–1968, and RC papers continue to be used as the primary support for silver halide based photographic color display materials. RC materials offer the advantages of reduced absorption of processing solutions, faster drying times, less usage of water and processing chemicals, energy savings, reduced curl, and the availability of a variety of surface textures.

Although the polyethylene coating prevents the absorption of solutions into the paper from the front and back sides, solutions can still absorb into the paper from the edges. Additives can be added to the paper pulp to minimize or prevent this.

The transmission of water through the paper can be reduced by adding a number of materials, such as polyethylene terephthalate, polybutyl terephthalate, acetates, polyethylene vinyl acetate, ethylene vinyl acetate, methacrylate, polyethylene methylacrylate, acrylates, acrylonitrile, polyester ketone, polyethylene acrylic acid, polychlorotrifluoroethylene, polytetrafluoroethylene, amorphous nylon, polyhydroxamide ether, and metal salts of ethylene methacrylic acid copolymers.

In addition to the use of various polymers to prevent water absorption, other polymeric materials can be used to prevent oxygen from diffusing through the layers. These materials are known as oxygen barrier layers. Polyvinyl alcohol, polyvinylidene chloride, and aliphatic polyketone polymers can be used for this purpose. Also used are polymers of acrylonitrile, alkyl acrylates or methacrylates, alkyl vinyl esters or ethers, cellulose acetates, polyesters, polyamide, and polycarbonate.

Front (Emulsion) Side. Modern RC paper supports contain a variety of other additives such as pigments, colorants, dyes, stabilizers, and optical brighteners. The print (emulsion) side usually contains a pigment to provide image sharpness, reflectivity, and opaqueness. The white pigment most commonly used is anatase or rutile TiO_2. One of the problems in using TiO_2 is that it can accelerate the photooxidative degradation of polyolefins (8). This effect can be reduced by the incorporating UV stabilizers in the TiO_2-containing layers. Hindered amine light stabilizers, such as those based upon 2,2,6,6-tetramethylpiperidine can be used for this purpose. These compounds form stable nitroxyl radicals that interfere with the photooxidation of polyolefins in the presence of oxygen. Antioxidants such as hindered phenols and organic alkyl and aryl phosphites can also be used.

Blue colorants are generally added to compensate for the slight yellow color of gelatin. Because they are generally used in the extruded polymer layer, thermally stable materials such as pigments are used. Some of these are phthalocyanines, Chromophthal, Irgazin, and Irgalite blue pigments.

Another technique for providing a blue color is the use of optical brighteners. Optical brighteners are colorless, fluorescent, organic compounds that absorb UV light and emit it as visible blue light. Examples of optical brighteners are derivatives of 4,4′-diaminostilbene-2,2′-disulfonic acid, coumarin derivatives such as 4-methyl-7-diethylaminocoumarin, 1,4-bis(*o*-cyanostyryl)benzol and 2-amino-4-methyl phenol.

Back Side. The back side (side opposite the emulsion) is composed of a polymer similar to that used on the front (emulsion) side of the paper. The back side, however, has a rough surface. This allows the back side to be written on and can improve the transport properties through processing equipment. This roughness can be achieved by adding particulate materials whose particle size is between 0.20 and 10.0 μm. Some of the materials that can be used for this are inorganic compounds such as TiO_2, SiO_2, $CaCO_3$, $BaSO_4$, Al_2O_3, kaolin, and mica. Crosslinked polymer beads or incompatible block copolymers that are mixed with the polyolefin matrix may also be used. Surface roughness can also be achieved by embossing a pattern onto the polymer backing.

The back side of the support also contains electrically conductive compounds (antistats) that reduce the chance of exposing the material by static charges. Examples of antistatic coatings include conductive salts, colloidal silica, fatty amine or ammonium compounds, and phosphate esters.

DuraLife. Recently researchers at Eastman Kodak developed a multilayer laminated paper support that has greatly improved photographic properties. They claim that this new support is "the most significant advancement in 50 years of photographic paper" (9). This new paper improves optical properties (surface smoothness, sharpness, brightness, and opacity), physical properties (tear resistance, stiffness, curl, gloss, and acceptance of writing), and process efficiency (edge penetration reduction). This support uses a composite material that consists of multilayered coextruded, biaxially oriented polyolefin sheets laminated to the top and bottom sides of a cellulose paper core. The top sheet contains a white pigment, an optical brightener, a UV stabilizer, and a unique microvoided polyolefin layer.

The following is the overall structure of this paper in a color photographic format:

Overcoat
Light-sensitive AgX imaging layers
Five layer biaxially oriented, microvoided polyolefin sheet + TiO_2, blue tint, and optical brightener
Ethylene plastomer + TiO_2
Cellulose paper base + TiO_2 + blue dye
Ethylene plastomer
Biaxially oriented polyolefin sheet
Antistatic coating

This is the substructure of the top polyolefin composite sheet upon which the emulsion layers are coated:

Layer	Contents	Thickness (μm)
L1	LDPE + blue tint	1.0
L2	Polypropylene + TiO_2	8.0
L3	Voided polypropylene	20.0
L4	Polypropylene + TiO_2	8.0
L5	Polypropylene	1.0

One very unique aspect of this support is the use of microvoided layers to provide whiteness and opacity. Layer L3 shown in the table is composed of a substructure that combines layers of alternating high and low refractive index. The high refractive index ($n = 1.49$) is due to the matrix polypropylene layers. Each polypropylene layer is 1.8–2.5 μm thick. Voids between layers give rise to low refractive index ($n = 1.0$). The voids are 0.8–1.5 μm thick. Approximately 15 such sublayers layers make up this layer. This alternating structure is more opaque than the commonly used polyethylene containing TiO_2.

This invention (10) contains an integral emulsion bonding layer, that consists of a low-density polyethylene (LDPE) skin (layer L1) on the biaxially oriented sheet. Because emulsions adhere well to LDPE, this layer eliminates the need for CDT. This layer also contains blue tints that correct for the native yellowness of the gelatin. Concentration of these tints in a thin layer, reduces the amount of the tint materials needed.

EMULSION CHARACTERISTICS

Requirements

The exact nature of the silver halide emulsions used for photographic color display materials are dictated by some of the stringent product specifications. The need for rapid processing requires emulsions that develop quickly and exert minimal restraint on development. This necessitates emulsions of low bromide and iodide content, because bromide and iodide restrain development. For environmental reasons, the quantity of effluent should be minimized, and the effluent should be environmentally benign. Photographic color display materials require very low minimum densities (D_{min}), typically less than 0.1. This requirement is the most critical for reflection print materials, where higher D_{min} is more noticeable and objectionable. High sensitivity (speed) is always desirable, particularly for the blue record of photographic color display materials. Another important requirement is that the sensitivity (photographic speed) remain constant across a wide range of exposure times. This is known as flat reciprocity. It is also very important that these photographic materials maintain their characteristics in spite of the way they are stored, even at high temperatures or humidities. Two of the most important "keeping" properties are the variations of photographic speed between manufacture and exposure of the material, known as "raw stock keeping," and between exposure and processing of the material, known as "latent image keeping." Stable dyes that are not prone to fading are also highly desirable.

One of the most stringent requirements for photographic print elements is that they possess very low minimum density. The D_{min} requirements for color paper are typically less than 0.1. Minimum density levels that are acceptable for film products would be extremely objectionable for a print material when viewed against a white reflective support. These minimum density requirements are achieved by judiciously selecting high chloride emulsions and employing antifoggants. A variety of compounds can be added during the precipitation process to reduce fog. Some compounds that are used for this purpose include hydrogen peroxide, peroxy acid salts, disulfides (11), and mercury compounds (12).

Another important requirement of color display materials is that they maintain their sensitometric characteristics as a function of time. This ability is called "keeping." There are two major types of keeping, latent image keeping and raw-stock keeping. Latent image keeping (LIK) performance is generally measured in terms of observed variations of photographic speed as a function of the time delay between imagewise exposure and processing. Raw-stock keeping is important to stabilize silver halide

emulsions against changes such as those due to storage at high temperature and humidity. Some techniques for improving the raw-stock keeping have recently been described (13).

A number of compounds have been described which can be used to reduce the amount of dye fade. A compound that has both a hindered amine and a hindered phenol in the same molecule prevented the deterioration of yellow dye images by heat, humidity, and light (14). Light fade of magenta dyes was reduced using spiroindanes (15) or chromans that had a hydroquinone diether or monoether as a substituent (16). The light stability of cyan dyes can be improved by using benzotriazole ultraviolet absorbers.

Theoretically, the exposure E of a photographic element is given by the product of the exposure intensity (I) and the exposure duration (time). This is known as the "reciprocity law." Actual photographic materials usually deviate somewhat from this ideal behavior at very long or very short exposure times. This deviation from the ideal reciprocity law is known as "reciprocity law failure." Typically, for the kinds of emulsions used in color paper, the speed and contrast decrease as the exposure intensity increases (or exposure time decreases). This is known as high-intensity reciprocity failure (HIRF). Higher exposures must be employed to compensate for reduced exposure times. One of the most important developments in color paper technology has been the discovery of iridium doping for controlling reciprocity behavior. This technology is discussed further in the section on dopants.

One of the most significant challenges in silver halide output technology is achieving sufficient sensitivity in the blue sensitive layer ("blue speed"). Blue speed is a critical factor for several reasons. Incandescent light sources are normally used to expose these types of photographic materials. The energy output of incandescent light sources decreases as wavelength decreases. In other words, there is less blue light to begin with than green or red light. The color negative film through which the light must pass before exposing the photographic printing paper usually has masking dyes that give it an orange tint. These masking dyes absorb much of the blue light. So, the small amount of blue light emitted by the incandescent light source is reduced even further by the masking dyes in the color film. As described before, at least three light sensitive emulsion layers are used to capture a photographic image. In the structure used most typically, the blue sensitive emulsion is placed below the red and green sensitive emulsion layers. Although it offers many other advantages, this layer order severely reduces the amount of light available to the blue layer because of the light scattering and absorption that occurs in the layers above. So, the amount of blue light available to expose the blue sensitive layer is limited by the incandescent light source and the masking dyes, as well as the presence of the red and green sensitive emulsion layers.

Silver Halide Precipitation, Structure, and Technology

For many reasons, the silver halide emulsions that are used for color paper are composed primarily of AgCl. Silver chloride is more soluble than silver bromide or silver iodide. Chloride ion in the development solution has less restraining action on development than bromide or iodide. For these reasons, chloride emulsions can be processed quickly, which allows photofinishers to increase their throughput. Furthermore, chloride emulsions use smaller quantities of processing solutions, which gives rise to better ecological compatibility and reduced processing effluent.

The silver halides used for color display materials are generally precipitated in the presence of ripening agents. Ripening agents (ripeners) are compounds that enhance the solubility of silver halide grains and are used primarily to increase the grain size and to make geometrically uniform grains. However, the presence of ripeners during precipitation can also affect the morphology of the silver halide grains. In general, the silver halide grains have rounded corners and edges when precipitated in the presence of ripeners. Additionally, residual ripener may be adsorbed on the surface of the emulsion grain and may not be removed during washing and filtration. This residual ripener may be carried over to subsequent emulsion preparation steps. Such remaining ripener may cause undesirable sensitization effects and create fog during storage.

Halide Conversion. The chloride emulsions used in color photographic display materials typically contain a small amount of bromide. This bromide is usually introduced following precipitation of the AgCl emulsion. This process is known as "halide conversion." Photographic emulsion performance can be a very sensitive function of the exact nature of the bromide distribution on the AgCl grains.

Silver chloride emulsions can be converted to silver chlorobromide emulsions in a variety of ways. Almost any source of bromide can be used. The three major types of bromide sources are soluble bromide (KBr), small grain (Lippmann) silver bromide (17), and organic bromides. The bromide distribution on the grains can be modified, for example by adsorbing an organic compound on the surface of the AgCl grain before introducing of the bromide source (18).

Recently, this bromide conversion process was studied dynamically using a reflection spectrophotometric technique (19). This technique has a number of advantages, such as the ability to study the conversion process *in situ*, the sensitivity to small amounts (<5 % of the amount of chloride) of bromide, and the opportunity to study reaction products and reactants simultaneously. This study showed that the halide conversion rate depends on many factors, such as the source of bromide and the presence and concentration of ripeners and restrainers. The rate of halide conversion is much faster using soluble bromide sources (KBr) than insoluble ones (Lippmann AgBr). Ripeners increase the halide conversion rate, whereas restrainers suppress it.

Iodide. In the last several years, a number of patents have been issued to Eastman Kodak on incorporating iodide into chloride emulsions. Although it can not be known for sure which, if any, of these technologies are incorporated into Kodak products, this technology seems to be important and heavily researched.

Recently, it has been shown that incorporating small amounts (0.1–0.6 mole % iodide, based on total silver within the host grains) of locally high iodide concentrations (sometimes referred to as "dump iodide") below the surface of an AgCl emulsion can greatly increase the sensitivity (speed) of the emulsion (20). This speed enhancement is primarily attributable to the intentional inclusion and specific placement of iodide within the host grains.

The disruptive effect of iodide ions on the crystal lattice is suspected of playing an important role in this increased photographic sensitivity. Because iodide ions are much larger than chloride ions, the crystal cell dimensions of silver iodide are much larger than those of silver chloride. For example, the crystal lattice constant of silver iodide is 5.0 Å compared to 3.6 Å for silver chloride. These local high iodide concentrations cause local increases in the crystal lattice size and apparently create traps that increase photographic efficiency.

Although chloroiodide emulsions are effective in improving speed, a disadvantage is that they can increase the minimum density. The production of density in unexposed areas is known as fog. It is believed that this fog is caused by iodide ion released during development that migrates to unexposed grains and disrupts the crystal lattice structure sufficiently to render these additional grains developable. This results in increased minimum density.

Chloroiodide induced fog can be reduced by blending silver iodochloride grains with small amounts of a second grain population immediately before coating (21). The second grain population is composed of very small grain (Lippmann), unsensitized silver halide that is more soluble than silver iodide. This second grain population, it is thought, acts as a "sink" for iodide liberated during development and prevents it from creating fog by reacting with the primary grain population.

As described for bromide conversion, a number of different sources of iodide can be used to introduce iodide into silver halide grains during precipitation. Iodide ion can be added as a soluble salt, such as an alkali or alkaline earth iodide salt (20–22), an iodate (23), or a triiodide salt; sparingly soluble iodine, by ripening the fine silver iodide grains of a Lippmann emulsion; by cleaving iodide ions from an organic molecule (24); or by using an amine complex of a silver iodide salt.

Iodide Source

Soluble Iodide Salts. One of the easiest ways of introducing iodide is as a soluble salt, such as an ammonium or alkali metal iodide salt. Silver and/or chloride ion may be added at the same time or following iodide addition. Soluble iodide salts facilitate adding iodide rapidly, which is required to produce the necessary local high iodide concentrations

There is a disadvantage in using soluble iodide sources, however. Soluble iodides can cause mixing sensitivity problems in large-scale precipitation of iodochloride emulsions. The rate of reaction between iodide ion and silver ion can be much faster than the rate of dispersion of the soluble iodide reactant. The dispersion rate depends on the amount of iodide added, various mixing related parameters, and the kettle volume. In these precipitations, grains that happen to impinge on the point of iodide ion introduction encounter higher iodide ion concentrations than the remainder of the grains. This problem can result in uneven distribution of the iodide ion from grain to grain and from batch to batch. The resulting silver iodochloride emulsions can vary significantly in photographic performance and can be very difficult to manufacture consistently.

Iodate **(IO_3^-).** One way that reportedly improves the control over iodide ion uniformity is to use iodate (IO_3^-) as the source of iodide (23). Iodide can be released from iodate by reacting it with a mild reducing agent, such as sulfite. The reaction of iodate with sulfite is as follows:

$$IO_3^- + 3SO_3^- \rightarrow I^- + 3SO_4^=$$

Unfortunately, sulfite is photographically active, and is known as a grain ripening agent, that is, it tends to speed the ripening out of smaller grains onto larger grains and the preferential solubilization of grain edge and corner structures. This can have the undesirable effect of changing the shape of the grains. Furthermore, the use of iodate (IO_3^-) ion to release iodide (I^-) anion is relatively inefficient and requires three sulfite anions to release a single iodide (I^-) anion. So, a relatively large amount of sulfite may be necessary, which can cause other deleterious photographic effects.

Organic Iodides (R–I). Another way to improve the control over iodide ion uniformity is to use an organic compound that releases iodide when it reacts with a nucleophile (25). The nucleophiles, however, can produce other harmful photographic effects. Some of the nucleophiles that have been used are hydroxide, sulfite, or ammonia. Hydroxide and ammonia are alkaline species that increase the emulsion pH and can produce fog. This occurs because the alkaline conditions favor reduction of Ag^+ to Ag^0. If enough Ag^0 atoms are produced in close proximity to each other, the grain can spontaneously develop, independent of its exposure. This is sometimes referred to as reduction fog or R-typing. Sulfite is a known silver halide grain ripening agent that can cause changes in grain morphology (see previous discussion).

Triiodide **(I_3^-).** Another source of iodide that appears to offers some advantages is triiodide (26). Triiodide is water soluble and provides an efficient, environmentally friendly, and convenient source of iodide from readily available materials. Triiodide does not require using additional photographically active materials for iodide release, using organic solvents, or generating organic by-products. Furthermore, triiodide generates iodide ion under mild conditions that avoid the risk of reduction fog and grain ripening.

When added to photographic emulsions, triiodide slowly releases iodine and iodide as shown in the following equation:

$$I_3^- \rightarrow I^- + I_2$$

The reaction reportedly goes to completion, driven by the constant removal of the iodide ion by incorporation in the silver halide grains.

The delay in iodide ion release allows the triiodide to be well dispersed before incorporation of iodide into the silver halide grains. This reduces the variation in iodide content between grains and results in a more robust emulsion that has improved photographic performance.

Silver Halide Amine Complexes **($[Me_2NH_2]_n[AgICl_n]$).** Similar to the use of Lippmann AgI as a silver source, silver halide amine complexes of the type $[Me_2NH_2]_n[AgICl_n]$ have been used as single source materials for incorporating iodide into silver chloride (27). Iodide is incorporated by introducing the precursor material into an aqueous medium of silver chloride crystals, where iodide is released from the complex as silver iodide.

Doping. One of the most actively patented silver halide technologies is the use of dopants in silver halides. The dopant related patent activity has been intense and has been conducted by almost all photographic manufacturers. Dopants can modify almost any photographic property and can be used in extremely low amounts (as low as a few atoms per silver halide grain!). Display materials, in particular, have benefited from dopant technology. Once again, it is impossible to know exactly which of these dopants is incorporated in any particular product. There seems to be little doubt, however, that dopant chemistry is one of the most important areas of research in silver halide technology, particularly in display materials.

A dopant has been defined as an ion or complex that is incorporated in the silver halide lattice in a dispersed manner (28). This definition excludes compounds that exist in separate phases or are adsorbed onto the surface of silver halides.

Initially, it was thought that dopants operate simply by replacing a single silver ion by an ion of another metal. This viewpoint evolved to one where a hexacoordinated anionic dopant complex $(MX_{6-n}L_n)^{m-}$ replaced the silver halide lattice unit $(AgX_6)^{5-}$. Recently, it has been shown that very large, complicated organic ligands can also be incorporated intact in the silver halide lattice as part of a transition-metal dopant complex. To accommodate such large ligands, it is necessary to consider replacing even larger portions of the silver halide lattice $(Ag_nX_m)^{p-}$ by the organometallic dopant complex. This technology has given rise to a great deal of control over the electronic properties of the dopant complex.

Dopants can be incorporated in a silver halide lattice in a number of ways. Most commonly, a water soluble solution of the dopant is added during the precipitation process. Another convenient method of incorporating dopants into a AgCl emulsion is to dope a small grain (Lippmann) AgBr emulsion and use this to effect both the halide conversion and doping of the AgCl substrate. In this manner, both the dopant and the silver bromide can be added to the host AgCl grain in a single step.

Iridium (Ir). Group 8–10 hexahalo coordination complexes create deep electron traps in silver halides (29–31). Deep electron traps trap electrons permanently or for relatively long periods of time. Deep electron traps are useful for controlling reciprocity failure. Most notable among these metals is iridium, which has been used extensively as a dopant to control the reciprocity behavior of photographic emulsions (32,33). Iridium hexachloride complexes are commonly used as dopants to reduce reciprocity law failure in silver halide emulsions. Many iridium dopants are effective for this purpose (34). Although iridium is incorporated in the silver halide grain as Ir^{3+}, Ir^{4+} complexes are preferred because they are more stable (35).

Organic Ligands. In addition to simple metal ions, organometallic complexes can be incorporated in silver halide grains (36) to produce photographic performance modifications not possible by incorporating the transition-metal ion alone. It is also possible to incorporate organometallic complexes that have large organic moieties (37) and obtain photographic performance that can be attributed specifically to the presence of the organic ligand or ligands. Surprisingly, the selection of organic ligands in these complexes is not limited by steric considerations, long considered important for dopant incorporation in the silver halide lattice. For practical reasons, the organometallic complex should be thermally and structurally stable and exhibit at least very slight water solubility. This technology provides an excellent opportunity to tailor photographic performance to meet specific requirements.

The organic ligands used in these complexes can be selected from a wide range of organic families, including substituted and unsubstituted aliphatic and aromatic hydrocarbons, secondary and tertiary amines (including diamines and hydrazines), phosphines, amides (including hydrazides), imides, nitriles, aldehydes, ketones, organic acids (including free acids, salts, and esters), sulfoxides, and aliphatic and aromatic heterocycles containing chalcogen and pnictide (particularly nitrogen) atoms in the heterocyclic ring.

These organometallic dopant complexes can be effective in extremely small concentrations. Depending on the specific dopant and intended use, concentrations of 10^{-9}–10^{-4} moles of dopant per mole of silver can be used.

Applications

Reduced Dye Desensitization. It has been shown that dye desensitization can be reduced when silver halide grains are doped with group 8 organometallic complexes that containing very electron withdrawing ligands, such as pseudohalide (CN, SCN, SeCN) ligands (37). These complexes are effective in concentrations of 10^{-8}–10^{-5} moles of dopant per mole of silver.

Speed Reduction. An interesting application of organometallic dopants reduces photographic speed with minimal or no alteration in photographic contrast. Organometallic cobalt complexes that have electron withdrawing ligands are useful for this purpose (37). These complexes are effective in concentrations of 10^{-9}–10^{-6} moles of dopant per mole of silver.

Reciprocity, Heat Sensitivity, and Latent Image Keeping. Aromatic heterocyclic organometallic complexes of the forms shown following can be used to reduce high intensity reciprocity failure (HIRF), reduce low intensity reciprocity failure (LIRF), reduce thermal sensitivity, and improve latent image keeping (LIK) simultaneously. The preferred heterocycles for this application are azoles and azines, specifically thiazoles, thiazolines, and pyrazines. These complexes are effective in concentrations of $10^{-9}-10^{-3}$ moles of dopant per mole of silver:

$$[MX_x(H_2O)_yL_z]$$

M is a group 8 or 9 metal (preferably iron, ruthenium, or iridium),

X is a halide or pseudohalide (preferably Cl, Br, or CN)

$x = 3-5$,

$y = 0, 1$

L is an organic ligand

$z = 1, 2$

Bimetallic complexes that have bridging organic ligands can also be used:

$$[MX_5LM'X_5]$$

M and M′ are group 8 or 9 metals (preferably iron, ruthenium, or iridium, or any combination)

X is a halide or pseudohalide (preferably Cl, Br, or CN)

L is a bridging organic ligand, particularly those containing nitrogens, such as pyrazine, bipyridine, and phenanthroline

Reciprocity. The organometallic thiazole complexes following are particularly useful for controlling reciprocity at very short exposure times, such as those used in digital printing (38).

$$[IrX_{6-m}L_m]^{-3-m}$$

$m = 0-1$

X is a halide or pseudohalide (preferably Cl, Br, or CN)

L = thiazole

Contrast-Increasing Dopants (NZ). Rhodium hexahalides are well known and widely employed dopants for increasing photographic contrast. These dopants are generally used in concentration ranges of $10^{-6}-10^{-4}$ moles of rhodium per mole of silver. Substituting an organic ligand for one or two of the halide ligands in rhodium hexahalide can result in a more stable complex (37).

Group 8 nitrosyl or thionitrosyl complexes are also useful for increasing contrast (40–41). They are optimally incorporated below the surface of the grain to avoid reducing the grain sensitivity.

Shallow Electron Trapping (SET) Dopants. Some dopants can be used to increase the sensitivity of photographic emulsions. This was first reported by Keevert et al. using an organometallic complex containing cyano ligands as a dopant in high chloride grains (39). A general class of dopants that can be used for this purpose are those that provide shallow electron trapping sites. A shallow electron trap is one that holds photoelectrons only temporarily, as opposed to "deep" electron traps, which trap electrons permanently. Shallow electron trapping dopants require a net valence that is more positive than +1 (the valence of the silver ion they replace), a filled highest occupied molecular orbital (HOMO), and a lowest occupied molecular orbital (LUMO) higher in energy than the lowest lying AgX conduction band (40–41). Many such complexes have been described (41) and are effective in concentrations ranging from 10^{-7} to 10^{-4} moles of dopant per mole of silver.

Shallow electron traps improve photographic sensitivity by reducing some of the inefficiencies inherent in silver halides. It is thought that the dopant lowers the energy of the conduction band at the site of the dopant and temporarily attracts photoelectrons at the dopant site. This localization of electrons at the trap sites reduces the chances of various electron loss and recombination processes, thus improving photographic efficiency.

Sensitization

Chemical Sensitization. To increase their photographic efficiency, silver chloride emulsions used for photographic color display materials can be sensitized by three different methods, either individually or in combination with each other (42). Sulfur sensitization employs sulfur-containing compounds such as thiosulfates, thioureas, or mercapto compounds that can react with silver. Generally, sulfur sensitization is conducted in the presence of a gold-containing compound. Sulfur and gold may also be contained in the same compound. An example of this used by Eastman Kodak is colloidal aurous sulfide (43). Reduction sensitization uses reducing agents, such as stannous salts, amines, or borane derivatives (44).

Chemical sensitization may also be conducted in the presence of other compounds such as thiocyanates (45), thioethers (46), azaindenes, azapyridazines, and azapyrimidines (47). Chemical sensitization can take place in the presence of spectral sensitizing dyes (48) or can be directed to specific sites or crystallographic faces on the silver halide grain (49).

Spectral Sensitization. Intrinsically, silver chloride emulsions absorb very little visible light. Spectral sensitizing dyes need to be added to silver halides to impart visible light sensitivity. This process is known as spectral sensitization. A large number of dyes have been developed for this purpose (50), and generally vary widely from manufacturer to manufacturer and product to product. Dyes that have sensitizing maxima at wavelengths throughout the visible and infrared spectrum and have a great variety of spectral sensitivity curve shapes are known. The choice and relative proportions of dyes depends on the region of the spectrum to which sensitivity is desired and on the shape of the spectral sensitivity curve desired. Most typically, polymethine dyes are used, particularly cyanines and merocyanines.

Normally, several sensitizing dyes are used, at least one per color record. It is also possible to use combinations

of dyes in a single color record, where each has a different absorption maximum, to achieve spectral sensitivity whose maximum is between the sensitizing maxima of the individual dyes.

Combinations of spectral sensitizing dyes that result in a phenomenon known as supersensitization can also be used (51). Supersensitization occurs when the spectral sensitization of a combination of dyes is greater in some spectral region than that from any concentration of one of the dyes alone or what would be predicted from the additive effect of the dyes. Supersensitization can be achieved by using selected combinations of spectral sensitizing dyes and other addenda such as stabilizers and antifoggants, development accelerators or inhibitors, coating aids, brighteners, and antistatic agents.

Spectral sensitizing dyes can be added at any stage during the preparation of the emulsion. They may be added at the beginning of or during precipitation (52), before or during chemical sensitization (53), before or during emulsion washing (54), or can be mixed in directly before coating (55). Small amounts of iodide can be adsorbed on emulsion grains to promote aggregation and adsorption of spectral sensitizing dyes (56). Depending on their solubility, spectral sensitizing dyes can be added to an emulsion as a solution in water or in such solvents as methanol, ethanol, acetone, or pyridine; dissolved in surfactant solution (57); or as a dispersion (58).

Couplers. Compounds added to an emulsion layer form the image dye when they react during processing with oxidized developer solution. These image dye forming compounds are known as couplers. Couplers are defined as 4-equivalent or 2-equivalent depending on the number of atoms of Ag^+ required to form one molecule of dye. Two-equivalent couplers are preferred because they reduce the amount of silver required. Couplers have different chemical structures, depending on the color image dye [cyan (59), magenta (60), or yellow (61)] they are to produce. As for sensitizing dyes, the couplers used vary considerably from manufacturer to manufacturer and product to product.

When a coupler reacts with oxidized developer, a portion of the coupler molecule is released. This fragment is known as a coupling-off group. Because this group is released only at the locations where there is an image, it can provide a number of useful functions (62). After being released from the coupler, coupling-off groups can accelerate or inhibit development or bleaching, influence dye hue or formation, accomplish color correction, or facilitate electron transfer. Coupling-off groups can modify the reactivity of the coupler and can affect either the layer containing the coupler or other layers in the photographic material.

Antifoggants. As mentioned before, one of the most important requirements for photographic color display materials is that they provide very low D_{min} and avoid fog. Many compounds can be added to stabilize the photographic characteristics and to help prevent the generation of fog during manufacture, storage, or processing (63). These compounds are known as antifoggants or stabilizers. Examples of these types of compounds are azolium salts (64), heterocyclic mercapto compounds, mercaptotetrazoles (65) (particularly 1-phenyl-5-mercaptotetrazoles and substituted derivatives thereof), azaindenes such as tetraazaindenes (66) [especially 4-hydroxy-substituted (1,3,3a,7)tetraazaindenes], dichalcogenides (67), benzenethiosulfonic acids, and benzenesulfinic acids. Antifoggants can be added to an emulsion at many stages of the manufacturing process. They are usually added just before coating or during chemical or spectral sensitization.

Similarly, compounds can be added during the precipitation process to prevent fog formation. These agents include hydrogen peroxide, peroxy acid salts, disulfides (68), mercury compounds (69), and *p*-quinone (70).

Other Addenda. In addition to the compounds discussed previously, many other addenda can be added to emulsions designed for photographic color display materials. These addenda can be used for a variety of purposes, including whitening, development acceleration, antihalation, and gelatin hardening.

Whitening agents such as stilbenes, triazines, oxazoles, or coumarins can be added to photographic emulsions. These whitening agents may be contained in any layer. If the whitening agent is not water-soluble, a dispersion of it may be used.

Compounds that increase the rate of development can be added to a photographic material. Benzyl alcohol was commonly used as a development accelerator, but its use has been decreasing for environmental reasons. Other compounds that can be used are cationic compounds and neutral nitrate salts (71), polyethylene glycols, and thioethers (72).

Compounds can be added to increase the hardness of gelatin to improve the physical properties of a photographic color material. Bisvinylsulfonyl compounds are particularly useful as hardeners. Other compounds, known as "fast acting" hardeners (73), can also be used for this purpose.

DIGITAL

One of the fastest growing markets for silver halide photographic output materials is that for producing hard copy images from digital images. This includes photofinishing systems that scan (and manipulate) processed film, then write the images digitally onto silver halide materials. Many of the basic principles of electronic printing have been discussed by Hunt (74).

Some of the advantages of color silver halide materials for digital photographic output are the image quality, the look and feel of the material, the permanence and durability of the images, the large installed base of processing equipment, high productivity, and low media cost. Silver halide materials can record continuous-tone, very high resolution images at very low cost.

Digital output onto silver halide materials provides all of the advantages inherent in silver halide materials

and the additional capabilities of digital image processing enhancements such as improvements in image structure and defect removal, in addition to the more conventional balance and tone scale modifications. In digital photofinishing, each pixel can be manipulated and optimized throughout the process. This allows using the full latitude for all media, and can even correct some normal failures encountered in photography. Based on the growth in digital imaging and the advantages provided by silver halide based output materials, the future for these materials looks very bright.

Silver halide materials have been used as output materials for a number of years. Initially, the graphic arts market exposed silver halide materials by using lasers for proofing or making digital separations. Film recorders, which used digital exposing onto either positive or negative photographic materials, have also been available for a number of years. Most of the early technologies for exposing silver halide materials used CRTs to generate the exposures. Other digital exposing technologies are based on LEDs, LCDs, or lasers (gas or solid state). In 1987, Kodak introduced a monochrome laser printer for the health imaging industry (75). Kodak also developed a device in the late 1980s which used a trio of a red, green, and blue lasers to expose silver halide materials.

Although there have been examples of digital exposures onto silver halide materials, it has only been recently that color silver halide materials have been specifically designed for digital exposure.

Many of the properties of digital exposing devices necessitate developing an output material which is specifically designed for that purpose, rather than using materials intended for optical printing. Some of these critical digital exposing device-dependent properties are spectral energy output, flare, fringing, exposure time, exposure overlap, the number of pixels or colors exposed at a time, and multiple exposures. Other non-device-dependent factors driving the development of color silver halide materials for digital output are the maximum image size available, color gamut, resolution (spatial and color), printing speed, media cost, and optimization of the output materials for both text and image display.

To generate optimal print quality from digital exposure devices, output media must satisfy a number of requirements. Both the exposure (power) required to achieve the maximum density and the spectral sensitivity of the media must be designed to match the exposing device. The exposure times used by digital exposure devices can be extremely short (as short as nanoseconds), necessitating carefully optimized reciprocity control at these times. The media must be able to generate sufficient contrast for all three color records at the exposure times used. Minimum density and color gamut are also important media parameters.

The contrast of the media is a very important parameter. The contrast needs to be large enough that the maximum density can be achieved at approximately 1.2 log E more than the energy required to get just above $D_{\min}$. Otherwise, so much energy is needed to achieve the high densities that scattering of the light within media causes flare or fringing. These problems can be seen visually in images as blurring or fuzzy text, particularly in dark text on a light background or vice versa. However, if the contrast is too high, amplification of digital artifacts, such as lines or bands in the writing direction caused by nonuniformities of the exposing device, can be seen. High contrast can also give rise to density jumps or image contouring, caused by insufficient digital resolution to generate smooth density changes. It is also important that the contrast of the three color records be appropriately matched; otherwise, color flare or fringing can result, which is much more objectionable than neutral flare or fringing. The lowest contrast color record will be the first to display flare or fringing.

The spectral sensitivity of the media and the spectral distribution of the exposing device also require careful matching. If the spectral overlap is poor, color contamination and color gamut reduction will occur. Productivity loss will also occur, requiring higher energies, which may cause flare or fringing.

Photographic papers for digital exposures also require that there be no chemical cross talk between layers. This means that the oxidized developer from one layer cannot be allowed to migrate to adjacent layers, thus forming image dye where it is not desired. This problem is greatest at maximum densities, where the highest concentration of oxidized developer is generated.

Another approach is to design materials that are compatible with both optical and digital exposing devices. The ability of a material to be used from nanoseconds to hundreds of seconds is truly impressive. Konica has been able to make emulsions that have these reciprocity characteristics by using transition-metal doping, optimization of chemical sensitization, and new emulsion stabilizer technology (76). Fluorescent brighteners are used to achieve low $D_{\min}$. Fast developing emulsions and improved couplers are used to achieve high $D_{\max}$. Konica has been supplying such materials since 1993 when they released Konica Color QA paper. Their newer materials employ high contrast in the middle- to high-density regions and low contrast in the low-density region.

DYE BLEACH (ILFOCHROME)

Dye bleach technology is another method of forming color photographic display images, particularly from positive originals. This technology is used, for example, in Ilfochrome materials. In the dye bleach process, preformed dyes (as opposed to those formed by coupling in the development process) are destroyed in an imagewise process (77).

BIBLIOGRAPHY

1. Research Disclosure, Item 36544, Section XI. Layers and layer arrangements.
2. *Digital imaging with high chloride emulsions*, U.S. Pat. 5783373, 1998, J. Z. Mydlarz, J. A. Budz, and E. L. Bell (to Eastman Kodak Company).
3. H. Gernsheim, *The History of Photography*, Oxford University Press, London, New York, 1955, p. 32.
4. H. Gernsheim, *Endeavour* **1**, 19 (1977).

5. R. L. Heidke, L. H. Feldman, and C. C. Bard, *J. Imaging Sci. Technol.*, **11**, 3 (1985).
6. G. Kolf, *Br. J. Photogr.* **127**(13) 296–299 (1980).
7. A. I. Woodward, *J. Appl. Photogr. Eng.* **7**(4), 117–120 (1981).
8. T. F. Parsons, G. G. Gray, and I. H. Crawford, *J. Appl. Photogr. Eng.* **5**(2), 110–117 (1979).
9. R. Bourdelais, T. Giarrusso, *IS&T 11th Int. Symp. Photofinishing Technol.*, 2000, pp. 49–56.
10. *Superior photographic elements including biaxially oriented polyolefin sheets*, U.S. Pat. 6030742, 2000, R. P. Bourdelais et al., (to Eastman Kodak Company).
11. *Photographic emulsion stabilized with bis(p-acylamidophenyl)disulfides*, U.S. Pat. 3397986, 1968, A. G. Millikan and A. H. Herz (to Eastman Kodak Company).
12. *Photographic emulsions containing mercury salts*, U.S. Pat. 2728664, 1955, B. H. Carroll and T. F. Murray (to Eastman Kodak Company).
13. *Light-sensitive silver halide photographic material*, U.S. Pat. 5116723, 1992, M. Kajiwara and H. Ono (to Konica Corporation); *Silver halide emulsions with improved heat stability*, U.S. Pat. 5670307, 1997, R. Lok (to Eastman Kodak Company); *Silver halide photographic elements containing dioxide compounds a stabilizers*, U.S. Pat. 5693460, 1997, R. Lok (to Eastman Kodak Company); *Combination of dithiolone dioxides with gold sensitizers in AgCl photographic elements*, U.S. Pat. 5756278, 1998, R. Lok (to Eastman Kodak Company).
14. *Recording material for color photography*, U.S. Pat. 4268593, 1981, D. G. Leppard and M. Rasberger (to Ciba-Geigy AG).
15. *Japanese Pat. Appl. (OPI)* 159,644/81.
16. *Japanese Pat. Appl. (OPI)* 89,835/80.
17. Method of manufacturing silver halide emulsion, U.S. Pat. 5254456, 1993, S. Yamashita, S. Takada, and S. Shibayama (to Fuji Photo Film Co., Ltd.).
18. *Photographic light-sensitive material and method of developing the same*, U.S. Pat. 4865962, 1989, K. Hasebe et al., (to Fuji Photo Film Co., Ltd.).
19. B. E. Kahn, et al., IS&T 2000 International Symposium on Silver Halide Technology, 20; B. E. Kahn, et al., *J. Imaging Sci. Technol.*, in press.
20. *Photographic print elements containing cubical grain silver iodochloride emulsions*, U.S. Pat. 5726005, 1998, B. T. K. Chen, J. L. Edwards, R. Lok, and S. H. Ehrlich (to Eastman Kodak Company).
21. *Photographic print elements containing cubical grain silver iodochloride emulsions*, U.S. Pat. 5728516, 1998, J. L. Edwards, B. T. Chen, and S. H. Ehrlich (to Eastman Kodak Company).
22. *Cubical grain silver iodochloride emulsions and processes for their preparation*, U.S. Pat. 5736310, 1998, B. T. K. Chen, J. L. Edwards, R. Lok, and S. H. Ehrlich (to Eastman Kodak Company); *Composite silver halide grains and processes for their preparation*, U.S. Pat. 5792601, 1998, J. L. Edwards, B. T. Chen, E. L. Bell, and R. Lok (to Eastman Kodak Company).
23. *Process for the preparation of silver halide emulsions having iodide containing grains*, U.S. Pat. 5736312, 1998, S. Jagannathan et al., (to Eastman Kodak Company).
24. *Silver halide photographic light-sensitive material*, U.S. Pat. 5389508, 1995, S. Takada, M. Yagihara, H. Okamura, H. Kawamoto, and M. Kikuchi (to Fuji Photo Film Co., Ltd.).
25. *Silver halide photographic light-sensitive material*, U.S. Pat. 5389508, 1995, S. Takada, M. Yagihara, H. Okamura, H. Kawamoto, and M. Kikuchi (to Fuji Photo Film Co., Ltd.); *Silver halide photographic emulsion and a photographic light-sensitive material*, U.S. Pat. 5418124, 1995, Y. Suga, M. Kikuchi, M. Yagihara, H. Okamura, and H. Kawamoto (to Fuji Photo Film Co. Ltd.); *Silver halide photographic emulsion and light-sensitive material using the same*, U.S. Pat. 5525460, 1996, Y. Maruyama, M. Yagihara, H. Okamura, H. Kawamoto, and M. Kikuchi (to Fuji Photo Film Co., Ltd.); *Method of preparing silver halide photographic emulsion, emulsion, and light-sensitive material*, U.S. Pat. 5527664, 1996, M. Kikuchi, M. Yagihara, H. Okamura, and H. Kawamoto (to Fuji Photo Film Co., Ltd.).
26. *Preparation of silver chloride emulsions having iodide containing grains*, U.S. Pat. 6033842, 2000, R. Lok, B. T. Chen, and W. W. White (to Eastman Kodak Company).
27. Preparation and use of a dimethylamine silver chloroiodide complex as a single source precursor for iodide incorporation of silver chloride crystals, U.S. Pat. 5866314, 1999, T. L. Royster, Jr., S. Jagannathan, J. A. Budz, and J. Z. Mydlarz (to Eastman Kodak Company).
28. M. T. Olm, R. S. Eachus, W. G. McDugle, and R. C. Baetzold, *Int. Symp. Silver Halide Technol.*, 2000, pp. 66–70.
29. R. S. Eachus, R. E. Graves, and M. T. Olm, *J. Chem. Phys.* **69**, 4580–4587 (1978).
30. R. S. Eachus, R. E. Graves and M. T. Olm, *Physica Status Solidi A* **57**, 429–38 (1980).
31. R. S. Eachus and M. T. Olm, *Annu. Rep. Prog. Chem. Sect. C., Phys. Chem.* **83**(3), 3–48 (1986).
32. B. H. Carroll, *Photogr. Sci. Engi.* **24**(6), 265–267 (1980).
33. Asami; *Silver halide photographic emulsion containing a gold salt and a polyalkylene oxide*, U.S. Pat. 3901711, 1975, K. E. Iwaosa and I. N. Seigo (to Mitsubishi Paper Mills, Ltd., JA); *High contrast scanner photographic elements employing ruthenium and iridium dopants*, U.S. Pat. 4828962, 1989, N. E. Grzeskowiak and K. A. Penfound (to Minnesota Mining and Manufacturing Company); *Silver halide emulsions having improved low intensity reciprocity characteristics and processes of preparing them*, U.S. Pat. 4997751, 1991, S. H. Kim (to Eastman Kodak Company).; *Silver halide photographic emulsion prepared with silver halide grains formed in the presence of a water soluble iridium compound and a nitrogen-containing heterocyclic compound*, U.S. Pat. 5134060, 1992, H. Maekawa, M. Kajiwara, M. Miyoshi, and M. Okumura (to Konica Corporation).; *Selenium and iridium doped emulsions with improved properties*, U.S. Pat. 5164292, 1992, B. R. Johnson and P. J. Wightman (to Eastman Kodak Company).; *Silver halide emulsion and photographic material using same*, U.S. Pat. 5166044, 1992, M. Asami (to Fuji Photo Film Co., Ltd.).; *Silver halide photographic material which contains an iron dopant and substantially no silver iodide*, U.S. Pat. 5204234, 1993, M. Asami (to Fuji Photo Film Co., Ltd.).
34. *Photographic emulsion containing transition metal complexes*, U.S. Pat. 5474888, 1995, E. L. Bell (to Eastman Kodak Company).; *Photographic emulsion containing transition metal complexes*, U.S. Pat. 5480771, 1996, E. L. Bell (to Eastman Kodak Company).; *Photographic emulsion containing transition metal complexes*, U.S. Pat. 5500335, 1996, E. L. Bell (to Eastman Kodak Company).; *Photographic silver halide emulsion containing contrast improving dopants*, U.S. Pat. 5597686, 1996, G. L. MacIntyre and E. L. Bell (to Eastman Kodak Company).
35. *Preparation of silver halide emulsions containing iridium*, U.S. Pat. 4902611, 1990, I. H. Leubner and W. W. White.

36. *Internally doped silver halide emulsions*, U.S. Pat. 4835093, 1989, G. A. Janusonis, et al., (to Eastman Kodak Company); *Photographic emulsions containing internally modified silver halide grains*, U.S. Pat. 4933272, 1990, W. G. McDugle, et al., (to Eastman Kodak Company).; *Photographic emulsions containing internally modified silver halide grains*, U.S. Pat. 4937180, 1990, A. P. Marchetti, W. G. McDugle, and R. S. Eachus (to Eastman Kodak Company).; *Photographic emulsions containing internally modified silver halide grains*, U.S. Pat. 4945035, 1990, J. E. Keevert, Jr., W. G. McDugle, and R. S. Eachus (to Eastman Kodak Company).; *Photographic emulsions containing internally modified silver halide grains*, U.S. Pat. 4981781, 1991, W. G. McDugle, J. E. Keevert, and A. P. Marchetti (to Eastman Kodak Company).; *Photographic emulsions containing internally modified silver halide grains*, U.S. Pat. 5037732, 1991, W. G. McDugle, A. P. Marchetti, J. E. Keevert, M. S. Henry, and M. T. Olm (to Eastman Kodak Company).
37. *Silver halide emulsions with doped epitaxy*, U.S. Pat. 5462849, 1995, T. Y. Kuromoto, et al., (to Eastman Kodak Company).
38. *High chloride emulsions doped with combination of metal complexes*, U.S. Pat. 6107018, 2000, J. Z. Mydlarz, E. L. Bell, and M. S. Graham (to Eastman Kodak Company).
39. Keevert et al.; *Photographic emulsions containing internally modified silver halide grains*, U.S. Pat. 4945035, 1990, J. E. Keevert, Jr., W. G. McDugle, and R. S. Eachus (to Eastman Kodak Company).
40. *Epitaxially sensitized ultrathin tabular grain emulsions*, U.S. Pat. 5494789, 1996, R. L. Daubendiek, et al., (to Eastman Kodak Company).; *Ultrathin tabular grain emulsions with novel dopant management*, U.S. Pat. 5503970, 1996, M. T. Olm, et al., (to Eastman Kodak Company).; *Ultrathin tabular grain emulsions containing speed-granularity enhancements*, U.S. Pat. 5503971, 1996, R. L. Daubendiek, et al., (to Eastman Kodak Company).
41. Research Disclosure, vol. 367, November 1994, Item 36736.
42. *Research Disclosure, vol. 120, April, 1974, Item 12008; Research Disclosure, vol. 134, June, 1975, Item 13452*; *Photographic emulsion and material and process used in the preparation thereof*, U.S. Pat. 1623499, 1927, S. E. Sheppard and R. F. Punnett (to Eastman Kodak Company, Rochester, NY); *Process of manufacturing photographic silver halide emulsions and products obtained thereby*, U.S. Pat. 1673522, 1928, O. Matties, P. Wulff, W. Dieterle, and B. Wendt (to I. G. Farbenindustrie Aktiengesellchaft).; *Photographic materials*, U.S. Pat. 2399083, 1946, C. Waller, R. B. Collins, and E. C. Dodd (to Ilford, Ltd.).; *Photographic emulsions sensitized with salts of metals of group VIII of the periodic arrangement of the elements*, U.S. Pat. 2448060, 1948, W. F. Smith and A. P. H. Trivelli (to Eastman Kodak Company).; *Photographic silver halide emulsions sensitized with water-insoluble gold compounds*, U.S. Pat. 2642361, 1953, R. E. Damschroder and H. C. Yutzy (to Eastman Kodak Company).; *Synergistic sensitization of photographic systems with labile selenium and a noble metal*, U.S. Pat. 3297446, 1967, J. S. Dunn (to Eastman Kodak Company, Rochester, NY); *Stabilization of synergistically sensitized photographic systems*, U.S. Pat. 3297447, 1967, P. A. McVeigh (to Eastman Kodak Company, Rochester, NY); *Photographic emulsions containing certain silver-gelatin concentrations during chemical ripening*, U.S. Pat. 3565633, 1971, G. H. Klinger and M. V. Cwikla (to GAF Corporation, New York, NY); *Photographic element containing monodispersed, unfogged, dye sensitized silver halide grains metal ions sensitized internally and the use thereof in reversal process*, U.S. Pat. 3761267, 1973, P. W. Gilman, Jr., R. G. Raleign, and J. A. Merrigan (to Eastman Kodak Company, Rochester, NY); *Silver halide grains and photographic emulsions*, U.S. Pat. 3772031, 1973, C. R. Berry, S. J. Marino (to Eastman Kodak Company, Rochester, NY); *Silver halide photographic emulsion sensitized with a heterocyclic compound containing 4-sulfur atoms*, U.S. Pat. 3857711, 1974, R. Ohi, M. Sugiyama, Y. Nakajima, M. Horie, and S. Kobayashi (to Fuji Photo Film Co., Ltd., Ashigara-Kamigun, Kanagawa, JA); *Silver halide emulsions and elements including sensitizers of adamantane structure*, U.S. Pat. 3901714, 1975, E. N. Oftedahl (to Eastman Kodak Company, Rochester, NY); *Phosphine sensitized photographic silver halide emulsions and elements*, U.S. Pat. 3904415, 1975, E. N. Oftedahl (to Eastman Kodak Company, Rochester, NY); *Preparation of photographic silver halide materials*, UK Pat. 1315755, 1973, C. E. McBride.; *Sensitive silver halide photographic materials*, UK Pat. 1396696, 1975, M. J. Simons.
43. Research Disclosure 37154.
44. *Chemical sensitization of photographic emulsions*, U.S. Pat. 2518698, 1950, W. G. Lowe and J. E. Jones (to Eastman Kodak Company).; *Chemical sensitization of photographic emulsions*, U.S. Pat. 2739060, 1956, W. G. Lowe, J. E. Jones, and H. E. Roberts (to Eastman Kodak Company).; *Chemical sensitization of photographic emulsions*, U.S. Pat. 2743182, 1956, W. G. Lowe, J. E. Jones, and H. E. Roberts (to Eastman Kodak Company).; *Chemical sensitization of photographic emulsions*, U.S. Pat. 2743183, 1956, W. G. Lowe, J. E. Jones, and H. E. Roberts (to Eastman Kodak Company).; *Chemical sensitization of photographic emulsions*, U.S. Pat. 2983609, 1961, C. F. H. Allen and W. G. Lowe (to Eastman Kodak Company).; *Silver halide photographic emulsions containing linear polyamine sensitizing agents*, U.S. Pat. 3026203, 1962, V. C. Chambers, Jr. and A. E. Oberth (to E. I. Du Pont de Nemours and Co.).; *Amine borane as fogging agent in direct positive*, U.S. Pat. 3361564, 1968, J. H. Bigelow and C. R. Burt (to E. I. Du Pont de Nemours and Co.); *Polyhedral borane fogged direct-positive silver halide emulsion containing and organic sulfoxide*, U.S. Pat. 3782959, 1974, J. H. Bigelow (to E. I. Du Pont de Nemours and Co.); *Method of manufacturing silver halide emulsion*, U.S. Pat. 5254456, 1993, S. Yamashita, S. Takada, and S. Shibayama (to Fuji Photo Film Co., Ltd.).; *Photographic element exhibiting improved speed and stability*, U.S. Pat. 5399479, 1995, R. Lok (to Eastman Kodak Company).; *Photographic sensitivity increasing alkynylamine compounds and photographic elements*, U.S. Pat. 5413905, 1995, R. Lok, C. R. Preddy, and J. W. Harder (to Eastman Kodak Company).; *Class of compounds which increases and stabilizes photographic speed*, U.S. Pat. 5500333, 1996, J. N. Eikenberry, R. Lok, and C. Chen (to Eastman Kodak Company).; *Silver halide photographic material and process for its preparation*, Eur. Pat. 0 407 576, 1990, U. Shigeharu.; *Silver halide photographic light-sensitive material*, Eur. Pat. 0 552 650, 1993, O. Hirofumi; Oftedahl et al., Research Disclosure, vol. 136, August 1975, Item 13654.
45. *Photographic silver halide emulsions sensitized with water-insoluble gold compounds*, U.S. Pat. 2642361, 1953, R. E. Damschroder and H. C. Yutzy (to Eastman Kodak Company).
46. *Chemical sensitization of photographic emulsions*, U.S. Pat. 2521926, 1950, W. G. Lowe and J. E. Jones (to Eastman Kodak Company).; *Polythiaalklyenediols as sensitizers for photographic silver halide emulsions*, U.S. Pat. 3021215, 1962, J. L. R. Williams and B. C. Cossar (to Eastman Kodak Company).; *Silver halide emulsions containing hexathiocane thiones as sensitizers*, U.S. Pat. 4054457, 1977, J. H. Bigelow (to E. I .Du Pont de Nemours and Co.).

47. *Method for stabilizing X-ray emulsions to red safelights*, U.S. Pat. 3411914, 1968, C. Dostes (to Eastman Kodak Company, Rochester, NY); *Stabilized photographic silver halide composition*, U.S. Pat. 3554757, 1971, Y. Kuwabara and S. Sugita (to Konishiroku Photo Industry Co., Ltd.); *6-Amidinothio -5,7-Dihydroxy-s-triazolo[2,3-a]pyrimidino compounds as stabilizers for silver halide emulsion*, U.S. Pat. 3565631, 1971, M. Oguchi, Y. Kuwabara, and K. Mogaki (to Konishiroku Photo Industry Co., Ltd.); *Silver halide emulsions and elements including sensitizers of adamantane structure*, U.S. Pat. 3901714, 1975, E. N. Oftedahl (to Eastman Kodak Company, Rochester, NY); *Silver halide photographic light-sensitive material*, U.S. Pat. 4897342, 1990, M. Kajiwara, M. Miyoshi, and K. Onodera (to Konishiroku Photo Industry Co., Ltd.).; *Silver halide photographic emulsions*, U.S. Pat. 4968595, 1990, S. Yamada, et al., (to Fuji Photo Film Co., Ltd.).; *Process for preparing silver halide emulsion and silver halide X-ray photographic material containing said emulsion*, U.S. Pat. 5114838, 1992, S. Yamada (to Fuji Photo Film Co., Ltd.).; *Silver halide photographic emulsion*, U.S. Pat. 5118600, 1992, S. Yamada and M. Kuwabara (to Fuji Photo Film Co., Ltd.).; *Process of preparing for photographic use high chloride tabular grain emulsion*, U.S. Pat. 5176991, 1993, C. G. Jones and T. L. Osborne-Perry (to Eastman Kodak Company).; *Silver halide photographic material and method for processing the same*, U.S. Pat. 5190855, 1993, I. Toya, R. Inoue, and T. Ito (to Fuji Photo Film Co., Ltd.).; *Silver halide photographic material*, Eur. Pat. 0 554 856, 1993, T. Hioki and R. Nishimura.

48. *Light sensitive halide material with variable contrast*, U.S. Pat. 3628960, 1971, H. A. Philippaerts, H. J. Corluy, H. Gernet, and G. E. W. Schulz (to Gevaert-AGFA N.V., Mortsel, Belgium); *Sensitized high aspect ratio silver halide emulsions and photographic elements*, U.S. Pat. 4439520, 1984, J. T. Kofron, et al., (to Eastman Kodak Company); *Photographic element exhibiting reduced sensitizing dye stain*, U.S. Pat. 4520098, 1985, R. E. Dickerson (to Eastman Kodak Company); *Radiographic elements exhibiting reduced crossover*, U.S. Pat. 4639411, 1987, R. L. Daubendiek, R. E. Dickerson, and J. E. Kelly (to Eastman Kodak Company); *Method for manufacturing a silver halide emulsion*, U.S. Pat. 4693965, 1987, M. Ihama and T. Tani (to Fuji Photo Film Co., Ltd.); *Silver halide photographic material*, U.S. Pat. 4791053, 1988, T. Ogawa (to Fuji Photo Film Co., Ltd.); *High sensitivity light-sensitive silver halide photographic material with little stain*, U.S. Pat. 4925783, 1990, I. Metoki, C. Honda, and Y. Wada (to Konica Corporation).; *Color photographic recording material and a process for the preparation of a photographic silver halide emulsion*, U.S. Pat. 5077183, 1991, H. Reuss, B. Mucke, and H. Kampfer (to Agfa Gevaert Aktiengesellschaft).; *Process for producing silver halide photographic material*, U.S. Pat. 5130212, 1992, I. Morimoto, K. Hattori, K. Ueda, and H. Ushiroyama (to Konica Corporation).; *Method for preparing photographic emulsion*, U.S. Pat. 5141846, 1992, K. E. Fickie and M. R. Lee (to Polaroid Corporation).; *Silver halide light-sensitive photographic material*, U.S. Pat. 5192652, 1993, M. Kajiwara and M. Miyoshi (to Konica Corporation).; *Method of manufacturing silver halide emulsion and a color photographic material having the emulsion manufactured by the method*, U.S. Pat. 5230995, 1993, M. Asami (to Fuji Photo Film Co., Ltd.).; *Silver halide photographic light-sensitive material*, U.S. Pat. 5238806, 1993, Y. Hashi (to Fuji Photo Film Co., Ltd.).; *High sensitive silver halide photographic light-sensitive material*, Eur. Pat. 0 354 798, 1990, K. Goan and T. Marui.; *Silver halide color photographic material.*, Eur. Pat. 0 509 519, 1992, Y. Kashi.; *Method for manufacturing a silver halide emulsion*, Eur. Pat. 0 533 033, 1993, M. Kubotera and M. Kajiwara.; *Silver halide photographic light-sensitive material.*, Eur. Pat. 0 556 413, 1993, T. Marui.; *A silver halide photographic emulsion and a photographic light-sensitive material*, Eur. Pat. 0 562 476, 1993, Y. Suga, M. Kikuchi, H. Okamura, M. Yagihara, and H. Kawamoto.

49. *Sensitized high aspect ratio silver halide emulsions and photographic elements*, U.S. Pat. 4439520, 1984, J. T. Kofron, R. E. Booms, C. G. Jones, J. A. Haefner, H. S. Wilgus, and F. J. Evans (to Eastman Kodak Company); *Novel silver halide crystals with two surface types*, UK Pat. 2038792, 1980, E. F. Haugh. E. L. Kitts, and D. J. Mickewich (to Du Pont); *Process for producing a silver halide photographic material*, Eur. Pat. 0 302 528, 1989, H. Mifune and S. Takada.

50. Research Disclosure, item 36544, Section V. Spectral sensitization and desensitization.

51. P. B. Gilman, Jr., *Photographic Science and Engineering*, **18**(4), 418–430 (1974).

52. E. J. Wall, *Photographic Emulsions; their preparation and coating on glass, celluloid and paper, experimentally and on the large scale*, American Photographic Publishing Co., Boston, 1929, p. 65; *Prevention of dye wandering in photographic emulsions*, U.S. Pat. 2735766, 1956, G. D. Hill (to Eastman Kodak Company).; *Light sensitive halide material with variable contrast*, U.S. Pat. 3628960, 1971, H. A. Philippaerts, H. J. Corluy, H. Gernet, and G. E. W. Schulz (to Gevaert-AGFA N.V.); *Pre-precipitation spectral sensitizing dye addition process*, U.S. Pat. 4183756, 1980, D. J. Locker (to Eastman Kodak Company); *Silver halide precipitation and methine dye spectral sensitization process and products thereof*, U.S. Pat. 4225666, 1980, D. J. Locker and R. L. Daubendiek (to Eastman Kodak Company); *Method for producing a silver halide photographic emulsion.*, Eur. Pat. App. 301508, 1989, T. Tani and T. Ikeda; Research Disclosure, vol. 181, May, 1979, Item 18155.

53. *Light sensitive halide material with variable contrast*, U.S. Pat. 3628960, 1971, H. A. Philippaerts, H. J. Corluy, H. Gernet, and G. E. W. Schulz (to Gevaert-AGFA N.V., Mortsel, Belgium); *Controlled site epitaxial sensitization*, U.S. Pat. 4435501, 1984, J. E. Maskasky (to Eastman Kodak Company); *Sensitized high aspect ratio silver halide emulsions and photographic elements*, U.S. Pat. 4439520, 1984, J. T. Kofron, et al., (to Eastman Kodak Company); *Photographic element exhibiting reduced sensitizing dye stain*, U.S. Pat. 4520098, 1985, R. E. Dickerson (to Eastman Kodak Company).

54. *Silver halide photographic material*, Eur. Pat. App. 287100, 1988, M. Asami, N. Ohshima and S. Ohtani (to Fuji Photo Film Co., Ltd.).

55. *Production of photographic material*, U.S. Pat. 2912343, 1959, R. B. Collins and S. F. W. Welford (to Ilford, Ltd.).

56. *Photographic element exhibiting reduced sensitizing dye stain*, U.S. Pat. 4520098, 1985, R. E. Dickerson (to Eastman Kodak Company).

57. *Process for producing photographic emulsions*, U.S. Pat. 3822135, 1974, T. Sakai, A. Sato, N. Yamamota, and S. Kondo (to Fuji Photo Film Co., Ltd., Ashigara-Kamigun, Kanagawa, JA).

58. Japanese published Patent Application (Kokai) 24185/71; *Method of spectrally sensitizing photographic silver halide emulsions*, U.S. Pat. 3469987, 1969, J. M. Owens, J. L. Graham, and E. A. Pilsworth (to Eastman Kodak Company, Rochester, NY).

59. "Farbkuppler–Eine Literature Ubersicht," published in Agfa Mitteilungen, Band III, pp. 156–175 (1961).; *Acylaminophenol photographic couplers*, U.S. Pat. 2367531,

1945, I. F. Salminen, P. W. Vittum, and A. Weissberger (to Eastman Kodak Company).; *Acylamino phenols*, U.S. Pat. 2423730, 1947, I. F. Salminen and A. Weissberger (to Eastman Kodak Company).; *1-Naphthol-2-carboxylic acid amide couplers for color photography*, U.S. Pat. 2474293, 1949, A. Weissberger, I. F. Salminen, and P. W. Vittum (to Eastman Kodak Company).; *Diacyclaminophenol couplers*, U.S. Pat. 2772162, 1956, I. F. Salminen and C. R. Barr (to Eastman Kodak Company).; *Photographic color couplers containing fluoroalkylcarbonamido groups*, U.S. Pat. 2895826, 1959, I. F. Salminen, C. R. Barr, and A. Loria (to Eastman Kodak Company).; *Cyan color former for color photography*, U.S. Pat. 3002836, 1961, P. W. Vittum and A. Weissberger (to Eastman Kodak Company).; *Magenta-colored cyan-forming couplers*, U.S. Pat. 3034892, 1962, R. J. Gledhill and J. Williams (to Eastman Kodak Company).; *Silver halide photographic material*, U.S. Pat. 4883746, 1989, Y. Shimada, H. Fukuzawa, and S. Ichijima (to Fuji Photo Film Co., Ltd.).

60. Farbkuppler–Eine Literature Ubersicht, published in Agfa Mitteilungen, Band III, pp. 126–156 (1961).; *Pyrazolone coupler for color photography*, U.S. Pat. 2311082, 1943, H. D. Porter and A. Weissberger (to Eastman Kodak Company).; *Pyrazolone coupler for color photography*, U.S. Pat. 2343703, 1944, H. D. Porter and A. Weissberger (to Eastman Kodak Company).; *Acylated amino pyrazolone couplers*, U.S. Pat. 2369489, 1945, H. D. Porter and A. Weissberger (to Eastman Kodak Company).; *Halogen-substituted pyrazolone couplers for color photography*, U.S. Pat. 2600788, 1952, A. Loria, A. Weissberger, and P. W. Vittum (to Eastman Kodak Company, Rochester, NY); *Photographic color couplers containing mono n- alkyl groups*, U.S. Pat. 2908573, 1959, W. M. Bush, I. F. Salminen, and J. R. Thirtle (to Eastman Kodak Company).; *Photographic emulsion containing pyrazolone magenta-forming couplers*, U.S. Pat. 3062653, 1962, A. Weissberger, A. Loria, and I. F. Salminen (to Eastman Kodak Company).; *Magenta-forming couplers*, U.S. Pat. 3152896, 1964, R. J. Tuite (to Eastman Kodak Company, Rochester, NY); *Silver halide emulsions containing a stabilizer pyrazolone coupler*, U.S. Pat. 3519429, 1970, G. J. Lestina (to Eastman Kodak Company, Rochester, NY).

61. Farbkuppler–Eine Literature Ubersicht, published in Agfa Mitteilungen, Band III, pp. 112–126 (1961); *Nondiffusing sulfonamide coupler for color photography*, U.S. Pat. 2298443, 1942, A. Weissberger (to Eastman Kodak Company).; *Colored couplers*, U.S. Pat. 2407210, 1946, A. Weissberger, C. J. Kibler, and P. W. Vittum (to Eastman Kodak Company).; *Benzoylacet-o-alkoxyanilide couplers for color photography*, U.S. Pat. 2875057, 1959, F. C. McCrossen, P. W. Vittum, and A. Weissberger (to Eastman Kodak Company).; *Yellow forming couplers*, U.S. Pat. 3265506, 1966, A. Weissberger and C. J. Kibler (to Eastman Kodak Company, Rochester, NY); *Silver halide emulsion containing two-equivalent yellow dye-forming coupler*, U.S. Pat. 3447928, 1969, A. Loria (to Eastman Kodak Company, Rochester, NY)

62. *Colored couplers*, U.S. Pat. 2455169, 1948, D. B. Glass, P. W. Vittum, and A. Weissberger (to Eastman Kodak Company, Rochester, NY); *Photographic color reproduction process and element*, U.S. Pat. 3227551, 1966, C. R. Barr, J. Williams, and K. Whitmore (to Eastman Kodak Company).; *4-Acyloxy-5-pyrazolones*, U.S. Pat. 3432521, 1969, A. Loria (to Eastman Kodak Company, Rochester, NY); *Photographic silver halide elements containing two equivalent cyan couplers*, U.S. Pat. 3476563, 1969, A. Loria (to Eastman Kodak Company, Rochester, NY); *Two-equivalent couplers for photography*, U.S. Pat. 3617291, 1971, G. W. Sawdey (to Eastman Kodak Company, Rochester, NY); *Silver halide emulsion containing acylamidophenol photographic couplers*, U.S. Pat. 3880661, 1975, P. T. Lau, I. F. Salminen, and L. E. Beavers (to Eastman Kodak Company, Rochester, NY); *Photographic silver halide emulsion containing 2-equivalent cyan coupler*, U.S. Pat. 4052212, 1977, H. Deguchi, et al., (to Konishiroku Photo Industry Co., Ltd.); *Dye image forming process*, U.S. Pat. 4134766, 1979, S. Kikuchi, et al., (to Konishiroku Photo Industry Co., Ltd.); *Two equivalent type colour couplers containing carbamoyloxy groups at their active positions and their use in photography*, UK Pat. 1466728, 1977.; *Photographic cyan colour couplers and silver halide materials*, UK Pat. 1531927, 1978.; *Photographic silver halide materials and developers containing couplers*, UK Pat. 1533039, 1978.; *Cyan couplers containing sulfonamide groups for use in photography*, UK Pat. 2006755A, 1978, M. Yagihara and Y. Yokada.; Ketomethylene yellow dye-forming couplers, UK Pat. 2017704A, 1979, P. T. S. Lau.

63. Research Disclosure, Item 36544, Section VII. Antifoggants and stabilizers.

64. *Photographic emulsion containing alkyl quaternary salts of thiazoles and the like as antifoggants*, U.S. Pat. 2131038, 1938, L. G. S. Brooker and C. J. Staud (to Eastman Kodak Company).; *Polymethylene-bis-benzothiazolium salts*, U.S. Pat. 2694716, 1954, C. F. H. Allen and C. V. Wilson (to Eastman Kodak Company, Rochester, NY); *Benzothiazolium compounds for controlling overdevelopment*, U.S. Pat. 3342596, 1967, J. L. Graham (to Eastman Kodak Company, Rochester, NY); *Silver halide emulsion containing an alkenyl benzothiazolium salt as stabilizer*, U.S. Pat. 3954478, 1976, N. Arai and R. Ohi (to Fuji Photo Film Co., Ltd.); *Quaternized tellurium salt fog inhibiting agents for silver halide photography*, U.S. Pat. 4661438, 1987, R. Przyklek-Elling, W. H. H. Gunther, and R. Lok (to Eastman Kodak Company).

65. *Photographic emulsion containing fog inhibitors*, U.S. Pat. 2319090, 1943, S. E. Sheppard and W. Vanselow (to Eastman Kodak Company).; *Improvers for photographic emulsions*, U.S. Pat. 2403927, 1946, J. D. Kendall, D. J. Fry, and J. D. Brooks (to Ilford, Ltd.).; *Antifoggant agents for photography*, U.S. Pat. 3266897, 1966, K. C. Kennard and D. F. Reard on (to Eastman Kodak Company, Rochester, NY); *Photographic emulsions containing mercapto development anti-foggants*, U.S. Pat. 3397987, 1968, G. W. Luckey and B. D. Illingsworth (to Eastman Kodak Company, Rochester, NY); *Photographic elements and processes lithographic silver halide element containing a 1. (amidophenyl)-5-mercaptotetrazole sensitizing agent and development process of using same*, U.S. Pat. 3708303, 1973, E. D. Salesin (to Eastman Kodak Company, Rochester, NY).

66. *Manufacture of photographic gelatin emulsions*, U.S. Pat. 2152460, 1939, E. J. Birr, H. Fricke, and G. Wilmanns (to Agfa Ansco Corp.); *Stabilizers for photographic emulsions*, U.S. Pat. 2444605, 1948, N. Heimbach and W. Kelly, Jr. (to General Aniline & Film Corporation); *Tetraazaindene compounds and photographic application*, U.S. Pat. 2933388, 1960, E. B. Knott (to Eastman Kodak Company); *Photographic silver halide emulsions stabilized with tetraazaindene compounds*, U.S. Pat. 3202512, 1965, L. A. Williams (to Eastman Kodak Company); *Process for the production of photographic emulsions*, U.K. Pat. 1338567, 1973, M. S. W. Nepker and K. D. Strube (to Wolfen Fotochemische Werke); *Nouvelles émulsions négatives aux halogénures d'argent protégées contre la désensibilisation par pression*, Fr. Pat. 2296204, 1976, C. G. Dostes, D.C. Dubromel, and J.-F. Le Guen (to Kofak Pathe); Research Disclosure, vol. 134, June, 1975, Item 13452; Research Disclosure vol. 148, August 1976, Item 14851.

67. *Stabilized photographic silver halide emulsions containing 5-thioctic acid*, U.S. Pat. 2948614, 1960, C. F. H. Allen, B. D. Illingsworth, and J. J. Sagura (to Eastman Kodak Company).; *Photographic emulsion stabilized with bis(p-acylamidophenyl)disulfides*, U.S. Pat. 3397986, 1968, A. G. Millikan and A. H. Herz (to Eastman Kodak Company).; *Amido substituted divalent chalcogenide fog inhibiting agents for silver halide photography*, U.S. Pat. 4607000, 1986, W. H. H. Gunther and R. Lok (to Eastman Kodak Company); *Divalent chalcogenide fog inhibiting agents for silver halide photography*, U.S. Pat. 4607001, 1986, R. Lok and W. H. H. Gunther (to Eastman Kodak Company); *Cyclic dichalcogenide fog inhibiting agents for silver halide photography*, U.S. Pat. 4861703, 1989, R. Lok, W. H. H. Gunther, and J. P. Freeman (to Eastman Kodak Company).; *Silver halide emulsion containing an organic selenium compound antifogging agent*, UK Pat. 1282303, 1973, R. Pollet and J. Willems.; *Photographic emulsions*, UK Pat. 1336570, 1973, J. C. Brown, D. J. Fry, and P. J. Keogh.

68. *Photographic emulsion stabilized with bis(p-acylamidophenyl) disulfides*, U.S. Pat. 3397986, 1968, A. G. Millikan and A. H. Herz (to Eastman Kodak Company).

69. *Photographic emulsions containing mercury salts*, U.S. Pat. 2728664, 1955, B. H. Carroll and T. F. Murray (to Eastman Kodak Company).

70. *Method of preparing photographic silver halide emulsions*, U.S. Pat. 3957490, 1976, M. J. Libeer and F. H. Claes (to Agfa-Gevaert N.V.).

71. Japanese Patent Publication No. 9503/69; *Photographic developer containing a pyridinium salt and process of development*, U.S. Pat. 2648604, 1953, L. G. Welliver and C. E. Johnson (to General Aniline & Film Corporation).

72. Japanese Patent Publication No. 9504/69; *Silver halide developers containing polyethylene glycols*, U.S. Pat. 2531832, 1950, W. A. Stanton (to E. I. Du Pont de Nemours and Co.).; *Silver halide developer compositions containing polyoxyalkylene ethers of hexitol ring dehydration products*, U.S. Pat. 2533990, 1950, R. K. Blake (to E. I. Du Pont de Nemours and Co.).; *Photographic element with colloid layer containing color former and nonionic wetting agent*, U.S. Pat. 2577127, 1951, A. B. Jennings and O. W. Murray (to E. I. Du Pont de Nemours and Co.); *Color developers containing polyethylene glycols*, U.S. Pat. 2950970, 1960, J. A. Schwan and C. M. Spath (to Eastman Kodak Company).; *Accelerators for reversal color development*, U.S. Pat. 3201242, 1965, J. A. Schwan and J. R. Dann (to Eastman Kodak Company).

73. *Light-sensitive photographic silver halide recording material*, U.S. Pat. 4418142, 1983, H. Langen, E. Wolff, and E. Ranz (to Agfa Gevaert Aktiengesellschaft); *Silver halide color photographic material*, U.S. Pat. 4618573, 1986, H. Okamura, H. Kobayashi, T. Nishikawa, and K. Tamoto (to Fuji Photo Film Co., Ltd.); *Silver halide color photographic material*, U.S. Pat. 4863841, 1989, H. Okamura, H. Kobayashi, T. Nishikawa, and K. Tamoto (to Fuji Photo Film Co., Ltd.).; *Method and composition for hardening gelatin*, U.S. Pat. 4877724, 1989, C. Y. Chen, E. E. Riecke, K. G. Harbison, and D. D. Chapman (to Eastman Kodak Company).; *Photographic recording material*, U.S. Pat. 5009990, 1991, R. Scheerer and F. Moll (to Agfa-Gevaert Aktiengesellschaft).; *Method and composition for hardening gelatin*, U.S. Pat. 5236822, 1993, E. E. Riecke, D. D. Chapman, C. Y. Chen, and K. G. Harbison (to Eastman Kodak Company).

74. R. W. G. Hunt, *The Reproduction of Colour*, 4e, 1987, pp. 306–307.

75. H. C. Bozenhard, B. Narayan, *IS&Ts 11th Int. Symp. Photofinishing Technol.*, 2000, pp. 86–89.

76. M. Miyoshi and H. Nakatsugawa, *IS&Ts 11th Int. Symp. Photofinishing Technol.*, 2000, pp. 59–60.

77. Research Disclosure, vol. 308, December 1989, Item 308119, Section VII. Color materials, paragraph H, and Research Disclosure, Item 36544, Section X. Dye image formers and modifiers, Subsection A. Silver dye bleach.

RF MAGNETIC FIELD MAPPING

JOSEPH P. HORNAK
Rochester Institute of Technology
Rochester, NY

INTRODUCTION

Radio-frequency (RF) magnetic field mapping is the imaging of oscillating magnetic fields in space. The title of this article implies RF mapping, however, the technique is applicable to both higher frequency microwaves and lower audio frequency electromagnetic waves. At a fixed point in space, the magnetic field, $B(t)$, associated with these waves varies sinusoidally with respect to time t at frequency ν according to the general wave equation

$$B(t) = B_1 \sin(2\pi\nu t), \tag{1}$$

where B_1 is the amplitude factor. It is this amplitude factor, B_1, that is mapped or imaged in one, two, or three spatial dimensions.

Five mapping techniques are presented in this article. They were developed for mapping the oscillating magnetic fields (standing waves) associated with inductive–capacitive–resistive (LCR) circuits, such as cavity resonators and resonant coils used in magnetic resonance spectroscopy and imaging. LCR circuits store energy alternately in the capacitive element as electrical fields and in the inductive element as magnetic fields. The oscillating magnetic fields form a standing wave within the volume of the inductor. The electrical fields are either zero or separated in space from the magnetic fields, and hence generally are not detected in these circuits. This point is worth noting because some of the techniques presented in this article cannot distinguish oscillating magnetic and electrical fields. Figure 1 is a schematic representation of a multiturn solenoid LCR circuit and the direction of B_1 at the center of the solenoid.

Because of the magnetic resonance application of these techniques, much of the magnetic resonance symbolism has been adopted, and some basic magnetic resonance theory will be presented. The reader is directed to the references for more detailed information on nuclear magnetic resonance (NMR) (1,2) and electron paramagnetic resonance (EPR) (3), and to the articles in this encyclopedia on TOMOGRAPHIC IMAGE FORMATION TECHNIQUES and ELECTRON PARAMAGNETIC RESONANCE (EPR) IMAGING.

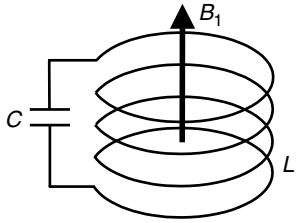


Figure 1. A schematic depiction of an LCR circuit and the direction of the B_1 field within the inductor.

The mapping techniques are divided into two general categories: probe and imaging techniques. Probe techniques produce field values at discrete points in space that can be used to generate a B_1 map. Imaging techniques produce an image directly. The method depends on the proportionality of the intensity of a pixel in this image to the magnetic field at the corresponding point in space. Although B_1 in tesla can be calculated from any of the techniques presented in this section, it is often reported in relative units.

PROBE TECHNIQUES

Probe techniques are based on a measurement that is made from a small probe placed in the oscillating magnetic field. The probe is moved to specific points in space, typically on a predetermined grid, and measurements are made. The measurements represent the field at specific points in space. These measurements are combined to produce a contour map or image of the B_1 field. The spatial resolution of probe techniques is limited by the size of the probe, the spatial density of measurements, and the accuracy of the positioning device.

Method of Perturbing Spheres

The method of perturbing spheres (4,5) is useful for measuring the absolute B_1 inside a resonant circuit such as a microwave cavity perfectly matched to a frequency source. This technique requires a mechanism for precisely measuring the resonant frequency of the cavity, as well as the full frequency width ($\Delta\nu$) at half-power of the absorption signal from the cavity. A gold or silver conductive sphere of radius r is supported by a thin, low-dielectric, nonconducting rod (typically glass or Teflon). The sphere can be made of any conductive material. However, gold or silver are typically used to minimize resistive losses due to the sphere and to improve sensitivity. The sphere is placed inside the resonator and the B_1 field in tesla at the location of the sphere is given by Eq. (2), where W is the power lost in the cavity and μ the permeability constant of the material in the cavity.

$$B_1 = \frac{1}{2}\left\{\left(\frac{\nu^2 - \nu_o^2}{\nu_o^2}\right)\frac{\mu W}{\pi^2 \Delta\nu r^3}\right\}^{\frac{1}{2}}. \tag{2}$$

The sphere perturbs the resonant frequency ν_o of the cavity to frequency ν. Therefore, mapping ν allows calculating B_1. The B_1 sensitivity of this technique is limited by the frequency resolution of the device used to monitor

the resonant frequency of the cavity. The method of perturbing spheres has been applied to map the B_1 fields of X-band microwave cavity resonators (4) and loop-gap resonators (6) that have fields of view less than 2.5 cm. The technique is applicable to larger size, lower frequency resonators as well. The method of perturbing spheres has recently been applied to map the RF fields from large-volume, low-frequency ESR and Overhauser MRI resonators (7,8).

Note that the method of perturbing spheres is sensitive to both electrical and magnetic field components. The technique gives an accurate mapping of the magnetic fields only if the electrical and magnetic field components are well separated in space. Otherwise, the use of small conductors that have variable shapes is required to distinguish the contributions of the two components.

Pickup Coil Technique

Pickup coil techniques (9–16) measure a B_1 field by detecting the current induced in a coil of wire placed in an oscillating magnetic field (Fig. 2). A small N turn coil of wire of radius r is placed in an oscillating magnetic field of frequency ν. The average B_1 within the pickup coil and perpendicular to the plane of the coil is related to the maximum voltage $V_{\max}$ produced in the coil by Eq. (3):

$$B_1 = \frac{V_{\max}}{2\pi^2 r^2 \nu N}. \tag{3}$$

The orientation of the loop may be varied, and therefore, the orthogonal components of B_1 may be measured. The B_1 sensitivity is limited by the signal-to-noise ratio (SNR) in the voltage monitoring system. The smaller the coil, the better the spatial resolution, but the poorer the SNR and hence B_1 sensitivity. This technique has been applied to 2×10^3 cm^3 imaging coils used in magnetic resonance imaging (11). The technique may also be used to map the fields in nonresonant devices and in free space, such as those from radar antennae.

NMR Technique

The simplest magnetic resonance experiment, or pulse sequence as it is more commonly called, consists of a one-pulse sequence (17). In this pulse sequence, the net magnetization from the sample is rotated by an angle θ away from its equilibrium position along the $+Z$ axis. The rotational angle in degrees is given by

$$\theta = 2\pi\gamma B_1 \tau, \tag{4}$$

where γ is the gyromagnetic ratio for the signal bearing nucleus in (sT)$^{-1}$ and τ is the width in seconds of the RF pulse of amplitude B_1 tesla.

After rotation, the XY component of the magnetization precesses about the direction of the applied static magnetic field, which coincides with the $+Z$ axis. The signal induced in the resonator from the XY magnetization as it rotates about $+Z$ is called free induction decay (FID). The Fourier transform of the FID is a frequency domain spectrum of the sample. Recording a series of frequency domain spectra as a function of τ yields maxima in the signal when

$$\theta = n90^\circ, \tag{5}$$

and n is an odd integer. Knowing the τ value when $\theta = 90^\circ$ allows calculating B_1. This is a rather tedious but effective process if an exact B_1 value is needed. The B_1 sensitivity is limited by the ability to determine the maximum in the signal accurately as a function of τ.

A technique similar to this one, but using a spin-echo pulse sequence has been used to map the B_1 fields in *NMR* coils (18). Another variation on this technique is called the rotating-frame technique (19). A series of NMR spectra are recorded at different τ values. From Eq. (4), the difference in τ values, $\Delta\tau$, between successive maxima is $(\gamma B_1)^{-1}$. In all NMR techniques, the timing of the pulses in the sequence must be chosen to minimize effects from spin relaxation.

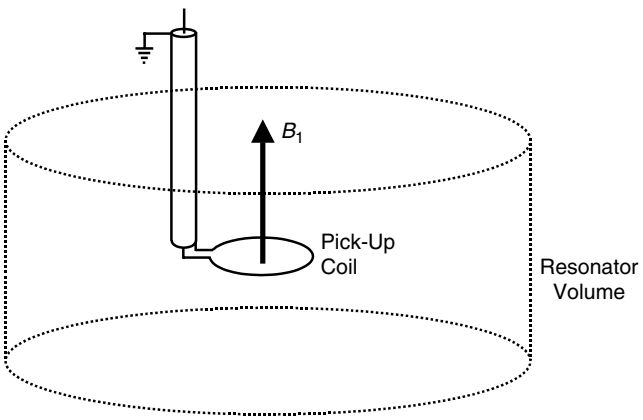


Figure 2. A schematic representation of a pickup coil oriented for maximum signal pickup in the B_1 field from a resonator.

IMAGING TECHNIQUES

Imaging techniques require using a magnetic resonance imager and are limited to imaging the B_1 field from an imaging coil or resonator connected to the imager. (The reader should refer to the article in this encyclopedia on MAGNETIC RESONANCE IMAGING (MRI) for more detailed information on the magnetic resonance imaging process and signal). The space in which the B_1 field is being measured must be filled by a nuclear magnetic resonance signal-bearing substance. This procedure implies that the fields that are being mapped are those in the signal-bearing substance. Some of the techniques presented require a homogeneous substance. Aqueous paramagnetic solutions; pure gels of gelatin, agar, polyvinyl alcohol, silicone, polyacrylamide, or agarose; organic doped gels; paramagnetically doped gels; and reverse micelle solutions have been proposed as hydrogen MRI phantoms (20). Each of these materials has its own application-specific advantages and disadvantages. When the material is water, the B_1 images can show a standing wave artifact. This artifact manifests itself as a stronger field at the locations of the standing waves. The artifact is most pronounced when the size of the object approaches

dimensions greater than half the wavelength of the radiation in water. The index of refraction of water is approximately 9 at 63 MHz, therefore the wavelength of this frequency in water is 53 cm. Materials that have high conductivity produce skin-effect artifacts. The B_1 field at the surface of a sphere of radius r filled by a conducting solution of resistivity ρ and permeability μ is B_1'. The field at any location x, measured from the surface along a radius of the sphere, is given by

$$B_1(x) = B_1' e^{-x/\delta}, \tag{6}$$

where δ is the skin depth of the photons of frequency ν in the solution:

$$\delta = \sqrt{\frac{\rho}{\pi \nu \mu}}. \tag{7}$$

At 63 MHz, the skin depth in normal saline is ~6 cm.

Solutions that minimize the conductivity and index of refraction, such as reverse micelle solutions of a water (8 mM $NiCl_2$)–AOT–decane solution, where AOT is the surfactant bis(2-ethylhexyl)sulfosuccinate sodium salt, have been proposed to minimize these artifacts (20). The imager produces an image or series of images, where the intensity of any pixel is an assumed function of B_1. Both the rotational and direct mapping techniques mentioned following are affected by inhomogeneities in the static magnetic field B_0 and nonlinearities in the gradient fields from the imager.

The spatial resolution of imaging techniques is limited by the voxel size in the image. This voxel size is determined by the selected tomographic slice thickness (Thk), the number of points in the image matrix ($NPTS_1 \times NPTS_2$), and the field of view (FOV). The dimensions of a voxel are $Thk \times FOV_1/NPTS_1 \times FOV_2/NPTS_2$. In practice, the SNR limits the voxel size.

Rotational Mapping

This technique is an imaging version of the probe NMR technique mentioned earlier (21). Instead of a point source sample, a volume sample is used. Images are recorded as a function of a series of rotational angles. The period of the oscillation in the signal as a function of the rotational angle is used to calculate B_1. In this technique, B_1 sensitivity is limited by the ability to measure the period of the oscillations in the signal intensity accurately.

A variation on this technique applies only a high-power pulse that rotates some of the magnetization by thousands of degrees. A single image of a spatially uniform signal-bearing sample will display a series of dark and bright stripes. Dark stripes correspond to θ values where n of Eq. 5 is even, whereas bright stripes correspond to odd values of n. If the region of minimum B_1 is known, B_1 can be calculated from Eqs. (4) and (5). This single image technique is not foolproof, because there is no a priori way to distinguish between a decreasing B_1 trend and an increasing trend.

An example of this technique is depicted in Fig. 3. An 18-cm diameter sphere full of 14 mM $NiCl_2$ was imaged in an inhomogeneous B_1 field from a 63-MHz magnetic resonance imaging coil. The original hydrogen magnetic resonance image shows bright bands of ~90° and ~270° rotational angles, and dark bands of ~0° and ~180° rotation. As the maximum rotational angle in the image increases, more bands are observed. Therefore, the accuracy of the technique increases when higher power RF or longer pulses are applied to the coil.

Direct Mapping

The amount of transverse magnetization and hence signal S produced by a one-pulse imaging sequence is proportional to the sine of the rotational angle (13). Substituting Eq. (4) for the rotational angle θ,

$$S \propto \sin(2\pi\gamma\tau B_1). \tag{8}$$

Therefore, measuring S as a function of B_1 for a few values on the sine curve will allow calculating B_1. This general procedure is more commonly implemented by using a spin-echo imaging sequence (13,23–26).

A spin-echo sequence consists of two RF pulses and a signal called the echo. The echo time, TE, is the time between the first RF pulse and the maximum in the echo. The two RF pulses rotate the magnetization by θ and then 2θ a time $TE/2$ later. The second pulse samples the magnetization present $TE/2$ after the first. The optimum signal is at $\theta = 90°$. The signal for this

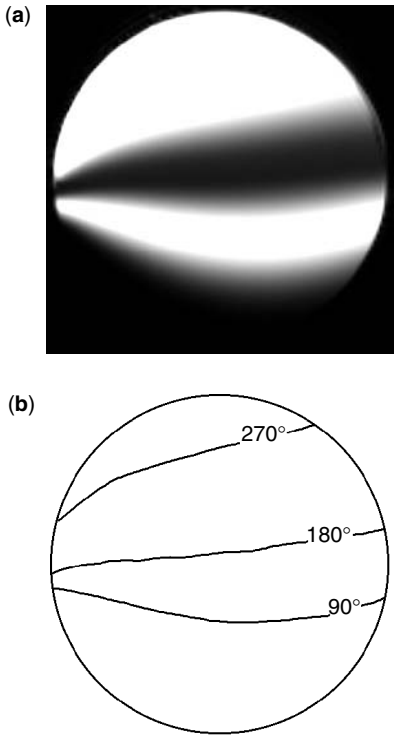


Figure 3. A magnetic resonance image (**a**) of a 18-cm diameter sphere full of 14 mM NiCl2 in an inhomogeneous B_1 field, and (**b**) a rotational angle map (with permission from Ref. 22).

sequence is proportional to $\sin^3\theta$, or substituting Eq. (3) for θ,

$$S \propto \sin^3(2\pi\gamma\tau B_1). \quad (9)$$

Images are taken at several B_1 values near $\theta = 90°$. Corresponding pixels in the images are fit with Eq. (9), and B_1 is determined. This process is repeated for every pixel in the image. Figure 4 displays the B_1 fields from a quadrature style bird cage coil (27) on a 1.5 tesla magnetic resonance imager. The fields were measured inside 27-cm diameter spheres full of 14 mM aqueous $NiCl_2$; a reverse micelle solution of water, AOT (a surfactant), and decane; and 154 mM aqueous NaCl containing 6 mM $NiCl_2$.

These three solutions exemplify a problem associated with the signal-bearing substance in a phantom. The inverted parabolic shape of the phantom full of 14 mM aqueous $NiCl_2$ is due to the standing wave artifact. The parabolic shape of the phantom full of 154 mM aqueous NaCl and 6 mM $NiCl_2$ is primarily due to the conductivity or skin-effect artifact associated with solutions of high electrical conductivity. The deviation from a perfect parabola is also due to the presence of a small standing wave artifact. The shape of the field in the phantom full of a reverse micelle solution more accurately represents the B_1 field in the imaging coil. The accuracy is improved because the standing wave and conductivity artifacts are minimized by the dielectric constant and conductivity of the reverse micelle solution (20).

Similar variations in B_1 have been reported for elliptical phantoms (28). The direct mapping technique does not require filling the resonator or coil with a homogeneous solution. In fact, B_1 fields have been measured inside the human head using this technique (29). Other imaging techniques have been proposed. The reader is directed to the scientific literature for information on them.

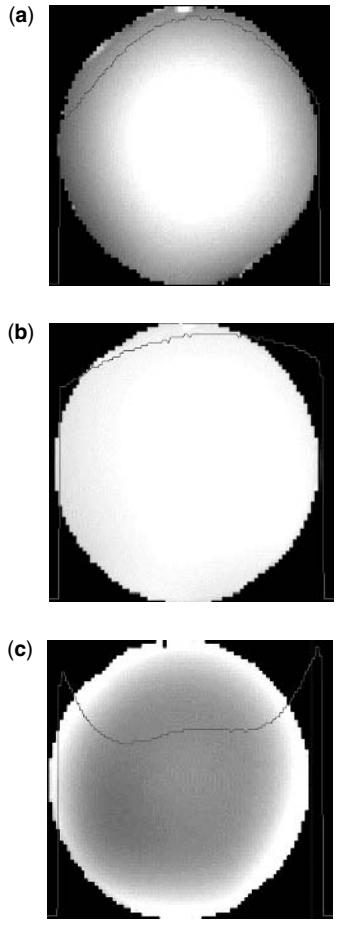


Figure 4. Images of the B_1 fields in a quadrature style bird cage coil on a 1.5 tesla magnetic resonance imager. The fields were measured inside 27-cm diameter spheres full of (**a**) 14 mM aqueous $NiCl_2$; (**b**) a reverse micelle solution of water, AOT (a surfactant), and decane; and (**c**) 154 mM aqueous NaCl and 6 mM $NiCl_2$. Red lines are plots of the B_1 field along a line drawn through the center of the sphere (15) (with permission from Ref. 22). See color insert.

ABBREVIATIONS AND ACRONYMS

AOT	the surfactant sodium bis(2-ethylhexyl)sulfosuccinate
B	magnetic field
B_1	amplitude of the RF magnetif field
δ	skin depth
δv	difference in frequency
EPR	electron paramagnetic resonance
FID	free induction decay
FOV	field of view
γ	gyromagnetic ratio
LCR	inductive-capacitive-resistive
MRI	magnetic resonance imaging
MHz	megahertz
mM	millimolar concentration
n	an integer
N	number of turns
ν	frequency
NMR	nuclear magnetic resonance
NPTS	number of points
$NiCl_2$	Nickel Chloride
r	radius
RF	Radio-frequency
SNR	signal-to-noise-ratio
τ	pulse width
θ	rotation angle
TE	echo time
Thk	slice thickness
V	voltage
W	power

BIBLIOGRAPHY

1. C. P. Schlicter, *Principles of Magnetic Resonance*, Springer-Verlag, NY, 1980.
2. E. Fukushima and S. B. W. Roeder, *Experimental Pulse NMR, A Nuts and Bolts Approach*, Addison-Wesley, Reading, MA, 1981.
3. C. P. Poole Jr., *Electron Spin Resonance*, Wiley, NY, 1983.
4. J. H. Freed, D. S. Leniart, and J. S. Hyde, *J. Chem. Phys.* **47**, 2762–2773 (1967).

5. E. L. Gintzon, *Microwave Measurements*, McGraw-Hill, NY, 1957, p. 448.
6. J. P. Hornak and J. H. Freed, *J. Magn. Resonance* **67**, 501–518 (1986).
7. M. Alecci, M. Ferrari, and C. L. Ursini, *Biophys. J.* **67**, 1274 (1994).
8. M. Alecci et al., *Phys. Med. Biol.* **43**, 1899–1906 (1998).
9. D. I. Hoult and P. C. Lauterbur, *J. Magn. Resonance* **34**, 425–433 (1979).
10. P. Mansfield and P. G. Morris, in J. Waugh, ed., *Advances in Magnetic Resonance*, Suppl. 2, Academic Press, NY, 1982.
11. C. Chen and D. J. Hoult, *Biomedical Magnetic Resonance Technology*, IOP, Bristol, 1989.
12. L. Fakri-Bouchet, C. Lapray, and A. Briguet, *Meas. Sci. Technol.* **9**, 1641–1646 (1998).
13. J. P. Hornak, J. Szumowski, and R. G. Bryant, *Magn. Resonance Med.* **6**, 158–163 (1988).
14. J-C. Ginefri, E. Durand, and L. Darrasse, *Rev. Sci. Instrum.* **70**, 4730–4731 (2001).
15. L. Darrasse and G. Kassab, *Rev. Sci. Instrum.* **64**, 1841–1844 (1993).
16. A. Leroy-Willig, L. Darrasse, J. Taquin, and M. Sauzade, *Magn. Resonance. Med.* **2**, 20–28 (1985).
17. J. P. Hornak, The Basics of NMR. © 2000 J. P. Hornak, *http://www.cis.rit.edu/htbooks/nmr/*.
18. D. I. Hoult, C. -N. Chen, and V. J. Sank, *Magn. Resonance Med.* **1**, 339–353 (1984).
19. J. Murphy-Boesch, G. J. So, and T. L. James, *J. Magn. Resonance* **73**, 293–303 (1987).
20. J. E. Roe, W. E. Prentice, and J. P. Hornak, *Magn. Resonance Med.* **35**, 136–141 (1996).
21. J. Murphy-Boesch, G. J. So, and T. L. James, *J. Magn. Resonance* **73**, 293–303 (1987).
22. J. P. Hornak, The Basics of MRI, © 2001, *http://www.cis.rit.edu/htbooks/mri/*.
23. *Method of Mapping the RF Transmit and Receive field in an NMR System.* US Pat. 5,001,428, (1991), J. K. Maier and G. H. Glover.
24. E. K. Insko and L. Bolinger, *J. Magn. Resonance Series A* **103**, 82–85 (1993).
25. G. H. Glover et al., *J. Magn. Resonance* **64**, 255–270 (1985).
26. G. J. Barker, S. Simmons, S. R. Arridge, and P. S. Tofts, *Br. J. Radiol.* **71**, 59–67 (1998).
27. C. E. Hayes et al., *J. Magn. Resonance* **63**, 622–628 (1985).
28. J. S. Sled and G. Bruce, *IEEE Trans. Med. Imaging* **17**, 653–662 (1998).
29. X. Li, Tissue parameter determination with MRI in the presence of imperfect radiofrequency pulses. M.S. Thesis. Rochester Institute of Technology, 1994.

S

SCANNING ACOUSTIC MICROSCOPY

BERNHARD R. TITTMANN
CHIAKI MIYASAKA
The Pennsylvania State University
University Park, PA

INTRODUCTION

This article is an overview of scanning acoustic microscopy, which is one of the more advanced ultrasonic imaging technologies. The scanning acoustic microscope (hereinafter called simply 'SAM') uses ultrasound to produce enlarged images of the microscopic structures of materials. In other words, the SAM is an instrument that subjects an object to ultrasound and detects the variations of the elastic properties of the object. The elastic properties are determined by the molecular arrangement of the material, molecule size, intermolecular force or the like, and each material has its unique characteristics. Accordingly, each object reacts in its own peculiar way when subjected to ultrasonic waves. The waves received from the object can be converted into an image. Because ultrasonic waves can penetrate into an object's interior, the surface and also the interior can be visualized.

The contents of this article are as follows: First, the need in industry for this process is described. Second, the history of the development of the SAM is illustrated in brief. Third, the basic theory of ultrasonic imaging used in the SAM is depicted, as well as the basic theory of the implements used in the SAM. Furthermore, the $V(z)$ curve method for material characterization is described. Selected images are given for a better understanding of the sections.

WHY SAM?

Recently, corresponding to growth in the development of material science and technology, the requirements for quality assurance of materials have become more stringent, especially in aerospace components and microelectronics. It is sometimes difficult for conventional materials to meet these new requirements completely. In this case, the use of new materials (composite materials, coated materials, thin films, metal alloys), is often discussed. However, the discussion comes to a close if no data for securing the reliability and safety of the new materials exist.

Generally speaking, when evaluating a material, test pieces for life estimation, quality, and safety analyses are prepared. Data showing material characteristics are typically collected by optical microscopes and scanning electron microscopes through destructive tests such as pull, compression, bending, torsion, creep, fatigue, impact, and corrosion. However, critical defects such as voids, delaminations, and debonding sites of the new materials often exist in the interior. Therefore, conventional microscopes cannot observe them directly. Hence, defects must be exposed by cutting or grinding the specimens in layers which causes the following problems.

1. The shape of the defects might be changed when cutting or polishing the specimens.
2. There is no guarantee that all defects are exposed.
3. Certain specimens are difficult to cut or grind.
4. When the sizes of the defects are less than 10 μm, it is difficult to see their 3-D distributions by grinding in layers.

The SAM is also very useful in the biomedical field where many new treatment methods and new medications have been developed. Obviously, their effects and side effects need to be studied qualitatively and quantitatively. Optical microscopes are often used for observing such data from tissues or cells taken from patients as specimens. However, the tissues or the cells need to be chemically stained for optical microscopic observation. These stains usually kill the cells. Therefore, when the stains are applied, it is difficult to understand the real effects of the treatment or the medicine because the transformation in the tissue may only be visible in living cells. Furthermore, when measuring the elastic properties of tissues by conventional techniques, large volumes of the tissues are required. It is not realistic to take such amounts from patients suffering from diseases.

It has been very popular to visualize internal structures of opaque objects nondestructively and to obtain the acoustic characteristics of specimens by instruments using the basic features of ultrasound (reflection, transmission, refraction, and diffraction). However, the frequencies of the ultrasound used for these instruments range typically from 20 kHz to 10 MHz, limiting the resolution of the process. Furthermore, most of their transducers do not have features to focus ultrasound so that the diameters of the ultrasonic beams would be small enough to form an image that has the resolution and/or the contrast of optical microscopes. In addition, traditional ultrasonic imaging cannot provide data on the elasticity of the small area of a specimen.

Therefore, scanning acoustic microscopy is a viable nondestructive method that has high-resolution defect visualization to determine elastic parameters in both the medical and the industrial fields. Because the SAM uses acoustic waves, it can penetrate nondestructively into optically opaque materials. In addition, the contrast mechanism of the SAM depends only on the reflection from objects that have different acoustic properties and, thus, chemical agents for staining the tissues or the cells are not necessary, so that living tissues or cells

may be clearly observed. Because the beam is focused (up to submicrometer size) and has a high frequency (10 MHz to 3 GHz), the resolution of the system is of the order of optical microscopes. In addition to image formation, the SAM can detect the amplitude and the phase of a reflected wave. By analyzing them, the elastic properties of a specimen can be quantitatively determined.

BRIEF HISTORY OF SAM

In 1949, Sokolov suggested the possibilities of SAM and indicated the advantages such a technology would have compared to conventional optical light microscopes. He made a remarkable statement in his article that a microstructure of a material could be microscopically visualized when using ultrasounds that have a frequency close to 3 GHz or more (1). However, the realization of a practical SAM was delayed until the development of electronic imaging, high-frequency processing, and an acoustic lens.

In 1973, Lemons and Quate developed a mechanically scanning acoustic transmission microscope, including two acoustic lenses (one for emitting focused ultrasound into a specimen and another for receiving the ultrasounds transmitted from the specimen) and a RF generator that sends a high frequency of 1 GHz or more (2). For this type, a specimen is located between the lenses, so that the thickness of the specimen is limited. To overcome the disadvantage, the reflection mode of a mechanically scanning acoustic microscope that had one acoustic lens was developed (3). This type of SAM has been widely accepted as standard hardware in industries because of its simple system configuration and capability of forming a highly resolved image.

In 1979, Weglein found that a change in the voltage of an acoustic lens, which is the so-called $V(z)$ curve, when defocusing the acoustic lens toward a specimen, is unique to the elastic properties of the specimen (4). This phenomenon was modeled with ray tracing by Parmon and Bertoni (5), and with Fourier optics by Atalar (6). They also, found that the velocity of surface acoustic waves traveling within the spot of an ultrasonic beam could be obtained by measuring the distance between the periods of the $V(z)$ curves. By the establishment of these basic theories, the SAM became the accepted apparatus for imaging and also for quantitative data acquisition.

After its development, the SAM was modified or newly designed to improve the resolution, contrast, penetration depth, and capability of quantitative data acquisition, as well as to design the SAM for a specific application.

Resolution

The resolution of an image formed by the SAM is determined by the frequency of the wave, the velocity of the coupling medium, and the lens geometry. Therefore, three approaches were pursued to increase the resolution.

The first approach was to raise the acoustic frequency. By this approach, Hadimiooglu and Quate achieved a resolution of 0.2 μ by operating at about 4.4 GHz, and the specimen was in boiling water to minimize the attenuation of the coupling medium (7). The second approach was to use low-velocity coupling media. Using the second approach, referred to as cryogenic microscopy, Foster and Rugar were able to obtain a resolution of 20 nm using superfluid liquid helium as a coupling medium at a temperature of 0.2 °K (8).

These approaches are excellent for increasing resolution on the surface of a specimen, but not necessarily in its interior. In the first approach, the ultrasonic wave is attenuated in proportion to the square of its frequency; therefore, when the high-frequency ultrasonic wave is used as a probe, the wave will not penetrate inside the specimen, and the internal information may not be obtained. In the second approach, the acoustic impedances of the coupling medium and the specimen are very different, so that most of the ultrasonic waves are reflected at the interface between the coupler and the specimen, and it is unlikely that an image may be obtained from any significant depth (9). Hence, acoustic microscopy's most important and unique feature — namely, obtaining subsurface information — may not be realized by these two approaches.

The third approach, which was to design a new acoustic lens, has been considered the best approach for maintaining the most important feature of the SAM. By the third approach, the following new acoustic lenses were proposed. Chubachi et al. developed the concave transducer to increase resolution by removing spherical aberration (10). Davids et al., Atalar et al., Yakub et al., and Miyasaka et al. proposed new acoustic lenses that have restricted apertures using Rayleigh waves (11), a nonspherical aperture using Lamb waves (12), a pinhole aperture for near-field imaging (13), and a center-sealed aperture substantially using shear waves (14), respectively.

Contrast

Apart from resolution increase, contrast enhancement is another way to increase imaging quality. Daft et al., Chou et al. and Atalar et al. proposed an acoustic lens having a large transducer for phase-contrast imaging (15), a shear-polarized transducer for increasing the contrast of acoustic images (16), and a slit aperture for pseudoline focusing to image the anisotropy of materials (17), respectively.

Penetration Depth

In the field of plastic matrix composites, especially fiber-reinforced plastic (FRP) laminates, it is necessary to detect delaminations located at an interface between laminates. When inspecting microelectronic integrated circuits (IC), it is necessary to form a cross-sectional image to display the details of defects (e.g., voids, delaminations, microcracks) located within an IC package. Materials commonly used for these have highly attenuating properties, so that a frequency of 100 MHz or higher may not penetrate into a portion that must be observed. Therefore, to meet the requirements, a new SAM, which is an improved version of a C-scan imaging apparatus, was designed to use the time-of-flight method (pulse-wave mode) in the lower frequency

range of 10–100 MHz (18–21). This SAM can perform an A scan and a B scan as well as a C scan, so that conventional ultrasonic techniques can be used directly. Therefore, this type of SAM is the most widely used in various industries (22–26).

Quantitative Data Acquisition

Quantitative data (i.e., velocities of surface acoustic waves) obtained by a spherical lens, which is commonly used for acoustic imaging, may not provide all the information on the elastic anisotropy of materials. Kushibiki et al. developed a cylindrical lens (called line-focus lens) and a measuring method to overcome this issue (27,28).

Liang et al. developed a SAM that can obtain complex $V(z)$ curves, based on a nonparaxial formulation of the $V(z)$ integral, and a reflectance function of a liquid–solid interface by inverting the $V(z)$ curves formed at frequencies of 10 MHz or less (29). Endo et al. improved this SAM in terms of mechanical movement and increased frequency range (up to 3.00 GHz) to perform more precise measurement (30).

Kulik et al. developed a SAM for continuous-wave operation that obtained $V(z)$ curves having better signal-to-noise ratios (31).

Not adopting the $V(z)$ curve techniques, Tsukahara et al. developed a SAM, a so-called ultrasonic microspectrometer, for accurately measuring the acoustic reflection coefficient of isotropic and/or anisotropic materials that have frequencies of 100 MHz or more (32).

Modifications for Applications

Phase-Contrast Imaging. An acoustic image can be formed by using either the amplitude or phase of the ultrasonic signals reflected or transmitted from a specimen. The amplitudes were often used to form an image because of the simplicity of designing electric circuits. However, phase contrast can often provide significant information such as residual stress distribution.

Nikoonahad developed the tilted transducer to obtain differential phase contrast (33). Meeks et al. developed a phase-sensitive SAM, which operates across a large bandwidth into the GHz regime (34).

Use of Various Coupling Media. Water is commonly used as a coupling medium for the SAM. However, some specimens allow only dry conditions. Attal proposed liquid metals (e.g., mercury or gallium) as coupling media (35). Wickramasinghe and Petts developed a SAM using gases (e.g., argon and xenon) as coupling media under elevated pressure (36). To increase resolution, Karaki and Sakai developed an acoustic microscope using liquid nitrogen instead of superfluid liquid helium and obtained images at 1.0 GHz (37). Yamanaka et al. developed a low-temperature microscope to study a 0.5-mm thick epoxy bond between two 1-mm PMMA sheets at 40 MHz at $-50\,^{\circ}C$ using methyl alcohol as a coupling medium (38). These types of SAMs were designed for use strictly in a laboratory because they required environmental chambers for the coupling media.

For industrial use, the air-coupled SAM is desired, although its resolution is limited. The challenge is the transfer of acoustic energy across the solid/air interface. The calculated transmission coefficient between a steel transducer face plate and air is equal to only 1.797×10^{-5}. This means that 99.996% of the power is reflected back from the interface. Therefore, an appropriate transducer needed to be developed. Fox et al. manufactured transducers using silicone rubber as the front acoustic impedance-matching layer and operating frequencies at 1 and 2 MHz (39). By using such a transducer operating frequency at 1 MHz, the distance in air could be measured from 20 mm to 40 mm with an accuracy of 0.5 mm. Further improvements in transmission efficiency were shown by planting an acoustic impedance-matching layer that is composed of tiny glass spheres in a matrix of silicone rubber on piezoelectric elements (40,41). Reilly and Hayward reported air-coupling transducers using piezoelectric composites and operating with tone-burst waves whose frequency ranged from 250 kHz to 1.5 MHz. They were successful in producing transmitted signals (milivolt level) through a composite laminated honeycomb structure at 500 kHz (42). Shindel et al. developed a micromachined capacitance air transducer characterized by high bandwidths for evaluating composite materials (43). The micromachined capacitance air transducer could be operated at a frequency of 11.4 MHz (44). Bhardwadj et al. equipped a transducer for noncontact SAM (reflection type) (45).

Elevated Temperature. Observation of changes of interior portions of materials due to a temperature increase had been desired when using the SAM. Miyasaka and Tittman developed a SAM using operating temperatures up to $500\,^{\circ}C$ to observe a specimen located in motor oil (46). Ihara et al. developed a SAM using operating temperatures up to $920\,^{\circ}C$ for liquid metals (47). However, such issues as frequency increase still remain, to be addressed.

Nonlinear Imaging. The sharply convergent acoustic beam used in the SAM can provide sufficient intensity to produce strong nonlinear effects at microwave frequencies. The first nonlinear imaging was performed by using a mechanically scanned transmission acoustic microscope. Second-harmonic acoustic radiation generated in the vicinity of the beam focus is readily detected and used to form an image of an object placed in the focal plane (48). To obtain nonlinear data, the receiving transducer is tuned to either the second or third harmonic. When harmonics are produced, the resultant images have a complex appearance. However, a significant portion of the detected signal is due to nonlinearity not associated with specimen, but, for example, due to the water that serves as a coupling medium. To minimize these phenomena, Yeack et al. used a three-lens scanning acoustic microscope (transmission type) to achieve nonspecular oblique incidence. This enables mixing two independent inputs to occur only in the specimen, thus eliminating the contribution from, for example, the water coupling medium (49). Germain and Cheek used high-order harmonic acoustic radiation to form nonlinear images by using a mechanically scanned reflection acoustic microscope (50) and evaluated biological media (51). Tan et al. constructed a micron wavelength

acoustic microscope to observe third-order nonlinear products (52). More recently, nonlinear effects have been studied in pressurized superfluid helium in which resolution was improved beyond the diffraction limit (53–56). Most recently, several scientists studied nonlinear propagation applied to acoustic tomography (57–59).

PRINCIPLE OF THE SCANNING ACOUSTIC MICROSCOPE

Imaging Mechanism

In this section, the imaging mechanism of the mechanically scanned reflection microscope using water as a coupling medium is described.

The SAM has two modes. One is the so-called "tone-burst-wave mode" that uses tone-burst waves (see Fig. 1b), and operates substantially in the frequency range between 100 MHz and 3 GHz. In this frequency range, the wavelengths of the waves in water range from 15.0 to 0.5 μm (see Table 1). Therefore, an image formed in the tone-burst-wave mode is highly resolved. However, the penetration depth of the waves is limited by attenuation. The tone-burst-wave mode is for penetrations up to 300 μm. Hence, the tone-burst-wave mode is applicable for observing the surface and the subsurface of specimens. The other mode is the so-called pulse-wave mode that uses a spike pulse wave or only a few cycles (see Fig. 1a) and operates substantially in the frequency range between 10 and 100 MHz. Therefore, resolution of an image formed in this mode is limited. However, the pulse-wave mode is for penetrations up to 1 cm. Hence, the pulse-wave mode is usually applicable for observing the deep interior of specimens.

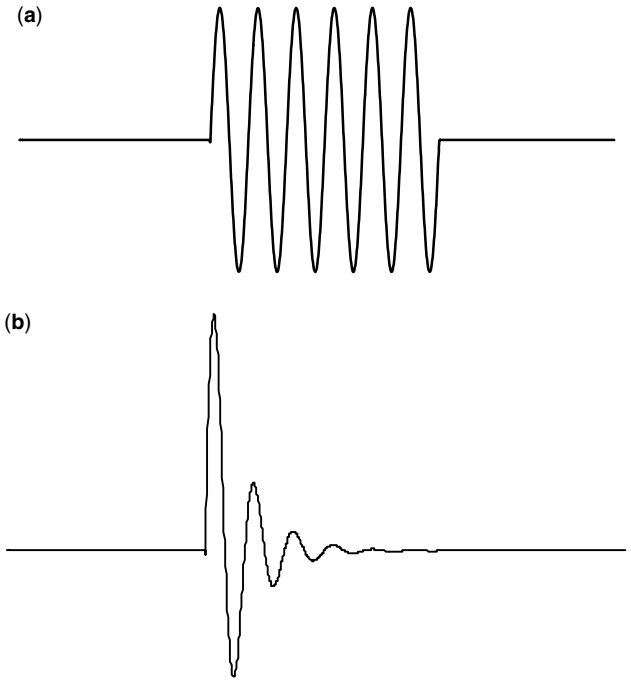


Figure 1. Waveform: (**a**) Tone-burst wave. (**b**) Pulse wave.

Table 1. Wavelength in Water at Various Frequencies

Frequency	Wavelength (μm)
100 MHz	15.0 μm
200 MHz	7.5 μm
400 MHz	3.7 μm
600 MHz	2.5 μm
800 MHz	1.8 μm
1.0 MHz	1.5 μm

Tone-Burst-Wave Mode. Figure 2 is a schematic diagram of the SAM (tone-burst-wave mode). The imaging mechanism of the tone-burst-wave mode in Fig. 2 is described here.

An electrical signal is generated by a RF tone-burst source. In the beginning, the Colpitts oscillator, or the like, was used as the RF tone-burst source. Nowadays, burst waves are gated out from continuous waves by a single pole double-throw (SPDT) switch for frequency stability. The output (i.e., voltage) of the source is approximately 10 V.

The electrical signal is transmitted to a piezoelectric transducer located on the top of a buffer rod through a circulator (or the SPDT switch). The electrical signal is converted to an acoustic signal (i.e., ultrasonic plane wave) by the transducer. The ultrasonic plane wave travels through the buffer rod to a spherical recess (hereinafter called simply "lens") located at the bottom of the buffer rod, wherein the lens is coated by the acoustic impedance-matching layer which is the so-called acoustic antireflective coating (hereinafter called simply "AARC"). The lens converts the ultrasonic plane wave to an ultrasonic spherical wave (i.e., ultrasonic beam). The ultrasonic beam is focused within the specimen and reflected from the specimen. The reflected ultrasonic beam, which carries acoustic information from the specimen, is again converted to an ultrasonic plane wave by the lens. The ultrasonic plane wave returns to the transducer through the buffer rod. The ultrasonic plane wave is again converted to an electric signal at the transducer. The voltage of the electric signal ranges from 300 mV to 1 V. When the operating frequencies range from 100 MHz to 1 GHz, the corresponding insertion losses are approximately 30 and 80 dB. Therefore, the electric signal must be amplified by 30 to 80 dB at the receiver. Furthermore, the electric signal is comprised of transmission leaks, internal reflections from the interface between the lens and the AARC, and reflections from the specimen. Therefore, the reflections must be selected by a rectangular wave from a double balanced mixer (DMB), called the first gate. Then, the peak of the amplitude of the electric signal is detected by a circuit, which includes a diode and a capacitor (i.e., the peak detection technique). The gate noise is removed by using the second gate existing within the first gate (the blanking technique). The peak-detected signal is stored in a memory through an analog-to-digital signal (hereinafter called simply "A/D") converter. The stored signal is again converted into an

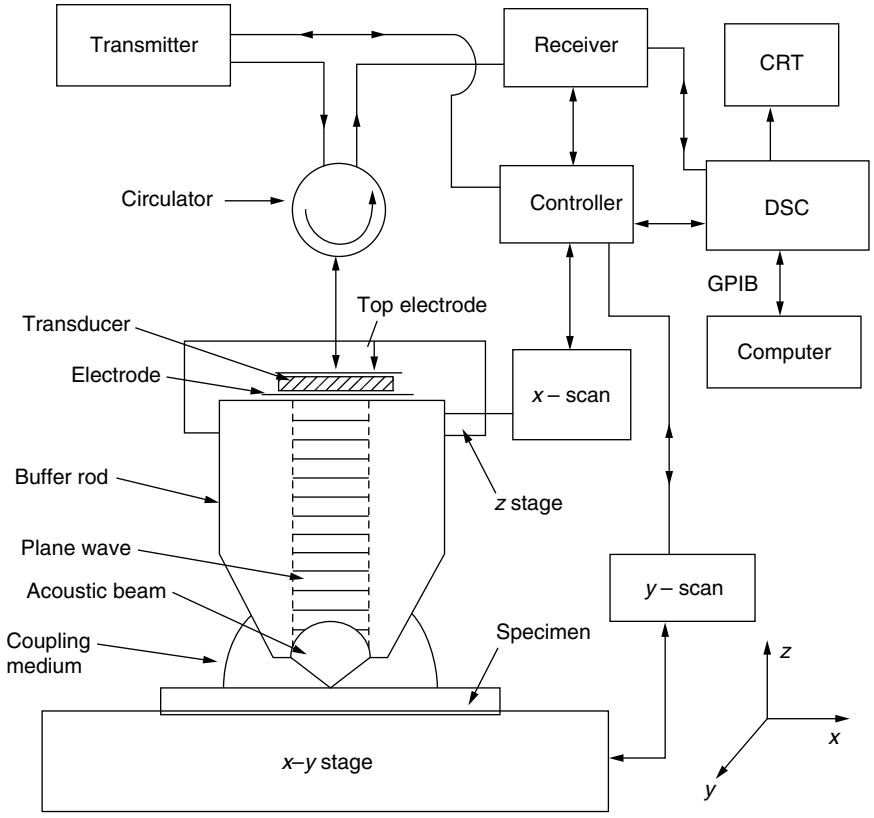


Figure 2. Schematic diagram of a scanning acoustic microscope (burst-wave mode).

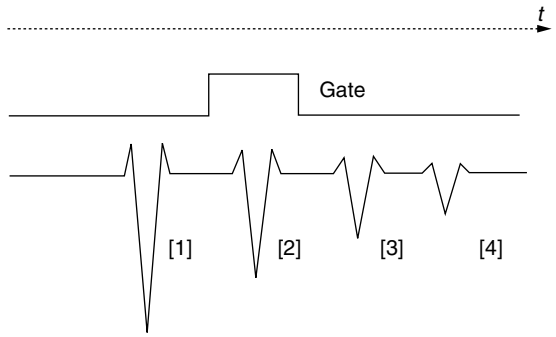
analog signal by a digital-to-analog signal (hereinafter called simply "D/A") converter. This flow of processes allows the information that is collected at a single spot on a specimen to be displayed as intensity at a certain point on the TV monitor.

To form a two-dimensional acoustic image, an acoustic lens and/or an x—y stage is mechanically scanned across a certain area of the specimen.

The acoustic lens can translate axially in the z direction to vary the distance between the specimen and the lens for subsurface visualization, that is, when visualizing the surface of the specimen, the acoustic lens is focused on the specimen (we denote $z = 0$ μm), and when visualizing a subsurface of the specimen, the acoustic lens is mechanically defocused toward the specimen (we denote $z = -x$ μm where x is the defocused distance).

Pulse-Wave Mode. The imaging mechanism for the pulse-wave mode is mostly the same as that for the tone-burst-wave mode. The significant differences are described here.

Suppose that a pulse wave emitted from an acoustic lens through a coupling medium (e.g., water) is focused onto the back surface of the specimen. Suppose that the pulse wave is strong enough to travel through the specimen and reflect back to the acoustic lens. If the specimen is a layered material, the timing of each reflection is different because the travel distance is different. As shown in Fig. 3, an oscilloscope may monitor both the amplitude and the delay of the pulse wave reflected from each plane of the specimen. Note that the output of an electrical



[1] Reflected waveform of front surface
[2] Reflected waveform of first interface
[3] Reflected waveform of second interface
[4] Reflected waveform of rear surface

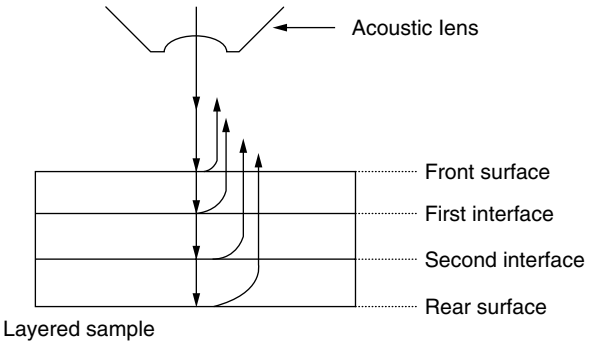


Figure 3. Timing of pulse waves reflected from a specimen, and a gate that selects the first interfacial reflection.

pulse signal generated from a RF source is approximately 50–300 V. The pulse wave can be focused through the specimen of interest. For example, when the first interface needs to be clearly visualized to detect a defect such as a delamination, the pulse wave is focused onto the first interface to maximize the amplitude of the pulse wave reflected from the first interface. Then, as shown in Fig. 3, the reflected pulse wave is electrically gated out to visualize the first interface as a horizontal cross-sectional image by horizontally scanning the acoustic lens or the x—y stage.

Acoustic Lens Design

Acoustic Lens for Tone-Burst-Wave Mode. The acoustic lens comprises a piezoelectric transducer and a buffer rod. The transducer is deposited on the top of the buffer rod, and a spherical recess covered by an acoustic antireflective coating is located at the bottom of the buffer rod.

Piezoelectric Transducer. The role of the piezoelectric transducer is to convert electric signals to acoustic signals and back again. Zinc oxide is typically used for the transducer at frequencies of 100 MHz or more. The transducer is sputtered onto the electrode (lower portion) that is deposited on the polished surface of the buffer rod. The electrode material is either chromium–gold or titanium–gold. The electrode (upper portion) is deposited on the zinc oxide using a mask plate to match the centers of the electrode (upper portion) and that of the lens.

When an electric signal is input onto the transducer, the transducer is excited at its resonance frequency f_r. The condition of the resonance is expressed as

$$\frac{2n-1}{f_r} = \frac{2d}{c}, \tag{1}$$

where d is the thickness of the transducer, n is a natural number, and c is the velocity of the longitudinal wave of the transducer.

Therefore, by modifying Eq. (1), the thickness d of the transducer is determined from

$$d = \left(\frac{2n-1}{2}\right)\left(\frac{c}{f_r}\right) = \left(\frac{2n-1}{2}\right)\lambda, \tag{2}$$

where λ is the wavelength of the acoustic wave.

However, the relationship between the acoustic impedance of the buffer rod and that of the transducer is also considered to determine the thickness. Let us denote the acoustic impedances of the buffer rod and the transducer as Z_1, and Z_2 respectively. Then, the relationship of these impedances can be either $Z_1 > Z_2$ or $Z_1 < Z_2$. The thickness d is determined as $\lambda/4$ when the relationship is $Z_1 > Z_2$, and $\lambda/2$ when $Z_1 < Z_2$.

Buffer Rod. The diameter of the buffer rod is determined to be larger than the diameter of the transducer. The length of the buffer rod is generally determined to be longer than the near-field region (Fresnel zone) calculated by

$$N = \frac{D^2}{4\lambda}, \tag{3}$$

where N is the near-field region, D is the diameter of the transducer, and λ is the wavelength of the material used for the buffer rod.

Considering acoustic energy loss, a material that has high velocity compared to that of a coupling medium and low attenuation (e.g., fused quartz, sapphire, or the like) is generally selected. The selection of the material directly relates to the design of the lens in terms of the spherical aberration that needs to be reduced.

Lens. The paraxial focal distance of the acoustic lens (denoted as "F_0") is approximately expressed as

$$F_0 = \frac{R}{(1-C')}, \tag{4}$$

$$C' = \frac{C_2}{C_1}, \tag{5}$$

where R is the radius of curvature of the surface of the lens, C_1 is the longitudinal wave velocity of the buffer rod, and C_2 is the longitudinal wave velocity of the coupling medium.

The radius of curvature of the surface of the lens is determined by considering the operating frequency and the attenuation coefficient. Table 2 shows the range of radii corresponding to various frequencies. The aperture angle of the lens (denoted as "θ_α") is determined by the focal distance. When the focal distance and the aperture angle are known, the radius of the aperture (denoted as "r") is determined by

$$r = F_0 \sin\left(\frac{\theta_\alpha}{2}\right). \tag{6}$$

The spherical aberration of an acoustic lens is illustrated in Fig. 4. When the acoustic wave from the lens is emitted into the specimen, the following equation is obtained by Snell's law:

$$C_1 \sin\theta' = C_2 \sin\theta, \tag{7}$$

where θ is the incident angle of the acoustic wave and θ' is the refracted angle of the acoustic wave.

Then, the zonal focal distance (denoted as "F") is expressed as

$$F = R\left\{(1-\cos\theta) + \frac{\sin\theta}{\tan(\theta-\theta')}\right\}. \tag{8}$$

Table 2. Range of Radius due to Frequency

Frequency	Radius
100 MHz	1–2 mm
200 MHz	500 μm–1 mm
400 MHz	500 μm
800–1,000 MHz	125 μm
1.5–3 GHz	50 μm

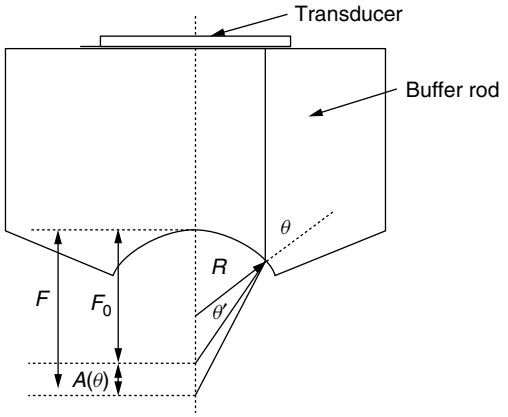


Figure 4. Calculation of the spherical aberration of the acoustic lens. $A(\theta)$ is the spherical aberration, F_0 is the paraxial focal distance of the acoustic lens, F is the zonal focal distance, R is the radius of curvature of the surface of the lens, θ is the incident angle of the acoustic wave, and θ' is the refracted angle of the acoustic wave.

Therefore, the spherical aberration denoted as "$A(\theta)$" is expressed as

$$A(\theta) = F_0 - F. \tag{9}$$

For example, when sapphire and water are used for a buffer rod and a coupling medium, respectively, $A(\theta)$ is calculated as $0.003\,R$, where R is the radius of curvature of the surface of the lens. For an acoustic lens operating at a frequency of 1.0 GHz, R is approximately 100 μm. Therefore, $A(\theta)$ is calculated as 0.3 μm. The value of wavelength of an ultrasonic wave that has a frequency of 1.0 GHz in water (1.5 μm) is small enough for a single spherical lens to form a focused image. Therefore, sapphire (z-cut) is considered the best material although it is difficult to make a spherical recess by mechanical polishing.

For reference, a comparison of the ratio of the spherical aberration and the radius (denoted as "$A(\theta)/R$") between an acoustic lens and an optical lens is shown in Fig. 5. The spherical aberration of an acoustic lens is much smaller than that of an optical lens.

Acoustic Anti Reflective Coating. When acoustic waves are emitted directly from sapphire into water that is used as coupling medium, 88% of the acoustic waves are reflected back from the interface between the lens and water because of their impedance mismatch. Therefore, it is necessary to coat the lens with an acoustic anti reflective coating (AARC). The thickness of the AARC should be $\lambda/4$, where λ is the wavelength of the acoustic wave, to have an optimized result, and the acoustic impedance Z is

$$Z = \sqrt{Z_1 Z_2}, \tag{10}$$

where Z is the impedance of the AARC, Z_1 is the impedance of the lens rod, and Z_2 is the impedance of the coupling medium.

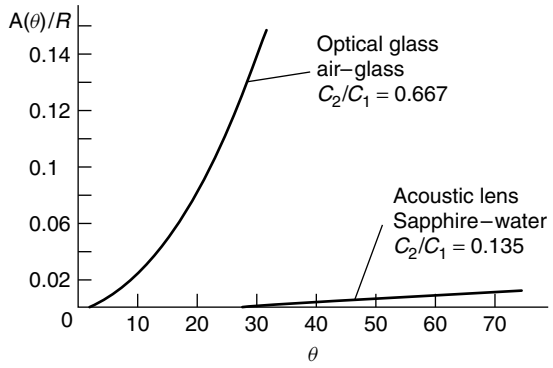


Figure 5. Spherical aberration of optical and acoustic lenses. $A(\theta)$ is the spherical aberration, R is the radius of curvature of the surface of the lens, C_1 is the longitudinal wave velocity of the buffer rod, and C_2 is the longitudinal wave velocity of the coupling medium.

Evaporated silicon oxide (SiO_2) is typically used as a material for the AARC although it does not completely satisfy Eq. (10).

Acoustic Lens for Pulse-Wave Mode. Most of the configurations of acoustic lenses for the pulse-wave mode are the same as those for the burst-wave mode. The significant differences are described here.

Piezoelectric Transducer. Lead zirconate titanate ceramic (PZT), lithium niobate ($LiNbO_3$), and polyvinyldifluoride, (PVDF) are typically used for transducer materials for operation of frequencies of 100 MHz or less.

Buffer Rod and Lens. This mode is used mainly for observing interior portions of a specimen. Therefore, it is important that the working distance (denoted as "W. D.") is designed to be long and the aperture angle of the lens is designed to be small. The working distance is defined as the distance from the position of the lens touching the surface of the specimen to the position of the lens when focused on the surface of the specimen. The maximum depth at which the material can be imaged is determined by the working distance. Table 3 shows the range of working distances corresponding to the range of frequencies. In the pulse-wave mode, the focal distance is designed to be long, so that the buffer rod is long. Considering its high cost, sapphire is usually not used for the buffer rod. Fused quartz, polymers, and ceramics are typically used.

Table 3. Range of Working Distances due to Frequency

Frequency	Working Distance
5–30 MHz	15–50 mm
30–50 MHz	10–20 mm
50–100 MHz	5–15 mm
More than 100 MHz	Less than 10 mm

Contrast and Application

Acoustic properties (i.e., reflection coefficient, attenuation, and velocity of acoustic wave) and the surface condition (i.e., surface roughness and discontinuities) of the specimen are factors in forming acoustic images.

Reflection Coefficient. The reflection coefficient at the surface of a specimen differs, depending on the materials. The difference in the reflection coefficient is the basic factor behind the formation of an acoustic image. The reflection coefficient is obtained by the reflectance function defined by the acoustic impedance and is determined by the type of material.

The acoustic impedance Z can be determined from

$$Z = \rho c, \tag{11}$$

where ρ is the density and c is the longitudinal wave velocity.

When acoustic waves are incident from liquid to solid at angle θ, some reflect from the solid, and some transmit into the solid (see Fig. 6). Let Z_1, Z_{2l} and Z_{2s} be the acoustic impedance of the liquid, the solid due to a longitudinal wave, and the solid due to a shear wave, respectively. Then, Z_1, Z_{2l} and Z_{2s} are expressed by

$$Z_1 = \frac{\rho_1 c_1}{\cos\theta}, \tag{12}$$

$$Z_{2l} = \frac{\rho_2 c_l}{\cos\theta_l}, \tag{13}$$

$$Z_{2s} = \frac{\rho_2 c_s}{\cos\theta_s}, \tag{14}$$

where ρ_1 is the density of the liquid, ρ_2 is the density of the solid, c_1 is the velocity of the longitudinal wave in the liquid, c_l is the velocity of the longitudinal wave in the solid, c_s is the velocity of the shear wave in the solid, θ is the reflection angle of the longitudinal wave in the liquid,

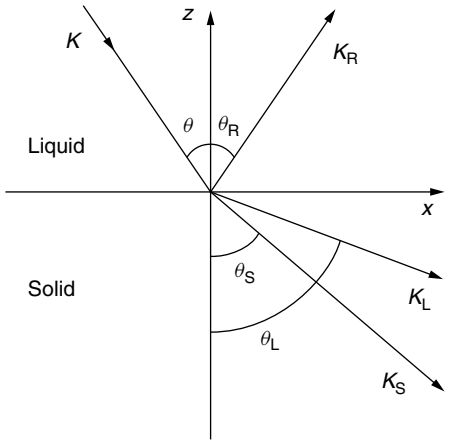


Figure 6. Reflection and transmission of a wave incident from a liquid to a solid. K, K_R, K_L, and K_S are wave vectors that represent the incident longitudinal wave, the reflected wave, the refracted longitudinal wave, and the mode-converted shear wave, respectively. θ, θ_R, θ_L, and θ_S are angles corresponding to the wave vectors.

θ_l is the refraction angle of the longitudinal wave in the solid, and θ_s is the refraction angle of the shear wave in the solid.

The following equations are obtained by Snell's law:

$$\sin\theta_l = \left(\frac{c_l}{c_1}\right)\sin\theta, \tag{15}$$

$$\sin\theta_s = \left(\frac{c_s}{c_1}\right)\sin\theta. \tag{16}$$

The reflectance function $\mathcal{R}(\theta)$ is expressed as

$$\mathcal{R}(\theta) = \frac{Z_{2l}\cos^2 2\theta_s + Z_{2s}\sin^2\theta_s - Z_1}{Z_{2l}\cos^2 2\theta_s + Z_{2s}\sin^2\theta_s + Z_1}. \tag{17}$$

When an acoustic wave is vertically incident from a liquid to a solid, the reflection coefficient $\mathcal{R}$ is determined as

$$\mathcal{R} = \frac{Z_{2l} - Z_1}{Z_{2l} + Z_1}. \tag{18}$$

From this equation, it can be seen that the larger the difference in acoustic impedance of two contacting materials, the greater the reflection coefficient at the interface.

Normally, water is used as coupling medium between the lens and a specimen. The acoustic impedance of water is lower than that of a solid; materials that have greater acoustic impedance show a higher reflection coefficient at the surface of a specimen, thus, forming image contrast. For example, suppose that an acoustic wave is perpendicularly incident onto fused quartz and sapphire immersed side by side in water. They have different elastic properties (their reflection coefficients are 0.632 and 0.873, respectively), but it is difficult to distinguish optically between them. However, the acoustic images show a significant difference in accordance with their reflection coefficients.

In addition, the acoustic impedance of a gas, such as air, is significantly lower. Thus, if a material has a void, a debonding, or a delamination, the acoustic wave will reflect strongly from the solid–gas interface. Many applications have been reported for visualizing voids, debondings, delaminations, inclusions, adhesion strength, powder distribution, deformation, and residual stress distribution, using this contrast mechanism. This nondestructive evaluation (NDE) technique is applicable to thin and thick layered films (60–62), electrical/electronic materials including integrated circuits (63–67), polymers (68–73), composite materials (74–76), ceramics (77–79), and other materials (80,81).

For example, Figs. 7 and 8 show debonding of reinforcing fibers from matrices. Figure 9 shows delaminations at the interfaces of carbon-fiber-reinforced plastics (C-FRP). Figure 10 shows delaminations and adhesion differences in a layered thin film. Figure 11 shows gaps around inclusions located at the interface of the substrate and the urethane coating. Figures 12 and 13 show deformations of composites. Figure 14 shows the powder distribution in composites.

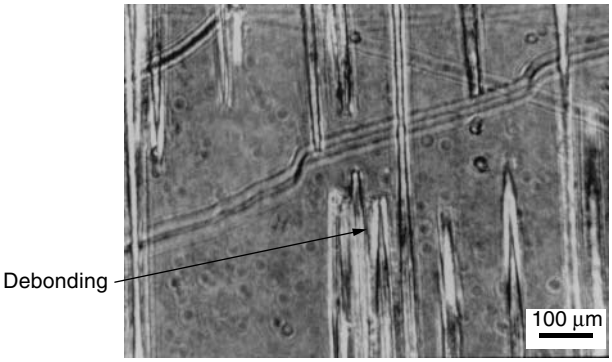


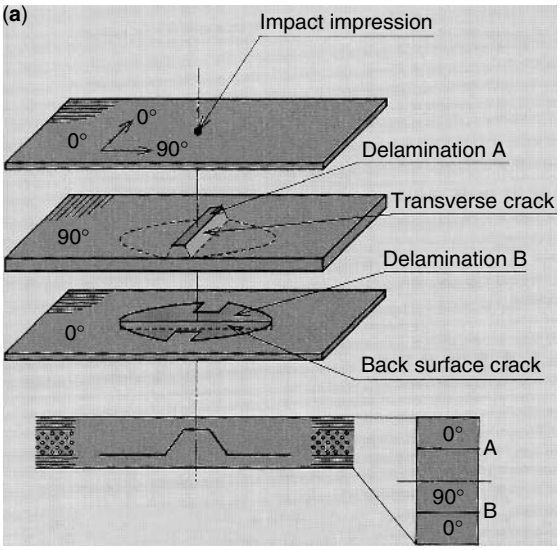
Figure 7. Debonding shown in metal matrix composites (MMC; fiber: carbon; matrix: aluminum alloy). The debondings are observed as bright regions at interfaces. Frequency: 200 MHz; defocusing distance: $z = -200$ μm.

Attenuation. When an acoustic wave travels within a specimen, its amplitude diminishes with travel distance and is finally not measurable. This phenomenon is called attenuation unique to the specimen and, mathematically, is described by

$$P_t = P_0 \exp(-\alpha z), \tag{19}$$

where P_t is the amplitude of a transmitted wave, P_0 is the amplitude of an incident acoustic wave, α is the attenuation coefficient, and z is the travel distance.

Figures 15 and 16 are acoustic images of living mouse cells (3T3) and a thinly sliced biological tissue (heart muscle). The acoustic impedance of these is close to that of water (or culture liquid). Therefore, virtually no contrast caused by the difference in reflection coefficient is displayed in these images. The contrast in the acoustic images of living cells or biological tissue is generated primarily from the difference in attenuation. Therefore, it is necessary to use a background composed of highly reflective material to make maximum use of the difference in attenuation. For example, when two types of substrates (one made of glass that has a longitudinal velocity of 5.6 km/s, a density of 2.49×10^3 kg/m^3, and a reflection coefficient of 0.581, and another of polystyrene that has corresponding values of 2.4 km/s, 1.05×10^3 kg/m^3, and 0.232, respectively) are used to image a biological specimen, the glass substrate theoretically provides better results. A variety of biological applications have been reported that use this contrast mechanism (82–93).



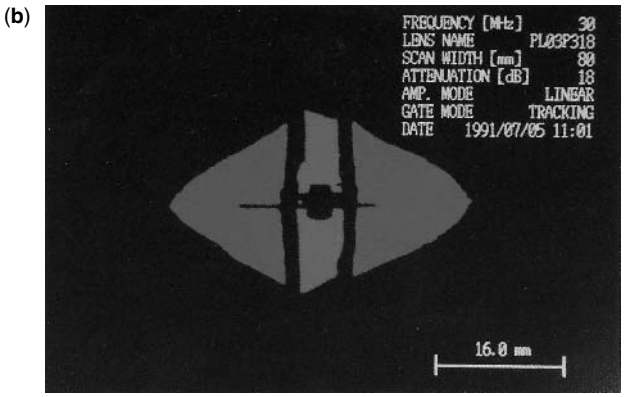


Figure 9. Delaminations at the interfaces of carbon-fiber-reinforced plastics (C-FRP). (**a**) Schematic diagram showing locations of delaminations. (**b**) Superimposed pulse-wave-mode images showing delaminations at interfaces. Delaminations were induced by the impact of energy of 1.71 J. Directions of fibers are shown in $[0^\circ_n/90^\circ_n]_{SYM}$.

Velocity of Surface Acoustic Waves (SAW). This parameter is very important for images obtained in the tone-burst-wave mode. Figure 17a is the acoustic image of polycrystalline manganese zinc ferrite that has an operating frequency of 400 MHz and has the acoustic lens focused on the surface ($z = 0$ μm). Also, Figs. 17b,c are the acoustic images of the same specimen where the acoustic lens is focused into the interior ($z = -25$ μm), and the same areas were scanned. Acoustic

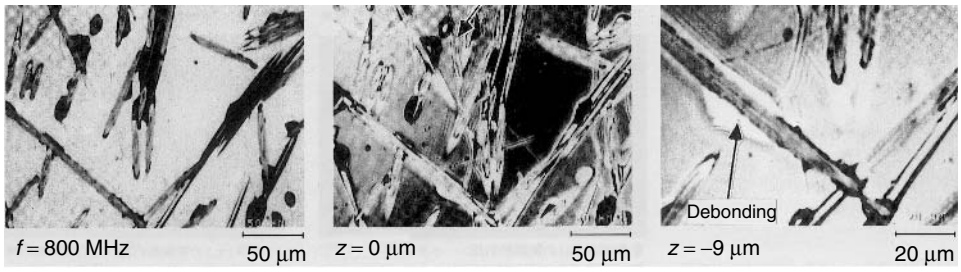


Figure 8. Debonding shown in plastic matrix composites (PMC; fiber: carbon; matrix: glass ceramics). The debondings are observed as bright regions at interfaces.

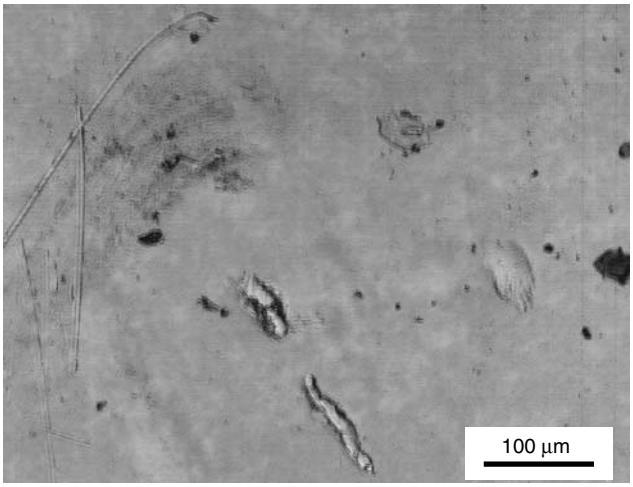


Figure 10. Delaminations and adhesion differences in a thin layered film. Frequency: 800 MHz. See color insert.

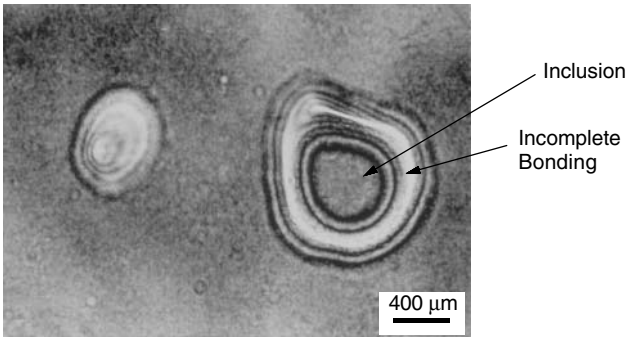


Figure 11. Inclusion at the interface of the urethane coating and the substrate. Frequency: 200 MHz.

lenses that have aperture angles of 120° and 60°, respectively, were used. Grains are not clearly seen in Figs. 17a,c, but are observed with significant contrast in Fig. 17b.

Furthermore, the contrast in the images formed by the acoustic lens that has the aperture angle of 120° changes in accordance with the position of the focal point controlled by the movement of the acoustic lens along the Z axis. This contrast mechanism is explained by the change of the receiving voltage in the transducer associated with the excitation of the SAW.

When the aperture angle of the acoustic lens is large, surface acoustic waves are generated because the incident angle goes beyond the Rayleigh critical angle. Therefore, the contrast in Fig. 17b is caused by the generation of the SAW. The aperture angle of the acoustic lens used for forming the image shown in Fig. 17c is not large enough to generate surface acoustic waves. Details will be given here.

V*(*z*) *Curve. Figure 18 shows a schematic diagram of an acoustic lens for expressing the $V(z)$ curve using an angular-spectrum approach (6).

When a plane acoustic wave that has an angular frequency ω is incident on the transducer plane, the acoustic field at the transducer plane is expressed as $u_0^+(x, y)e^{-i\omega t}$. Hereafter, we omit $e^{-i\omega t}$.

Acoustic fields of the planes (denoted as i, where $i = 0$, 1, 2, and 3) are expressed by a spatial distribution $u_i^{\pm}(x, y)$ where $i = 0$, 1, 2, and 3 or a frequency distribution $U_i^{\pm}(k_x, k_y)$, where $i = 0$, 1, 2, and 3. Superscripts $+$ or $-$ indicate the direction of field travel, that is in the $+z$ or $-z$ direction, respectively.

The angular spectrum at the transducer plane ($z = z_0$) is expressed as

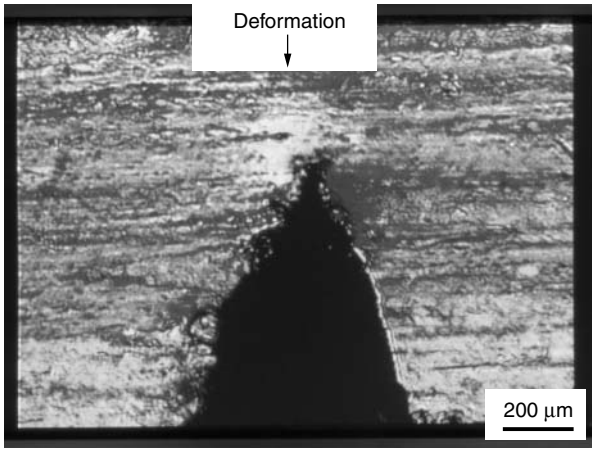


Figure 13. Deformation shown in metal matrix composites (MMC; whisker: SiC; matrix: aluminum alloy). Frequency: 600 MHz; defocusing distance: $Z = -10$ μm.

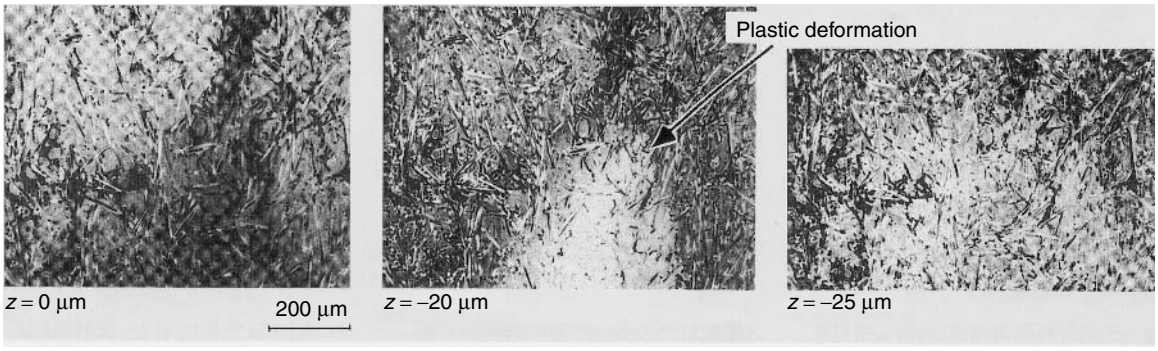


Figure 12. Deformation shown in plastic matrix composites (PMC; fiber: glass; matrix: nylon 6). Frequency: 200 MHz.

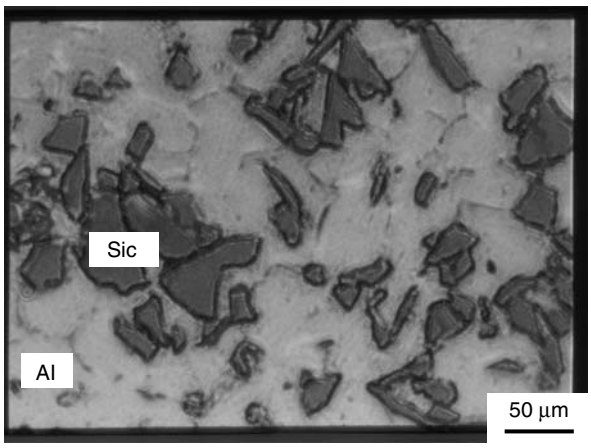


Figure 14. Powder distribution in metal matrix composites (MMC; powder: SiC; matrix: aluminum alloy). Frequency: 600 MHz; defocusing distance: $Z = -4$ μm.

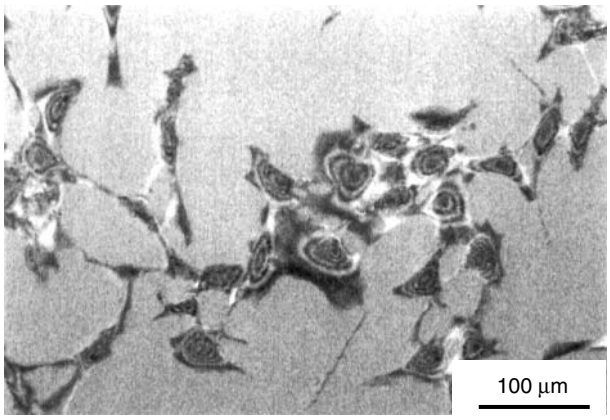


Figure 15. Living mouse cells (3T3). Frequency: 600 MHz; defocusing distance: $Z = -5$ μm; culturing medium: Dulebecco's modification of Eagle's medium: 1 × (Mod); temperature of the culturing medium: 37.5 °C. The cells that are grown on the glass substrate have enough acoustical reflectivity from the substrate.

$$
\begin{aligned}
U_0^+(k_x, k_y) &= \mathscr{F}[u_0^+(x,y)], \\
&= \iint_{-\infty}^{\infty} u_0^+(x,y) \exp[-i\{k_x x + k_y y + k_z(z - z_0)\}]\, dx\, dy, \\
&= \iint_{-\infty}^{\infty} u_0^+(x,y) \exp[-i(k_x x + k_y y)] \exp[ik_z(z - z_0)]\, dx\, dy, \\
&= \iint_{-\infty}^{\infty} u_0^+(x,y) \exp[-i(k_x x + k_y y)]\, dx\, dy. \qquad (20)
\end{aligned}
$$

Then, the relationships between the spatial and frequency distributions are generalized by the following equations:

$$U_i^{\pm}(k_x, k_y) = \mathscr{F}\{u_i^{\pm}(x,y)\}, \qquad (21)$$

$$u_i^{\pm}(k_x, k_y) = \mathscr{F}^{-1}\{U_i^{\pm}(x,y)\}. \qquad (22)$$

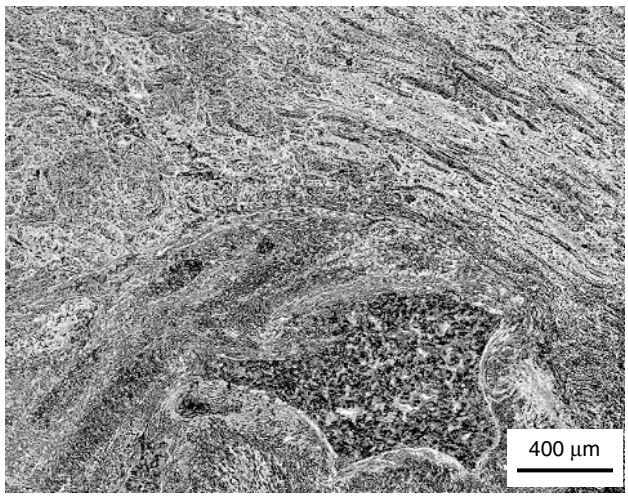


Figure 16. Human heart muscle. Frequency: 400 MHz. To prepare thin specimens, the tissue was embedded in paraffin and sectioned at 5 μm by the microtome. For scanning acoustic microscopy, the thinly sectioned specimen was affixed on the glass substrate, wherein the specimen was deparaffinized in xylene but not stained.

When ultrasonic waves travel from the transducer plane (z_0) to the back focal plane (z_1), only the phase related to the distance traveled ($|z_1 - z_0|$) is changed:

$$U_1^+(k_x, k_y) = U_0^+(k_x, k_y) \exp[ik_z(z_1 - z_0)]. \qquad (23)$$

The pupil function $\mathcal{P}(x,y)$, which shows the shape of the lens, is expressed as

$$\mathcal{P}(x,y) = \mathrm{circ}\left[\frac{(x^2 + y^2)^{1/2}}{R}\right], \qquad (24)$$

where R is the radius of the pupil and circ. is expressed as

$$\mathrm{circ}(r) = \begin{cases} 1, & r < 1 \\ 0, & r > 1 \end{cases}. \qquad (25)$$

Using the Fresnel approximation and the thin lens model, $u_2^+(x,y)$ is expressed as

$$
\begin{aligned}
u_2^+(x,y) = {} & \frac{\exp[ik_0 f(1 + \bar{c}^2)]}{i\lambda_0 f} \\
& \times \iint_{-\infty}^{\infty} u_1^+(x_1, y_1)\mathcal{P}_1(x_1 + x_2, y_1 + y_2) \\
& \times \exp\left[-i\frac{2\pi}{\lambda_0 f}(x_1 x_2 + y_1 y_2)\right] dx_1\, dy_1, \qquad (26)
\end{aligned}
$$

where k_0 is the wave number in the coupling medium, f is the front focal distance, λ_0 is the wavelength in the coupling medium, $\mathcal{P}_1$ is the pupil function from the lens to the specimen, and $\bar{c}$ is the ratio expressed by

$$\bar{c} = \frac{C_2}{C_1}, \qquad (27)$$

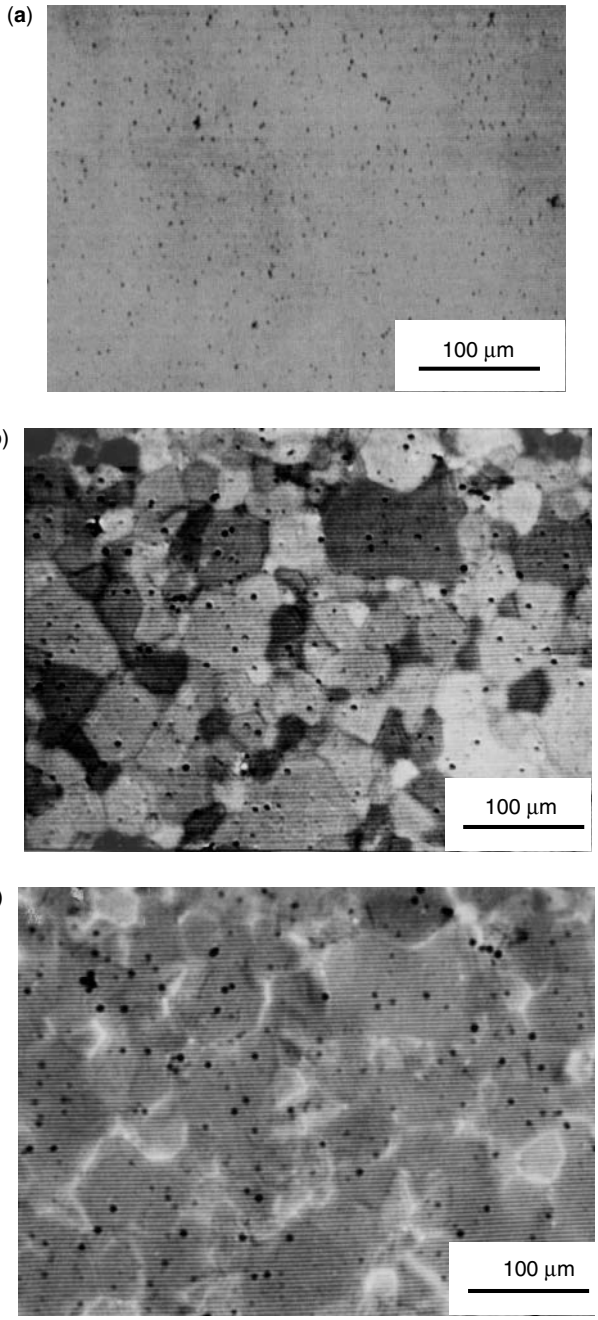


Figure 17. Manganese zinc ferrite. Frequency: 400 MHz; (**a**) $z = 0$ μm (aperture angle is 120°); (**b**) $z = -25$ μm (aperture angle is 120°); (**c**) $z = -25$ μm (aperture angle is 60°).

where C_1 is the longitudinal wave velocity of the buffer rod and C_2 is the longitudinal wave velocity of the coupling medium.

Supposing that $x_2 \ll x_1$ and $y_2 \ll y_1$ at the front focal plane, the pupil function from the lens to specimen is expressed as

$$\mathcal{P}_1(x_1 + x_2, y_1 + y_2) \cong \mathcal{P}_1(x_1, y_1). \tag{28}$$

Then Eq. (26) is rewritten as

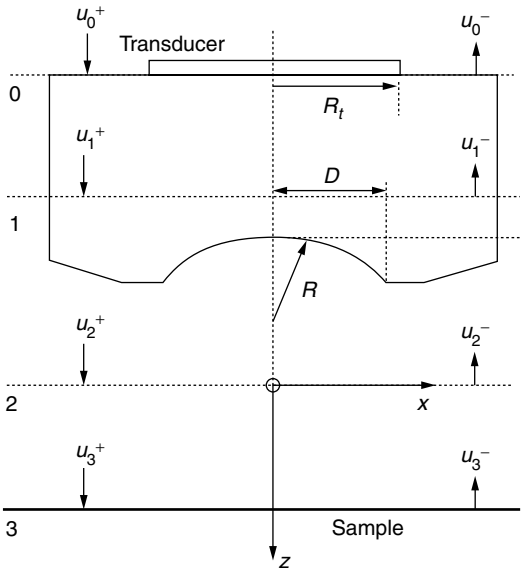


Figure 18. Schematic diagram of an acoustic lens for expressing the $V(z)$ curve by an angular-spectrum approach. 'Plane 0' is the transducer plane, 'Plane 1' is the back focal plane, 'Plane 2' is the front focal plane, 'Plane 3' is the surface of the specimen, and $u_i^{\pm}(x, y)$, where $i = 0$, 1, 2, and 3, is a spatial distribution.

$$u_2^+(x,y) = \frac{\exp[ik_0 f(1+\bar{c}^2)]}{i\lambda_0 f} \mathcal{F}[u_1^+(x,y)\mathcal{P}_1(x,y)] \Bigg|_{\substack{k_x = \frac{k_0 x}{f} \\ k_y = \frac{k_0 y}{f}}}, \tag{29}$$

$$U_2^+(k_x, k_y) = \mathcal{F}[u_2^+(x,y)]. \tag{30}$$

Combining Eqs. (29) and (30) gives

$$\begin{aligned} &U_2^+(k_x, k_y) \\ &= \mathcal{F}\left\{ \frac{\exp[ik_0 f(1+\bar{c}^2)]}{i\lambda_0 f} \mathcal{F}[u_1^+(x,y)\mathcal{P}_1(x,y)] \Bigg|_{\substack{k_x = \frac{k_0 x}{f} \\ k_y = \frac{k_0 y}{f}}} \right\} \\ &= i\lambda_0 f \exp[ik_0 f(1+\bar{c}^2)] u_1^+\left[-\left(\frac{f}{k_0}\right)k_x, -\left(\frac{f}{k_0}\right)k_y \right] \\ &\quad \times \mathcal{P}_1\left[-\left(\frac{f}{k_0}\right)k_x, -\left(\frac{f}{k_0}\right)k_y \right]. \end{aligned} \tag{31}$$

Suppose that $k_x^2 + k_y^2 \ll k_0$. Then, the following approximation is obtained:

$$k_z = (k_0^2 - k_x^2 - k_y^2) \cong k_0 - \frac{1}{2}\left(\frac{k_x^2 + k_y^2}{k_0}\right). \tag{32}$$

Then $U_3^+(k_x, k_y)$ is written as

$$U_3^+(k_x, k_y) = U_2^+(k_x, k_y) \exp[ik_0 z] \exp\left[-i\left(\frac{k_x^2 + k_y^2}{2k_0}\right) z\right], \tag{33}$$

$$z = |z_3 - z_2|, \tag{34}$$

where z_2 and z_3 are values along the Z axis at the front focal and the specimen planes, respectively.

$U_3^-(k_x, k_y)$ is the angular spectrum of the reflection wave at the interface between the specimen and the coupling medium. Therefore, it is written as

$$U_3^-(k_x, k_y) = U_3^+(k_x, k_y)\mathcal{R}\left[-i\left(\frac{k_x}{k_0}, \frac{k_y}{k_0}\right) z\right], \tag{35}$$

where $\mathcal{R}$ is the reflectance function [see Eq. (17)].

The angular spectrum of the reflection wave at the front focal plane is written as

$$U_2^-(k_x, k_y) = U_3^-(k_x, k_y) \exp[ik_0 z] \exp\left[-i\left(\frac{k_x^2 + k_y^2}{2k_0}\right) z\right]. \tag{36}$$

From Eqs. (33), (35), and (36), we obtain

$$\begin{aligned} U_2^-(k_x, k_y) &= i\lambda_0 f \exp[ik_0 f(1 + \overline{c}^2)] u_1^+\left[-\left(\frac{f}{k_0}\right) k_x, -\left(\frac{f}{k_0}\right) k_y\right] \\ &\quad \times \mathcal{P}_1\left[-\left(\frac{f}{k_0}\right) k_x, -\left(\frac{f}{k_0}\right) k_y\right] \exp(i2k_0 z) \\ &\quad \times \exp\left[-i\left(\frac{k_x^2 + k_y^2}{k_0}\right) z\right] \mathcal{R}\left(\frac{k_x}{k_0}, \frac{k_y}{k_0}\right). \end{aligned} \tag{37}$$

From Eq. (22), we obtain

$$u_2^-(x, y) = \mathscr{F}^{-1}[U_2^-(k_x, k_y)]. \tag{38}$$

Similarly to obtaining Eq. (26), we obtain

$$\begin{aligned} u_1^-(x, y) = \frac{\exp[ik_0 f(1 + \overline{c}^2)]}{i\lambda_0 f} &\times \iint_{-\infty}^{\infty} u_2^-(x_2, y_2) \mathcal{P}_2(x_1 + x_2, y_1 + y_2) \\ &\times \exp\left[-i\frac{2\pi}{\lambda_0 f}(x_1 + x_2, y_1 + y_2)\right] dx_2\, dy_2, \end{aligned} \tag{39}$$

where $\mathcal{P}_2$ is the pupil function from the coupling medium to the lens.

We use the same approximation as in Eq. (28):

$$\mathcal{P}_2(x_1 + x_2, y_1 + y_2) \cong \mathcal{P}_2(x_1, y_1). \tag{40}$$

Then, Eq. (39) is written as

$$u_1^-(x, y) = \frac{\exp[ik_0 f(1 + \overline{c}^2)]}{i\lambda_0 f} \mathcal{P}_1(x, y).\mathscr{F}[u_2^-(x, y)] \Bigg|_{\substack{k_x = \frac{k_0 x}{f} \\ k_y = \frac{k_0 y}{f}}}, \tag{41}$$

$$u_1^-(x, y) = \frac{\exp[ik_0 f(1 + \overline{c}^2)]}{i\lambda_0 f} \mathcal{P}_2(x, y) U_2^-\left(\frac{k_0}{f} x, \frac{k_0}{f} y\right), \tag{42}$$

$$\begin{aligned} u_1^-(x, y) = &- \exp\{i2k_0[z + f(1 + \overline{c}^2)]\} u_1^+(-x, -y) \mathcal{P}_1(-x, -y) \\ &\times \mathcal{P}_2(x, y) \exp\left[-i\left(\frac{k_0}{f^2}\right) z(x^2 + y^2)\right] \mathcal{R}\left(\frac{x}{f}, \frac{y}{f}\right). \end{aligned} \tag{43}$$

The variation of transducer voltage, as the position of the focal point varies is controlled by the movement of the acoustic lens along the z axis and is expressed as

$$V(z) = \iint_{-\infty}^{\infty} u_0^+(x_1, y_1) u_0^-(x_1, y_1)\, dx_1\, dy_1. \tag{44}$$

The wave number k_l in the buffer rod is expressed as

$$k_z' = (k_l^2 - k_x^2 - k_y^2)^{1/2}, \tag{45}$$

$$u_1^+(x_1, y_1) = u_0^+(x_1, y_1)^*.\mathscr{F}[\exp(ik_z' d)], \tag{46}$$

$$u_0^-(x_1, y_1) = u_1^-(x_1, y_1)^*.\mathscr{F}^{-1}[\exp(ik_z' d)], \tag{47}$$

$$d = |z_1 - z_0|, \tag{48}$$

$$V(z) = \iint_{-\infty}^{\infty} u_1^-(\xi, \eta) \left\{ \iint_{-\infty}^{\infty} u_0^+(x_1, y_1) \times \mathscr{F}^{-1}[\exp(ik_z' d)] \Big|_{\substack{x = x_1 - \xi \\ y = y_2 - \eta}} dx_1\, dx_2 \right\} d\xi\, d\eta. \tag{49}$$

Equation (45) is an even function. Therefore, the following relationship is obtained:

$$\mathscr{F}^{-1}[\exp(ik_z' d)] \Big|_{\substack{x = x_1 - \xi \\ y = y_1 - \eta}} = \mathscr{F}^{-1}[\exp(ik_z' d)] \Big|_{\substack{x = \xi - x_1 \\ y = \eta - y_1}}. \tag{50}$$

Using Eq. (50), Eq. (49) is rewritten as

$$V(z) = \iint_{-\infty}^{\infty} u_1^-(\xi, \eta) \left\{ u_0^+(\xi, \eta)^*.\mathscr{F}[\exp(ik_z' d)] \Big|_{\substack{x = \xi \\ y = \eta}} \right\} d\xi\, d\eta, \tag{51}$$

$$V(z) = \iint_{-\infty}^{\infty} u_1^-(\xi, \eta) u_1^+(\xi, \eta)\, d\xi\, d\eta. \tag{52}$$

Then, finally, we obtain

$$V(z) = -\exp\{i2k_0[z + f(1+\overline{c}^2)]\} \iint_{-\infty}^{\infty} u_1^+(-x,-y)u_1^+(x,y) \times \mathcal{P}_1(-x,-y)\mathcal{P}_2(x,y)\mathcal{R}\left(\frac{x}{f},\frac{y}{f}\right)\exp\left[-i\left(\frac{k_0}{f^2}\right) \times z(x^2+y^2)\right]dx\,dy. \quad (53)$$

Omitting the constant $\exp\{i2k_0[z + f(1+\overline{c}^2)]\}$, Eq. (53) is rewritten as

$$V(z) = \iint_{-\infty}^{\infty} u_1^+(-x,-y)u_1^+(x,y)\mathcal{P}_1(-x,-y)\mathcal{P}_2(x,y) \times \mathcal{R}\left(\frac{x}{f},\frac{y}{f}\right)\exp\left[-i\left(\frac{k_0}{f^2}\right)z(x^2+y^2)\right]dx\,dy. \quad (54)$$

Using $r-\theta$ coordination, the function of the voltage of the transducer, in accordance with the position of the focal point controlled by the movement of the acoustic lens along the Z axis, is rewritten as

$$V(z) = 2\pi \iint_{-\infty}^{\infty} u_1^+(r)^2\mathcal{P}_1(r)\mathcal{P}_2(r)\mathcal{R}\left(\frac{r}{f}\right) \times \exp\left[-i\left(\frac{k_0}{f^2}\right)zr^2\right]dr. \quad (55)$$

Equation (55) shows that the acoustic field of the back focal plane, the lens, the reflectance function, and the defocal distance are the main factors in the output voltage. Although the shape of the amplitude distribution of the acoustic field at the back focal plane is determined by the size of transducer and the frequency of the acoustic wave, generally speaking, its shape forms that of a Gaussian distribution (30). The pupil function is substantially constant. Therefore, the reflectance function affects the $V(z)$, which depends on the elastic properties of the material. Furthermore, the $V(z)$ changes in accordance with the defocusing distance.

Figure 19 shows the $V(z)$ curve for fused quartz. An acoustic lens that has an aperture angle of 120° and a working distance of 310 μm was used at an operating frequency of 400 MHz. The specimen was located in the water tank. The temperature of 22.3 °C in the coupling medium (distilled water) was measured by a thermocouple. The temperature was substantially stabilized (change less than ±0.1 °C). The movement of the acoustic lens along the Z axis was monitored by a laser-based measuring instrument. The transducer output voltage was periodic and had axial motion as the acoustic lens advanced from the focal plane toward the specimen. The contrast changed in accordance with the period.

Phase Change. Figure 20a,b show the amplitude and phase changes of the reflectance function due to incident angles of the acoustic waves from water to fused quartz, respectively. Figure 20a shows the relationship between the incident angle and the amplitude of the ultrasonic beam emitted from an acoustic lens operating at a frequency of 400 MHz onto the specimen (fused quartz) via the coupling medium (water). When the incident angle is close to the critical angle of the longitudinal wave (14.58°), the amplitude of the reflectance function suddenly becomes strong and becomes the maximum value at the critical angle. After passing the critical angle of the longitudinal wave, the amplitude decreases up to about 0.8 when the incident angle increases. Then, the incident angle is close to the critical angle of the shear wave (23.49°); the amplitude is again suddenly strong and becomes the maximum value at the critical angle. After passing the critical angle of the shear wave, the amplitude is constant.

Figure 20b shows the relationship between the incident angle and the phase of the ultrasonic beam emitted from an acoustic lens operating at a frequency of 400 MHz to the specimen (fused quartz) via the coupling medium (water). The phase changes a little at the critical angle of the longitudinal wave but changes significantly in the neighborhood of the Rayleigh critical angle. This significant phase change is the key factor in contrast change. Using a ray-tracing technique, this mechanism is explained as follows.

The period of this variation results from interference between the two components. Figure 21 shows that one component is spectrally reflected at normal incidence; the second undergoes a lateral shift on incidence and reradiates at the critical phase-matching angle

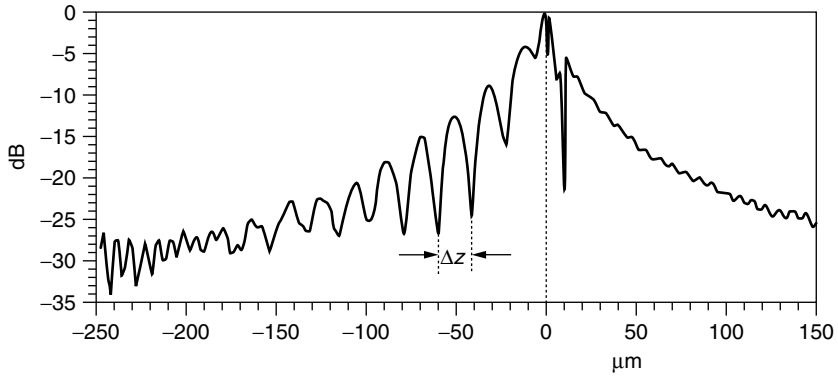


Figure 19. $V(z)$ curve for fused quartz; Specimen: fused quartz; coupling medium: distilled water; temperature of the coupling medium: 22.3 °C (change less than ±0.1 °C). Parameters of the acoustic lens are as follows: frequency: 400 MHz; aperture angle: 120°; and working distance: 310 μm.

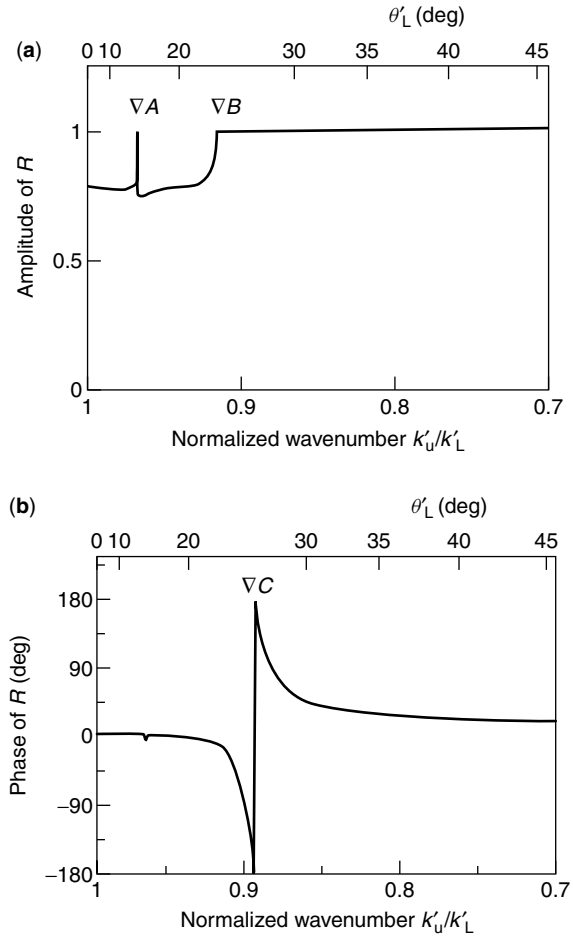


Figure 20. Reflectance function (**a**) Amplitude. (**b**) Phase. Specimen: fused quartz; coupling medium: distilled water; temperature of the coupling medium: 22.3 °C (change less than ±0.1 °C). Parameters of the acoustic lens are as follows: frequency: 400 MHz, aperture angle: 120°; and working distance: 310 μm.

for the SAW (also referred to as leaky Rayleigh waves).

When the acoustic wave is focused onto the surface of the specimen, the phases of acoustic waves that travel path (I) and path (II) are the same. Let this phase be denoted as Φ_f. When the acoustic lens is defocused toward the specimen at distance z, the phase changes of the waves that travel paths (I) and (II) are expressed as

$$\Phi_{(I)} = \Phi_f - \left(\frac{2\overline{OC}}{\lambda_W}\right)2\pi = \Phi_f - \frac{4\pi z}{\lambda_W}, \tag{56}$$

$$\begin{aligned}\Phi_{(II)} &= \Phi_f - \left(\frac{2\overline{AC}}{\lambda_W}\right)2\pi + \left(\frac{2\overline{AB}}{\lambda_R}\right)2\pi + \pi \\ &= \Phi_f - \frac{4\pi z}{\lambda_W \cos\theta_R} + \frac{4\pi \tan\theta_R}{\lambda_R} + \pi,\end{aligned} \tag{57}$$

where λ_W is the wavelength of the coupling medium (water), λ_R is the wavelength, and θ_R is the Rayleigh critical angle.

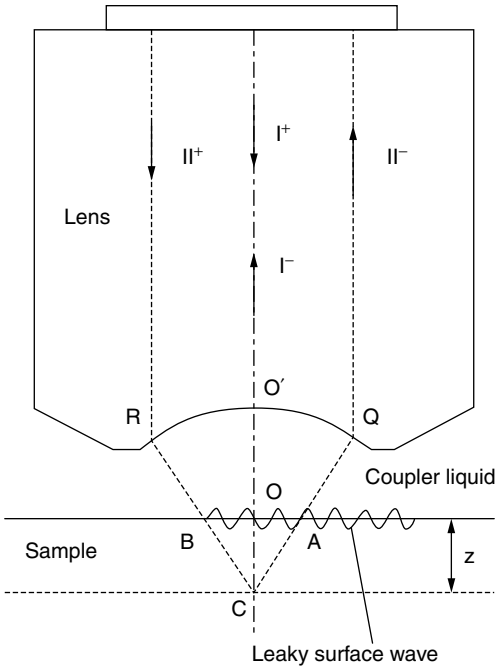


Figure 21. Cross-sectional geometry of spherical acoustic lens, explaining the mechanism of the $V(z)$ curves.

The phase difference is calculated by

$$\Delta\Phi = \Phi_{(II)} - \Phi_{(I)} = 4\pi z\left(\frac{1 - \dfrac{1}{\cos\theta_R}}{\lambda_W + \dfrac{\tan\theta_R}{\lambda_R}}\right) + \pi. \tag{58}$$

By Snell's law, we obtain the following equation:

$$\frac{V_W}{V_R} = \frac{\lambda_W}{\lambda_R} = \sin\theta_R, \tag{59}$$

where V_R is the SAW (e.g., Rayleigh wave) velocity on the surface of the specimen whose penetration depth is about one wavelength:

$$\lambda_R = \frac{\lambda_W}{\sin\theta_R} \tag{60}$$

Then, putting Eq. (59) into Eq. (58), we obtain the quantitative contrast factor as

$$\Delta\Phi = 4\pi z\left(\frac{1 - \cos\theta_R}{\lambda_W}\right) + \pi. \tag{61}$$

Theory of the SAW Velocity Measurement. Using the phase difference of two waves, the velocity of a surface acoustic wave (SAW) such as leaky Rayleigh waves, which propagate within a small area of a specimen, can be obtained (4,5).

When $(\Phi_{(II)} - \Phi_{(I)}) = (2n - 1)\pi$, where n is a natural number, the $V(z)$ curve is at a local minimum. Therefore, the period of the $V(z)$ is obtained as

$$\Delta z = \frac{\lambda_W}{2(1 - \cos\theta_R)}. \tag{62}$$

By rewriting Eq. (60), we obtain

$$V_R = \frac{V_W}{\sin\theta_R}. \tag{63}$$

λ_W is expressed as

$$\lambda_W = \frac{V_W}{f}, \tag{64}$$

where λ is the frequency of the acoustic lens.

Therefore, finally, from Eqs. (62–64) finally we obtain the following equation to calculate the SAW velocity:

$$V_R = \frac{V_W}{\sqrt{1 - \left(\frac{1}{2} \cdot \frac{V_W}{\Delta z \cdot f}\right)^2}}. \tag{65}$$

Many applications using this result have been reported for biological materials (94), thin films (95–101), coated materials (102–106), ceramics (107,108), and other materials (109–113).

Optimizing Measurement Precision. Assuming that $1 \gg \frac{V_W}{2f\Delta z}$, Eq. (65) is expressed approximately as

$$V_R \simeq \sqrt{V_W f \Delta z}. \tag{66}$$

To enhance the precision of the SAW velocity measurement by the $V(z)$ curve technique, both sides of Eq. (66) are differentiated after taking the logarithm of both sides. Then, the following equation is obtained:

$$\frac{dV_R}{V_R} = \frac{1}{2}\frac{dV_W}{V_W} + \frac{1}{2}\frac{df}{f} + \frac{1}{2}\frac{d\Delta z}{\Delta z} \tag{67}$$

Equation (67) shows that the errors in measuring the SAW velocity are the sum of the errors in the values of the velocity of the coupling medium, the frequency of the acoustic wave, and the distance of the period. Therefore, to minimize the measurement error, it is necessary to maintain a constant temperature for the coupling medium to stabilize the frequency of the acoustic wave and measure the movement of the acoustic lens along the Z axis accurately.

Many refinements in techniques for precision measurement have been reported (114–116).

Scattering. The scattering of acoustic waves can be divided into two categories: scattering caused by the particular condition of the specimen surface and scattering caused by nonhomogeneous particles inside the specimen.

The former covers discontinuities including those such as surface roughness. Although minor roughness (generally $\lambda/5$ or lower) is acceptable (117), large degrees of roughness affect the image directly (see Fig. 22). The image that is formed gives the impression that it is being viewed through frosted glass. For scanning acoustic microscopy, it is necessary either to consider the roughness in relation to the size of the area under observation or lower the frequency to minimize the influence of any roughness; some techniques for removing the effects of surface roughness have been reported (14,118,119).

The previous also applies to wave scattering caused by nonhomogeneous particles and must be kept in mind, especially for subsurface imaging. A specimen that contains a huge amount of scattering bodies is difficult to observe under high resolution and necessitates lowering the observation frequency.

Discontinuities. When a specimen includes an elastic discontinuity, such as an edge, a step, a crack, or a joint interface, an acoustic image of the elastically discontinuous and peripheral portions visualized by the scanning acoustic microscope shows unique contrast, for example, fringes or black stripes. As a model shown in Fig. 23, this type of contrast appears as an interference effect of surface acoustic waves incident on and reflected from elastic discontinuities (120–125).

Figure 24a is an image obtained by the scanning electron microscope (SEM) of a Cu/Si_3N_4 interface after the Si_3N_4 portion was given a micro-Vickers indentation. Figure 24b is the magnified SEM image of the portion that has a vertical radial crack which reached the Cu/Si_3N_4 interface. The SEM at higher magnification did not reveal any delamination. The assumption was that an opened

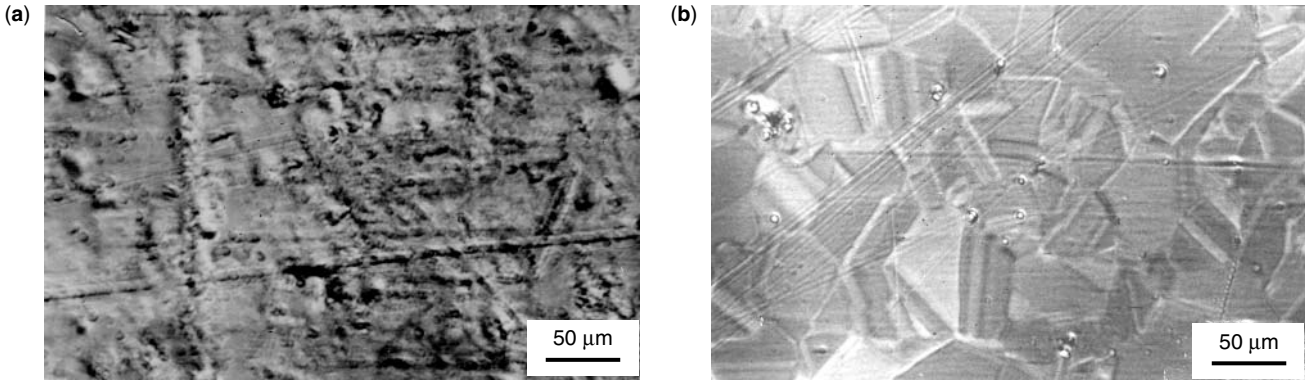


Figure 22. Surface roughness. (**a**) Incident acoustic waves are scattered by surface roughness of the specimen. Therefore, the interior portions are difficult to observe. (**b**) SAM image of the specimen that has a polished surface. Specimen: stainless steel (SUS 304); frequency: 800 MHz.

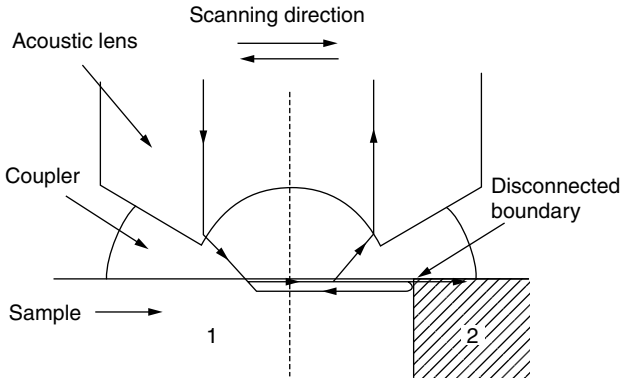


Figure 23. Model of Rayleigh wave excitation at a boundary.

delamination on the surface, soon closed due to change of time and relief of stress when the indentation was completed.

Figures 25a,b are SAM images. A frequency of 400 MHz was chosen for the visualization. The scanning width was 1.0 mm. In Figs. 25a,b, the acoustic lens was defocused by 40 and 100 μm, respectively. Without further preparation, the SAM gave the image of the same specimen at similar magnifications and revealed the delamination (see Fig. 25a). Moreover, the delamination was observed with enhanced contrast (Fig. 25b). This can be explained by the SAW scattering mechanism, that is, when an acoustic lens that has a high numerical aperture is defocused, an ultrasonic beam is incident at a certain range of angles onto the specimen. In this experiment, the aperture angle of the acoustic lens was 120°. At a certain incident angle, the SAW is excited, travels on the surface of the specimen, and is reflected back from the interface (see Fig. 23). These reflected SAWs give the enhanced contrast of discontinuities in a specimen in acoustic images.

Resolution

Two types of resolution must be considered for the SAM. One is a lateral resolution (notated as "Δr"), and the other is a vertical resolution (notated as "$\Delta\rho$"). These are

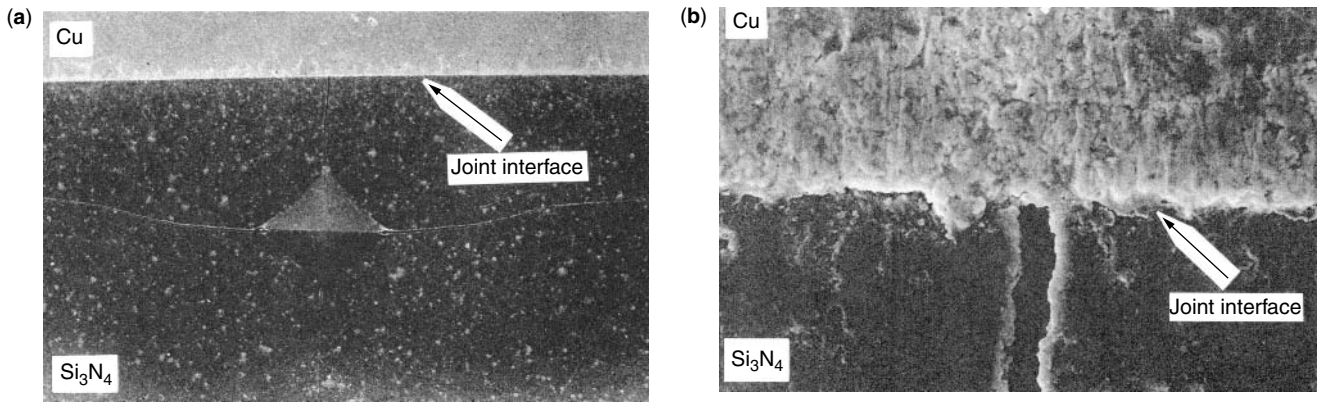


Figure 24. Scanning electron microscope (SEM) image. (**a**) Schematic view. (**b**) Magnified image. The micro-Vickers indentation was made in the Si_3N_4 portion to create the delamination at the interface between steel and Si_3N_4.

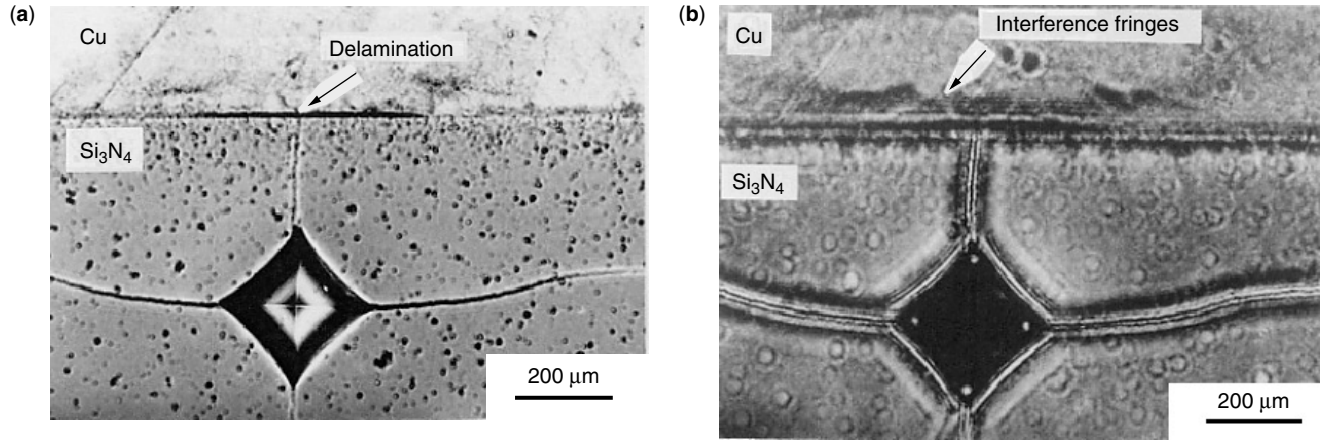


Figure 25. SAM images. (**a**) $Z = -40$ μm; (**b**) $Z = -100$ μm; frequency: 400 MHz.

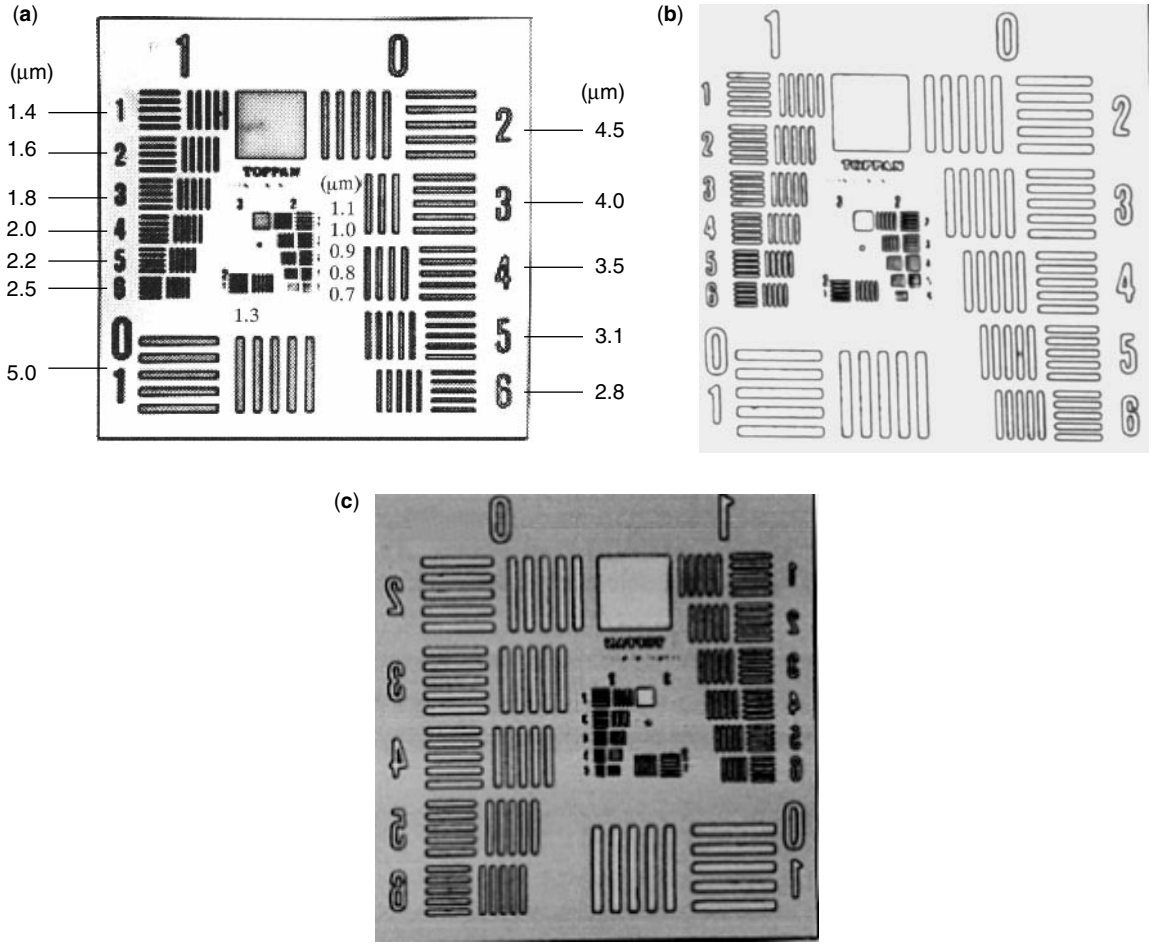


Figure 26. Resolution. (**a**) Optical image of a standard specimen that has patterns for measuring resolution for the scanning acoustic microscope in the C-scan mode using a tone-burst wave; (**b**) frequency: 1.0 GHz; (**c**) frequency: 400 MHz.

expressed as

$$\Delta r = F\lambda = F\left(\frac{v_w}{f}\right), \tag{68}$$

$$\begin{aligned}\Delta\rho = 2F^2\lambda &= (2F^2)\left(\frac{v_w}{f}\right),\\ &= 2\left(\frac{f_o}{D}\right)^2\left(\frac{v_w}{f}\right),\\ &= 2\left(\frac{1}{2\tan\theta}\right)^2\left(\frac{v_w}{f}\right),\\ &= \frac{1}{2(\tan\theta)^2}\left(\frac{v_w}{f}\right),\end{aligned} \tag{69}$$

where F is a constant related to lens geometry, λ is the wavelength in the coupling medium (water), f is the frequency of the wave generated by the transducer, v_w is the longitudinal wave velocity in the coupling medium, f_o is the focal distance of the lens, D is the diameter of the lens aperture, and θ is one-half of the aperture angle of the lens. Figure 26 shows surface resolutions at different frequencies. The resolutions in the images tend to be higher than the calculated resolutions.

CONCLUSION

The features for the SAM are summarized as follows:

1. capability of observing both the surface and subsurface of a specimen
2. high resolution
3. high contrast
4. capability of quantitative data acquisition

Although the SAM has unique features, it (especially the tone-burst-wave mode) has not been used in industry as much as conventional microscopes. It is hoped that dedicated work performed by acoustic microscopists will clear the obstacles to acceptance and allow the SAM to mature and prosper fully and that this article, which describes the fundamentals of scanning acoustic microscopy, is useful to a person to become one of them.

ABBREVIATIONS AND ACRONYMS

AARC	acoustic antireflective coating
A/D	analog-to-digital signal
C-FRP	carbon-fiber-reinforced plastic
D/A	digital-to-analog signal
DMB	double balanced mixer
FRP	fiber reinforced plastic
IC	integrated circutis
$LiNbO_3$	lithium niobate
mV	milli volt
NDE	nondestructive evaluation
PVDF	polyvinyldifluoride
PZT	lead zirconate titanate ceramic
RF	radio frequency
SAM	scanning acoustic microscope
SAW	surface acoustic wave
SEM	scanning electron microscope
SPDT	single pole double throw
V	volt
W.D.	working distance

BIBLIOGRAPHY

1. S. Y. Sokolov, *Academia Nauk SSSR Doklady*, **64**(3), 333–335 (1949).
2. R. A. Lemons and C. F. Quate, *Appl. Phys. Lett.* **24**, 163–165 (1974).
3. A. Atalar, C. F. Quate, and H. K. Wickramasinge, *Appl. Phys. Lett.* **31**, 791 (1977).
4. R. D. Weglein, *Appl. Phys. Lett.* **34**, 179–181 (1979).
5. W. Parmon and H. L. Bertoni, *Electron. Lett.* **15**, 684–686 (1979).
6. A. Atalar, *J. Appl. Phys.* **49**, 5,130–5,139 (1978).
7. B. Hadimioglu and C. F. Quate, *Appl. Phys. Lett.* **43**, 1,006–1,007 (1983).
8. J. S. Foster and D. Rugar, *Appl. Phys. Lett.* **42**, 869–871 (1983).
9. M. S. Muha, A. A. Moulthrop, G. C. Kozlowski, and B. Hadimioglu, *Appl. Phys. Lett.* **56**, 1,019–1,021 (1990).
10. N. Chubachi, J. Kushibiki, T. Sannomia, and Y. Iyama, IEEE, *Proc. Ultrasonic Symp.*, 1979, pp. 415–418.
11. D. A. Davids, P. Y. We, and D. Chizhik, *Appl. Phys. Lett.* **54**(17), 1,639–1,641 (1989).
12. A. Atalar, H. Koymen, and L. Degertekin, *Proc. IEEE Ultrasonics Symp.*, 1990, pp. 359–362.
13. B. T. Khuri-Yakub, C. Cinbis, and P. A. Reinholdtsen, *Proc. IEEE Ultrasonics Symp.*, 1989, pp. 805–807.
14. C. Miyasaka, B. R. Tittmann, and M. Ohno, *Res. Nondestructive Evaluation* **11**, 97–116 (1999).
15. C. W. W. Daft, J. M. R. Weaver, and G. A. D. Briggs, *J. Micros.* **139**(3), RP3–RP4 (1985).
16. B. T. Khuri-Yakub and C. -H. Chou, *Proc. IEEE Ultrasonics Symp.*, 1986, pp. 741–744.
17. A. Atalar, I. Ishikawa, Y. Ogura, and K. Tomita, *Proc. IEEE Ultrasonics Symp.*, 1993, pp. 613–616.
18. S. Ujihashi, K. Skanoue, T. Adachi, and H. Matsumoto, *Proc. Mech. Eng. Fourth Int. Conf. FRC'90*, 1990, pp. 157–162.
19. H. Matsumoto, T. Adachi, and S. Ujihashi, *Proc. Oji International Seminar on Dynamic Fracture*, Chou Technical Drawing Co., Ltd., 1990, pp. 174–182.
20. K. Kasano, T. S. Bea, and C. Miyasaka, *Proc., Japan-U.S. CCM-VII*, 1995, pp. 601–608.
21. A. M. Howard, *IEE Colloquium on 'NDT Evaluation of Electronic Components and Assemblies'*, 1990, p. 3.
22. L. Revay, G. Lindblad, and L. Lind, *CERT'90; Component Engineering, Reliability and Test Conference (Electron Components Inst.)*, 1990, pp. 115–122.
23. R. Bauer, M. Luniak, and L. Rebenklau, *Proc. Int. Symp. Microelectronics (IMAPS)*, 1997, pp. 659–664.
24. T. Adams, *Electron. Bull.* **16**(6), 51–55 (1998).
25. J. C. Bernier and L. Teems, *Proc. Int. Symp. Testing Failure Anal. (ISFTA)*, 1998, pp. 393–398.
26. B. Smith, *Am. Lab.* **31**(22), 18–21 (1999).
27. J. Kushibiki, K. Horii, and N. Chubachi, *Electron. Lett.* **19**, 404–405 (1983).
28. J. Kushibiki, K. Horii, and N. Chubachi, *Electron. Lett.* **19**, 404–405 (1983).
29. K. Liang, G. S. Kino, and B. T. Khuri-Yakub, *IEEE Trans.* **SU-32**, 213–224 (1985).
30. T. Endo, Y. Sasaki, T. Yamagishi, and M. Sakai, *Jpn. Appl. Phys.* **31**, 160–162 (1992).
31. A. Kulik, G. Gremaud, and S. Sathish, in *Acoustic Imaging*, vol. 17, Plenum Press, New York, 1989, pp. 71–78.
32. Y. Tsukahara, E. Takeuchi, E. Hayashi, and Y. Tani, *Proc. IEEE 1984 Ultrason. Symp.*, 1984, pp. 992–996.
33. M. Nikoonahad, *Electron. Lett.* **32**(10), 489–490 (1987).
34. S. W. Meeks et al., *Appl. Phys. Lett.* **55**(18), 1,835–1,837 (1989).
35. J. Attal, in E. A. Ash, ed., *Scanned Image Microscopy*, Academic Press, London, 1980, pp. 97–118.
36. H. K. Wickramasinghe and C. R. Petts, in E. A. Ash, ed., *Scanned Image Microscopy*, Academic Press, London, 1980, pp. 57–100.
37. K. Karaki and M. Sakai, K. Toda, ed., *Toyohashi Int. Conf. Ultrasonic Technol.*, Toyohashi, Japan, MYU Research, Tokyo, 1987, pp. 25–30.
38. K. Yamanaka, Y. Nagata, and T. Koda, *Ultrasonics Int.* **89**, 744–749 (1989).
39. J. D. Fox, B. T. Khuri-Yakub, and G. S. Kino, *Proc. IEEE 1983 Ultrasonics Symp.*, 1983, pp. 581–592.
40. T. Yano, M. Tone, and A. Fukumoto, *IEEE Trans. Ultrasonics Ferroelectronic Frequency Control* **34**(2), 222–236 (1987).
41. M. I. Haller and B. T. Khuri-Yakub, *Proc. IEEE 1992 Ultrasonics Symp.*, pp. 937–939.
42. D. Reilly and G. Hayward, *Proc. IEEE 1991 Ultrasonics Symp.*, pp. 763–766.
43. D. W. Schindel, D. A. Hutchins, L. Zou, and M. Sayer, *IEEE Trans. Ultrasonics Ferroelectronics Frequency Control* **42**, 42–51 (1995).
44. I. Ladabaum, B. T. Khuri-Yakub, and D. Spoliansky, *Appl. Phys. Lett.* **68**, 7–9 (1996).
45. M. C. Bhardwaj, I. Neeson, M. E. Langron, and L. Vandervalk, *Proc. 24th Conf. An Int. Conf. Eng. Ceram. Struct., American Ceramic, Materials, and Structures: A*, 2000, pp. 163–172.
46. C. Miyasaka and B. R. Tittmann, *Proc. IEEE 1998 Ultrasonics Symp.*, pp. 1,265–1,268.
47. I. Ihara, C. -K. Jen, and D. R. Franca, *Proc. IEEE 1998 Ultrasonics Symp.*, pp. 803–807.
48. R. Kompfner and R. A. Lemons, *Appl. Phys. Lett.* **28**(6), 295 (1976).

49. C. E. Yeack, M. Chodorow, and C. C. Cuttler, *Appl. Phys.* **51**(9), 4,631–4,636 (1977).
50. L. Germain and J. D. N. Cheek, *J. Acoust. Soc. Am.* **83**(3), 942–949 (1988).
51. L. Germain, R. Jacques, and J. D. N. Cheek, *J. Acoust. Soc. Am.* **86**(4), 1,560–1,565 (1989).
52. M. R. T. Tan, H. L. Ransom Jr., C. C. Cutler, and M. Chodorow, *Appl. Phys.* **57**, 4,931–4,935 (1985).
53. H. K. Wickramasinghe and C. E. Yeack, *Appl. Phys.* **48**(12), 4,951–4,954 (1977).
54. D. Rugar, *Appl. Phys.* **56**, 1,338–1,346 (1984).
55. K. Kraki, T. Saito, K. Matsumoto, and Y. Okuda, *Physica B* **165/166**, 131–132 (1990).
56. K. Kraki, T. Saito, K. Matsumoto, and Y. Okuda, *Appl. Phys. Lett.* **59**(8), 908–910 (1991).
57. V. A. Burov, I. E. Gurinovich, O. V. Rudenko, and E. Y. Tagunov, in *Acoustic Image*, vol. 22, Plenum, NY, 1996, pp. 125–130.
58. X. Gong and D. Zhang, in *Acoustic Imaging*, vol. 23, Plenum, NY, 1997, 601–605.
59. J. Synnevag and S. Holm, *Proc. IEEE 1998 Ultrasonics Symp.*, 1998, pp. 1,885–1,888.
60. R. C. Bray and C. F. Quate, *Thin Solid Films* **74**, 295–302 (1980).
61. C. C. Lee, C. S. Tsai, and X. Cheng, *IEEE Trans. Sonics Ultrasonics* **SU-32**(2), 248–258 (1985).
62. R. C. Addison, M. C. Somekh, J. M. Rowe, and G. A. D. Briggs, L. A. Ferrari, ed., *SPIE* **768**, 275–284 (1987).
63. R. D. Weglein, *IEEE Trans. Sonics Ultrasonics* **SU-30**, 40–42 (1983).
64. I. Ishikawa, H. Kanda, and K. Ktakura, *IEEE Trans. Sonics and Ultrasonics* **SU-32**(2), 325–331 (1985).
65. C. S. Tsai and C. C. Lee, *Proc. SPIE* **768**, 260–266 (1987).
66. J. Onuki, M. Koizumi, and I. Ishikawa, *Materials Trans. JIM* **37**(9), 1,492–1,496 (1996).
67. I. Ishikawa, H. Kanda, K. Ktakura, and T. Semba, *IEEE Trans. Ultrasonics Ferroelectrics Frequency Control* **UFCC-36**(6), 587–592 (1989).
68. R. L. Hollis, R. Hammer, and M. Y. Al-Jaroudi, *Appl. Phys.* **54**, 7,016–7,019 (1984).
69. M. Hoppe and J. Bereiter-Hahn, *IEEE Trans.* **SU-32**, 289–301 (1985).
70. J. Y. Duquesne et al., *Mat. Res. Soc. Symp. Proc.* **142**, 253–259 (1989).
71. A. F. Fagan, J. M. Bell, and G. A. D. Briggs, in A. C. Roulin-Moloney, ed., *Fractography and Failure Mechanisms of Polymers and Composites*, Elsevier Applied Science, London, 1989, pp. 213–230.
72. A. M. Sinton, G. A. D. Briggs, and Y. Tsukahara, in H. Shimizu, N. Chubachi, and J. Kushibiki, eds., *Acoustical Imaging*, 17th ed., Plenum, NY, 1989, pp. 87–95.
73. R. C. Addison, M. W. Kendig, and S. J. Jeanjaquet, in H. Shimizu, N. Chubachi, and J. Kushibiki, eds., *Acoustical Imaging*, 17th ed., Plenum, NY, 1989, pp. 143–152.
74. T. Adachi, M. Okazaki, S. Ujihashi, and H. Matsumoto, *JSME Int. J. Series A* **38**(3), 370–377 (1995).
75. H. Morita, T. Adachi, and H. Matsumoto, *J. Reinforced Plastics Composites* **16**(2), 131–143 (1997).
76. N. Takeda, C. Miyasaka, and K. Nakata, *Proc. 5th Int. Symp. Nondestructive Characterization Mater. Karuizawa*, 1991, pp. 813–824.
77. K. Yamanaka, *Appl. Phys.* **24**, 184–186 (1984).
78. A. Okada, C. Miyasaka, and T. Nomura, *JIM* **33**(1), 73–79 (1992).
79. B. Nongaillard et al., *NDT INT.* **19**(2), 77–82 (1986).
80. R. D. Weglein, *Acousts Imaging* **17**, 51–59 (1988).
81. H. Vetters, A. Meyyappan, A. Schlz, and P. Mayr, *Mater. Sci Eng.* **A122**, 9–14 (1989).
82. J. Hahn-Bereiter, C. H. Fox, and B. Thorell, *J. Cell Biol.* **82**, 767–779 (1979).
83. J. A. Hildebrand and D. Rugar, *J. Microsc.* **134**, 245–260 (1984).
84. N. Chubachi et al., *Acoust. Imaging* **16**, 1–9 (1987).
85. J. Bereiter-Hahn et al., *Acoust. Imaging* **17**, 27–38 (1988).
86. N. Akashi, J. Kushibiki, N. Chubachi, and F. Dunn, *Acoust. Imaging* **17**, 183–191 (1989).
87. J. Litniewski and J. Bereiter-Hahn, *J. Microsc.* **158**, 95–107 (1990).
88. J. Bereiter-Hahn and H. Luers, *Eur. J. Cell Biol.* **53**(Suppl. 31), 85 (1990).
89. H. Luers, K. Hillmann, J. Litniewski, and J. Bereiter-Hahn, *Cell Biophys.* **18**, 279–93 (1992).
90. P. A. N. Chandraratna et al., *Am. Heart J* **124**, 1,358–1,364 (1992).
91. G. A. D. Briggs, J. Wang, and R. Gundle, *J. Microsc.* **172**, 3–12 (1993).
92. J. Bereiter-Hahn and H. Luers, in N. Akkas, ed., *Mechanics of Actively Locomoting Cells*, ASI series 84, Springer, Heidelberg, New York, Berlin, 1994, pp. 181–230.
93. P. A. N. Chandraratna et al., *Acous. Imaging* **21**, 559–564 (1995).
94. T. Kundu, J. Bereiter-Hahn, and K. Hillmann, *Biophys. J.* **59**, 1,194–1,207 (1991).
95. J. Kushibiki, T. Ishikawa, and N. Chubachi, *Appl. Phys. Lett.* **57**(19), 1,967–1,969 (1990).
96. T. Endo, C. Abe, M. Sakai, and M. Ohono, *Proc. Ultrasonic Int. 1993 Conf.*, pp. 45–48.
97. D. Achenbach, J. D. Kim, and Y. C. Lee, in A. Briggs, ed., *Advances in Acoustic Microscopy*, 1st ed., Plenum, NY, 1995, pp. 153–208.
98. S. Parthasarathi, B. R. Tittmann, and R. J. Ianno, *Thin Solid Films* **300**, 42–52 (1997).
99. R. D. Weglein, *IEEE Trans. Sonics Ultrasonics* **SU-27**(2), 82–86 (1980).
100. J. Kushibiki and N. Chubachi, *Electron. Lett.* **23**(12), 652–4 (1987).
101. Y. Sasaki, T. Endo, T. Yamagishi, and M. Sakai, *IEEE Trans.* **UFFC-39**(5), 638–642 (1992).
102. R. D. Weglein, *IEEE Trans. Sonics Ultrasonics* **SU-32**(2), 225–234 (1985).
103. R. D. Weglein and A. K. Mal, *Surf. Coat. Technol.* **47**, 667–686 (1991).
104. J. Kushibiki, M. Miyashita, and N. Chubachi, *IEEE Photonics xTechnol. Lett.* **8**(11), 1,516–1,518 (1996).
105. J. Kushibiki, T. Kobayashi, H. Ishiji, and N. Chubachi, *Appl. Phys. Lett.* **61**(18), 2,164–2,166 (1992).
106. J. Kushibiki and M. Miyashita, *Jpn J. Appl. Phys.* **36**(7B), 959–961 (1997).
107. T. Narita, K. Miura, I. Ishikawa, and T. Ishikawa, *Jpn. Inst. Met.* 1,142–1,146 (in Japanese) (1990).
108. S. Tanaka and C. Miyasaka, *Residual Stress III*, 1991, pp. 278–283.

109. J. Kushibiki and N. Chubachi, *IEEE Trans. Sonics and Ultrasonics* **SU-32**(2), 189–212 (1985).
110. M. Obata, H. Shimada, and T. Mihara, *Exp. Mech.* **30**, 34–39 (1990).
111. Y. C. Lee, J. D. Kim, and D. Achenbach, *Ultrasonics* **32**(5), 359–365 (1995).
112. K. Yamanaka, *Electr. Lett.* **18**(14), 587–589 (1982).
113. S. M. Gracewski, R. C. Waag, and E. A. Schenk, *J. Acoust. Soc. Am.* **83**(6), 2,405–2,409 (1988).
114. Z. L. Li, *IEEE Trans.* **UFCC-40**(6), 680–686 (1993).
115. M. Okade and K. Kawashima, *Proc. QNDE* **14**, 1,883–1,889 (1994).
116. J. Kushibiki and M. Arakawa, *IEEE Trans.* **UFCC-45**(2), 421–430 (1998).
117. B. T. Khuri-Yakub, P. Reinholdtsen, and C. H. Chou, *IEEE 1985 Ultrasonics Symp.*, pp. 746–749.
118. P. Reinholdtsen and B. T. Khuri-Yakub, *Proc. IEEE 1986 Ultrasonics Symp.*, pp. 759–763.
119. M. Ohno, *Appl. Phys. Lett.* **55**, 832–823 (1989).
120. R. D. Weglein, *Electron. Lett.* **14**(20), 656–657 (1978).
121. C. Iett, M. G Somekh, and G. A. D. Briggs, *Proc. R. Soc. London* **A393**, 171–183 (1984).
122. S. Kojima, *Jpn. J. Appl. Phys.* **26**, 233–235 (1987).
123. K. Yamanaka and Y. Enomoto, *J. Appl. Phys.* **53**, 84–50 (1982).
124. T. Ghosh, K. I. Maslov, and T. Kundu, *Ultrasonics* **35**, 337–366 (1997).
125. M. Ohno, C. Miyasaka, and B. R. Tittmann, *Wave Motion* **33**, 309–320 (2001).

SCANNING ELECTROCHEMICAL MICROSCOPY

DAVID O. WIPF
Mississippi State University
Mississippi State, MS

INTRODUCTION

Scanning electrochemical microscopy (SECM) is one of a number of scanned probe microscopy (SPM) techniques invented after the demonstration of the scanning tunneling microscope (1). The use of an electrochemical process for image formation defines SECM. In most applications of the method, an ultramicroelectrode (UME) is used as the probe, and the probe signal is the faradic current that arises from electrolysis of solution species (2,3). In other applications, using an ion-selective electrode (ISE) as the probe provides a probe signal proportional to the logarithm of the activity of an ion in solution (e.g., pH). In SECM, the primary interaction between probe tip and sample is mediated by diffusion of solution species between the sample and the tip of the probe, this distinguishes SECM from other SPM methods that may use an electrochemically active probe. An electrochemically active probe permits a versatile range of experiments, whose essential aspect is chemical sensitivity or control of chemical processes that occur at a substrate surface (4–6).

Forming an image in an SPM technique requires reproducibly perturbing the probe signal by some aspect of the imaged substrate. There are two principal image-forming modes in SECM: feedback and generation/collection (GC). The feedback mode uses the faradaic current that flows from electrolysis of an intentionally added or naturally present mediator species (e.g., $Ru(NH_3)_6^{3+}$ or O_2) at an UME probe (4,7,8). Far from the substrate surface, the electrolytic current assumes a characteristic steady-state value, but moving the probe close to the surface disturbs the current. An electrochemical, chemical, or enzymatic reaction at the substrate can restore the mediator to its original oxidation state and produce a *positive*-feedback signal. The cartoon representation in Fig. 1a illustrates the process where an oxidized mediator undergoes reduction at the probe tip, diffuses to the substrate surface, and then is reoxidized by the substrate. The regeneration of the mediator in the tip–substrate gap causes the probe current to increase as the separation decreases and suggests the term feedback as a description. In contrast, *negative* feedback occurs when the substrate surface cannot regenerate the mediator and mediator diffusion is blocked. Depletion of mediator by tip electrolysis in the tip–substrate gap causes the probe signal to decrease as the gap width decreases (Fig. 1b). Imaging in the feedback mode provides topographic images of blocking or conducting surfaces. Images of many types of surfaces have been obtained in the feedback mode, including images of electrodes (9–11), polymer films (12–16), and immiscible liquid interfaces (17).

It is possible to manipulate the SECM imaging conditions to produce images that represent chemical and electrochemical activity. In the reaction-rate feedback mode (Fig. 1c), the overall mediator turnover rate at the substrate is limited by the rate of substrate–mediator electron transfer (ET) (18–20). Thus, the probe current signal contains kinetic and tip–substrate separation information. Images in the reaction-rate mode can be used to map variations in ET rate between the mediator at metallic electrodes (18,21) and enzymes in biological materials (22).

The generation/collection (GC) mode uses the probe to detect changes in the concentration of a chemical species at the surface of the imaged material (Fig. 1d). Ideally, the probe acts as a passive sensor to produce concentration maps of a particular chemical species near the substrate surface. GC mode imaging is described further by the type of sensing probe used. In amperometric GC imaging, the probe is an UME, that detects species by electrolysis. Amperometric GC, first reported as a method for mapping electrochemically active areas on electrodes (23), is used to make high-resolution chemical concentration maps of corroding metal surfaces (24–28), biological materials (29–33), and polymeric materials (15,34,35). In addition, measurements of ion fluxes through porous materials (36,37) such as skin (38–40) and dental material (37,41,42) are useful applications. In potentiometric GC imaging, the probe is an ISE, which has the advantage of increased sensitivity to nonelectroactive ions and improved selectivity for imaging a desired ion concentration. The literature contains descriptions of tip electrodes and experiments for measuring of many types of ions such as H^+ (i.e., pH) (43–45), K^+ (46), NH_4^+, and Zn^{2+} (45).

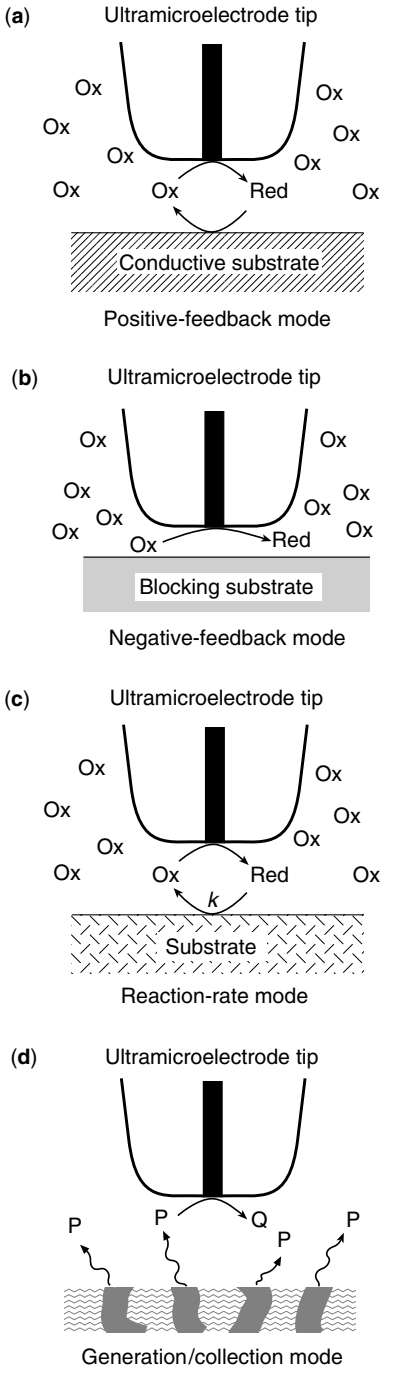


Figure 1. Schematic diagram of the image modes in SECM.

SIGNAL TRANSDUCTION

The probe in SECM is an electrode in an electrolyte solution, and the tip electrode response can be quantitatively described under most measurement conditions. The most well-understood response arises from a probe geometry where disk is embedded in an insulating plane. Other tip electrode geometries are also used, but the experimental data are generally treated semiquantitatively or empirically.

A description of the electrochemical behavior of the UME tip is required to understand the probe signal in the SECM feedback or amperometric GC mode. Voltammetry or amperometry is the basis of a large number of electrochemical techniques in which a potential is applied between an *auxiliary* and *working* electrode (47). Control of the potential is aided by a third, *reference*, electrode that maintains a constant reference voltage. In either voltammetry or amperometry, the signal is the faradaic current that flows as solution species are oxidized or reduced at the working electrode. A cyclic voltammogram (CV) illustrates this at an embedded-disk UME working electrode (Fig. 2, inset). The CV wave shown is a plot of the reduction current versus applied potential for a commonly used mediator ion, $Ru(NH_3)_6^{3+}$, at a 10-μm diameter tip electrode in unstirred solution. The reduction current reaches a limiting current value at potentials negative of the reduction potential, and the forward curve is retraced as the tip electrode potential is swept back to the starting potential. The current limit $i_{T,\infty}$ occurs when the rate of mediator diffusion to the electrode surface is at maximum velocity (i.e., when the mediator concentration is zero at the electrode surface) and, for a microdisk electrode embedded in an insulator, is given by the following equation (3):

$$i_{T,\infty} = 4\,nFDCa \tag{1}$$

where F is the faraday, C and D are the mediator concentration and diffusion coefficients, n is the number

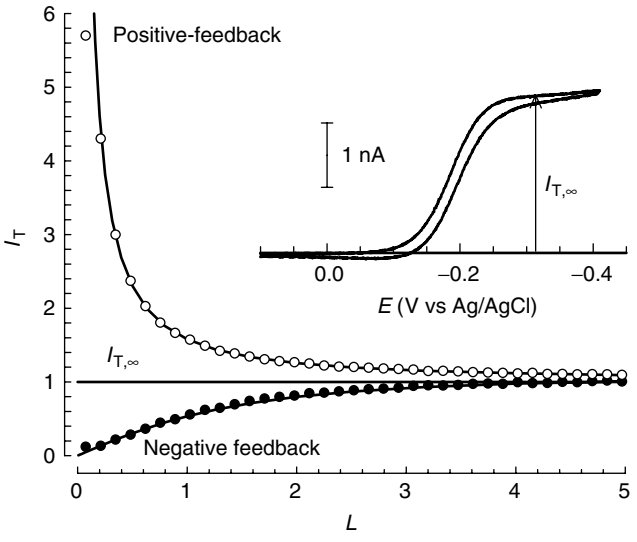


Figure 2. Theoretical (———) and experimental current vs. distance curves for positive (○○○) and negative (●●●) feedback using an embedded-disk tip electrode. Experimental conditions for positive feedback: 2 mM $Ru(NH_3)_6^{3+}$, pH 4.0 phosphate-citrate buffer, 10-μm diameter Pt tip, Pt substrate, tip potential (E_{tip}) is −0.3 V vs. Ag/AgCl reference, and substrate potential (E_{sub}) 0.0 V vs. Ag/AgCl. Negative feedback conditions are as in positive feedback except for use of a microscope slide substrate. Experimental current data are normalized by dividing by $i_{T,\infty}$, and distance scale is divided by the tip radius. (Inset) cyclic voltammogram for a 10-μm diameter Pt disk in 2 mM $Ru(NH_3)_6^{3+}$, pH 4.0 phosphate-citrate buffer at a scan rate of 100 mV/s.

of electrons transferred in the tip electrode reaction, and a is the disk radius. A characteristic of an UME is that the region of solution perturbed by diffusion is much larger than the electrode itself. This results in an efficient form of diffusional transport where the diffusion field assumes a nearly hemispherical zone around the electrode (3) and is the basis for the feedback mode of SECM imaging.

Feedback Mode

In the feedback mode, the tip electrode potential is adjusted so that the electrode current for a mediator is at the limiting current value. As the UME probe approaches the surface, the hemispherical diffusion field of the mediator is disturbed, and the imaging signal i_T changes from that observed at infinite distance (i.e., $i_{T,\infty}$). Two limiting cases, referred to as positive or negative feedback, are observed when the substrate either permits or blocks, respectively, ET to the mediator. Although no physical contact between probe and sample occurs in normal feedback imaging, a chemical interaction occurs because a small region of the substrate is subjected to a chemical environment different from the bulk. Imaging conditions are tuned by choosing a mediator that will produce a desirable interaction between the probe and substrate. However, a mediator must be chosen to avoid undesirable interactions with the sample, such as destruction of a sample by an oxidizing mediator.

The relationship between probe current and tip–surface separation can be quantitatively calculated for tip electrodes in the shapes of embedded disks (8), cones, and spherical sections (11,48–50). An approximate analytical equation of the current–distance relationship is available for an embedded-disk probe (49,51). For positive feedback,

$$I_T(L) = \frac{0.78377}{L} + 0.3315 \exp\left(\frac{-1.0672}{L}\right) + 0.68. \quad (2)$$

The probe current and distance are presented in normalized form: $I_T = i_T/i_{T,\infty}$ and $L = d/a$ is the normalized distance d between the probe tip electrode and substrate. For negative feedback,

$$I_T(L) = \frac{1}{0.292 + 1.515/L + 0.6553 \exp(-2.4035/L)}. \quad (3)$$

Comparing I–L curves between experimental and theoretical data for disk-shaped electrode probes generally produces excellent agreement, and curve fitting can be used to verify proper instrument operation. Figure 2 is a plot of a portion of the theoretical I–L curve compared to experimental data recorded at a 10-μm diameter microdisk probe. Note the increased distance sensitivity in positive feedback compared to negative feedback.

Probe geometry strongly influences the probe current in the feedback mode experiment (52). For accurate measurements, the probe generally has the shape of a truncated cone. Insulating material (e.g., glass or polymer) forms the shape of the cone, and the disk electrode is located in the center of the truncated end (Fig. 3). This shape allows a closer probe tip approach to the sample surface than would be otherwise possible with

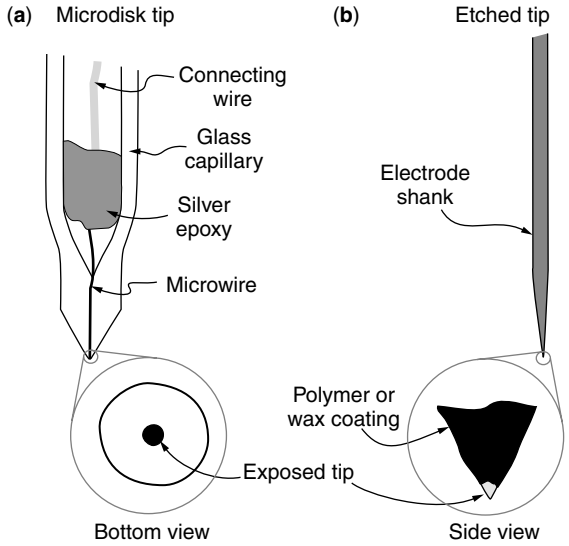


Figure 3. Schematic diagram for probes that have (**a**) an embedded disk and (**b**) a conical (etched) electrode geometry.

a large insulating radius. A truncated cone tip gives a different I–L curve than, say, a hemispherical tip. Thus, no single equation for an I–L curve includes all geometric possibilities. Equations (2) and (3) are strictly valid only for disk electrodes whose insulating radius is 10 times the electrode radius (a common configuration). For negative feedback, a smaller or larger insulating radius causes the probe I–L curve to descend less or more steeply, respectively (52,53) because both the sample surface and the insulating material provide the mediator blocking necessary for negative feedback. For example, at $L = 1$, $I_T = 0.534$, 0.302, and 0.213 for insulator-to-electrode radius ratios of 10, 100, and 1,000, respectively (8). The effect of insulating radius is less critical in positive feedback because the probe current is mainly due to regeneration of the mediator in the region of the substrate directly below the probe tip.

Probes that have a tip electrode size less than 1 μm are easier to make in a cone or spherical-segment (e.g., hemispherical) geometry (10,49,54–56). Unfortunately, the I–L sensitivity of these probes in the feedback mode is much less than that of the disk-shaped electrodes. Further, most construction methods for very small probes (e.g., electrochemically etched tips) do not produce a reproducible tip electrode shape (57). Good estimates of the tip electrode shape can be made by fitting the experimental to simulated I–L curves (49), but it is unlikely that fitting will be used for general imaging purposes. Thus, at present, non-disk-shaped tip electrodes should be used only to provide qualitative SECM feedback images.

Rastering the probe across the sample surface and recording the probe current produces topographic images of surfaces in the feedback mode. Both resolution and sensitivity increase with a decrease in tip–sample distance. Typically, the probe tip is positioned within one tip electrode radius of the sample. Because the probe response is different for insulating (blocking) and conducting surfaces, areas of both types can be

distinguished in feedback mode imaging. Conducting areas of a surface have a positive-feedback response and are identified by currents larger than $i_{T,\infty}$; insulating areas have a negative-feedback response when tip currents are less than $i_{T,\infty}$. Note that the nonlinear I–L relationship distorts the relationship to the actual topography of the images made in the feedback mode. However, a true topographic image can be made from the current image by employing theoretical I–L equations [e.g., Eqs. (2) and (3)] (7,58). The feedback image in Fig. 4a illustrates the concept of positive and negative feedback in a test image taken of an interdigitated electrode array. Au electrodes 3 μm wide are separated by 5-μm wide SiO_2 insulators. The primary contrast is the transition from positive to negative feedback; the Au electrodes exhibit positive feedback and the largest current. Topographic features such as damage to the Au electrodes are visible as slight depressions on the second and third band from the right.

Lateral and vertical resolution in the feedback mode depends on the probe electrode size and tip–sample separation (52,59). The ability to discriminate between insulating and conducting regions can be used as one estimate of lateral resolution. Scanning the probe over a conducting–insulating boundary ideally produces an instantaneous change in probe current (from positive to negative feedback). Experimentally, the change occurs across a region of finite width Δx. For zero tip–sample separation ($L = 0$), Δx equals the tip electrode diameter. For larger separations, the lateral resolution degrades as diffusion broadens the mediator concentration profile. Unfortunately, no comprehensive theoretical answer for lateral resolution has been published. An experimental study reported that the resolution function has the form (9)

$$\Delta x/a = 4.6 + 0.111\,L^2. \tag{4}$$

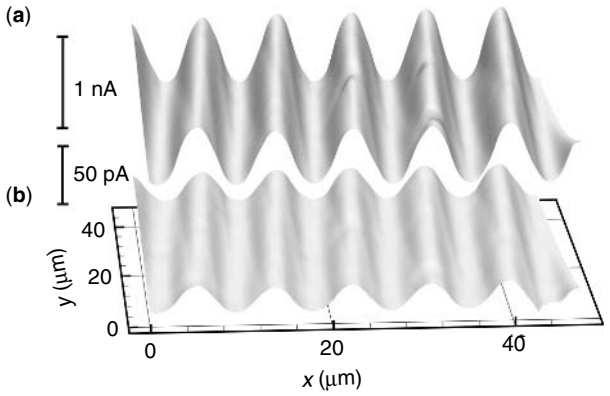


Figure 4. SECM images of an interdigitated array electrode that consists of 3-μm wide Au electrodes separated by 5-μm wide strips of SiO_2 insulator. Images acquired by using a 1.3-μm tip–substrate distance at a scan rate of 5 μm/s. Images are presented as shaded surface plots and larger tip currents appear as higher points on the image. (**a**) Feedback image experimental conditions: 2 mM $Ru(NH_3)_6^{3+}$, pH 3.2 phosphate-citrate buffer, 2-μm diameter Pt tip, $E_{tip} = -0.3$ V vs. Ag/AgCl reference, $E_{sub} = 0.0$ V, cathodic currents are positive. (**b**) GC image, experimental conditions: as in (**a**) except $E_{tip} = 0.0V$, $E_{sub} = -0.090$ V, and anodic currents are positive.

This equation is probably overconservative because the best resolution is 2.3 times the tip electrode diameter, which is contrary to intuition and other published data [see (60) for resolution superior to that predicted by Eq. (4)]. However, the functional form of resolution by distance seems more reasonable. Regions smaller than a tip electrode diameter cannot be resolved, but small active regions embedded in an insulating matrix can be recognized if the regions are separated by at least a tip electrode diameter.

The vertical resolution of the SECM depends primarily on the probe tip–sample separation and the dynamic range of the probe current measurement. Equations (2) and (3) allow estimating the minimum resolvable vertical change in distance at a disk-shaped probe tip electrode. For example, a 1% change in I occurs from a change from $L = 1$ to $L = 1.0137$. If this difference in current is measurable, it corresponds to a theoretical 0.068-μm change at a 5-μm radius tip electrode and 5-μm tip–substrate separation or a theoretical 0.014-μm change at a 1-μm radius tip electrode and a 1-μm separation.

Increasing the resolution of SECM feedback images is, in principle, possible by removing diffusional blurring by applying digital image processing techniques. Unfortunately, the nonlinear response in the SECM feedback mode has prevented a general formulation of a deconvolution function. It has been shown that approximate *deblurring* functions improve image contrast (61,62).

The ultimate resolution in SECM may have been demonstrated by Fan and Bard who reported a resolution approaching 1 nm for imaging biological molecules (63). To achieve this high resolution, they used an etched tungsten tip electrode and a very thin layer of water naturally present on the sample in a humid atmosphere. The thin layer of solution minimizes the blurring effect of diffusion.

Reaction-Rate Mode

The general case of feedback imaging, known as reaction-rate imaging, exists when the rate of ET at the probe tip electrode or sample is considered. Positive and negative feedback are limiting cases that occur when the ET rate is at the diffusion limit or zero, respectively. Between those limits, the I–L curves change from curves similar to positive feedback to what appears to be negative feedback.

Note that deviations from the theoretical positive-feedback curve may arise even when mediators that have very rapid ET kinetics are used. This is made clear by considering that the SECM response for a disk-electrode tip at very small tip–sample separations is approximately that of a twin-electrode, thin-layer cell (64,65). For this configuration, the probe current is

$$i_T = 4\,nFDCA/d. \tag{5}$$

where A is the tip electrode area. The very small separations in SECM produce impressively large values of the mass transfer coefficient D/d (47). As an example, for $d = 0.2$ μm and $D = 1 \times 10^{-5}$ cm^2 s^{-1}, the mass transfer coefficient is 0.5 cm s^{-1}. Heterogeneous rate constants for mediator tip electrodes or substrate ETs must be about 10 times larger to preserve the theoretical response

for positive feedback. Thus, the SECM can measure or detect change in heterogeneous rates for ET reaction at the tip electrode or the substrate, as long as the opposing electrode is not allowed to limit the overall reaction. This is not difficult for mediators with reasonably rapid ET rates because the opposing substrate or tip electrode can be poised at a potential sufficiently past the oxidation or reduction potential to drive the reaction to its diffusion-controlled limits. A beneficial consequence of rapid diffusional transport in the positive-feedback mode is that the probe is less sensitive to convection caused by, for example, probe movement during SECM imaging. Because of the high mass-transport rate and the ease of making measurements at steady state, the SECM is a very promising method for fundamental investigation of rapid interfacial ET rate constants. Calculations suggest that it is possible to use the SECM tip electrode to measure a rate constant as high as 10–20 cm s^{-1} under steady-state conditions (66).

When the mediator and sample are chosen to produce a kinetic limitation of the feedback process, a number of interesting experiments become possible. Calculations of I–L curves for quasi-reversible and irreversible ET kinetics are available (67–69) that permit quantitative measurements of ET rates at micron-sized regions of surfaces. At electronically conducting surfaces, kinetic information can be extracted in the tip-sized regions that allow direct imaging of active or passive sites on surfaces. For example, greater ET activity is easily observed on the Au region of a composite carbon/Au electrode (18). Feedback current also supplies potential independent kinetic information for chemical reactions that involve ET at the interface such as immobilized enzymatic activity or oxidative metal dissolution.

An example of a reaction-rate image is presented in Fig. 5. The SECM was used to prepare a pattern of oxidized regions on a glassy-carbon electrode using a direct-mode oxidation process (see later). Reduction of Fe^{3+} ions is catalyzed on oxidized carbon regions and the reaction-rate image is based on the increased reaction rate on oxidized carbon. Successive images of the same region were collected when the carbon electrode potential was at −200 (A), −400 (B), −600 (C), and −800 mV (D). At potentials less negative than −800 mV, the "face" pattern appears on the image due to a higher rate of ET in the oxidized regions compared to the native carbon. At −800 mV, the pattern disappears because the rate of ET in all regions of the carbon surface is at the diffusion limit and the contrast disappears. Note that topography is not the predominant contrast mechanism in the images, although a small pit in the electrode surface is clear in (D) and less so in the other images.

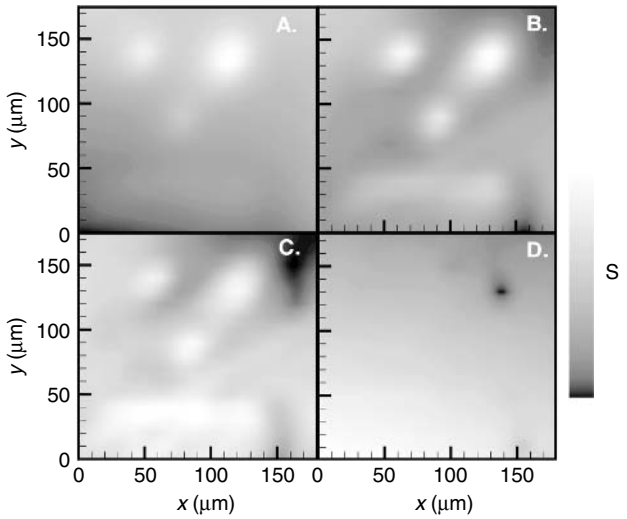


Figure 5. Reaction-rate SECM images of an oxidized carbon pattern on a glassy-carbon electrode. Imaging conditions: 10-μm Pt tip, $E_{tip} = 700$ mV vs. Hg/Hg_2SO_4 reference, 2 mM Fe^{2+}/1 M H_2SO_4, 2.5-μm tip-substrate distance, scan rate of 10 μm/s. Images are presented as gray-scale plots and larger cathodic tip currents are lighter shades. (A) $E_{sub} = -200$ mV, $S = 3.6$–5.0 nA. (B) $E_{sub} = -400$ mV, $S = 5.0$–7.3 nA. (C) $E_{sub} = -600$ mV, $S = 6.4$–8.7 nA. (D) $E_{sub} = -800$ mV, $S = 7.7$–9.3 nA. Conditions for SECM production of oxidized pattern: 10-μm Pt tip, distilled water solution, tip–substrate separation <1 μm, 0.46 μA is passed between tip (cathode) and substrate for 5 s to produce spots or while moving the tip at 2 μm/s to produce linear features.

Generation/Collection Mode

Using the GC mode is most appropriate when the sample itself produces an electrochemically detectable material. GC is ideally a passive mode where the probe is used as a microscopic chemical sensor. Examples include using the GC mode to investigate corroding metal surfaces (25,70), ionic and molecular transfer through porous material (71–73), and oxygen generation and uptake in living material (29,31,32). The GC mode has several advantages over the feedback mode: undesirable chemical interactions between mediator and sample are eliminated, chemical species produced by the sample can be identified by voltammetry, and both potentiometric and amperometric tip electrodes can be used to extend the range of detectable chemical species.

The major disadvantages of GC imaging are poor vertical sensitivity and lateral resolution. Unlike the feedback mode, there is not an artificially maintained diffusion gradient between probe tip and sample, and thus the concentration gradient of the sample-generated species may extend out to several hundred micrometers (74). In addition, the gradient is ill defined because natural convection, vibration, and probe movement can perturb the diffusion-based concentration profile. Unlike the feedback mode, the probe signal changes little near the surface, and this causes difficulty in locating the tip at the surface. Additional equipment such as an optical microscope is often required for positioning. Diffusional broadening also compromises lateral resolution. Consider that a species produced at a point on the surface will diffuse, on average, a distance x during time t:

$$x = \sqrt{2Dt}. \tag{6}$$

Thus, for example, sources smaller than 10 μm produce a diffusion *cloud* larger than themselves at times greater than 0.1 s. Closely spaced features overlap, and a diffuse background signal interferes with image interpretation.

Despite its poorer resolution, GC imaging is a popular method. The increased chemical sensitivity and selectivity are advantages. In addition, there is no sensitivity penalty for conical or hemispherical tip electrodes, so probe preparation is simplified. Techniques to improve resolution have also been developed. Imaging when the probe is as close as possible to the sample or stirring the solution minimizes diffusional broadening effects. The use of chemical scavengers can also be effective in reducing the extent to which a species diffuses into solution. An example is the addition of a small concentration of oxidizing agent to a sample solution; a reduced species produced by the sample can diffuse only a short distance before being destroyed by the added oxidant. Adjusting the scavenger concentration produces a much smaller reaction zone than the normal diffusion zone and minimizes diffusional broadening. Although not a general solution to diffusional broadening, a number of scavenging agents have been used or proposed: pH buffers, complexants, redox agents, and enzymes. Recent studies indicate that scavenging agents can also improve resolution in feedback imaging (a *chemical lens*) (75).

The GC image in Fig. 4b can be compared to the image of the identical region acquired in feedback mode (Fig. 4a). The 3-μm wide Au electrodes of the substrate were polarized to produce a small amount of $Ru(NH_3)_6^{2+}$ ion by reducing the $Ru(NH_3)_6^{3+}$ in solution. The probe electrode potential is set to oxidize $Ru(NH_3)_6^{2+}$, and thus the image shows regions of higher $Ru(NH_3)_6^{2+}$ ion concentration at the Au electrodes. The resolution in this image is similar to the feedback image primarily due to the close tip–substrate spacing (about 1.3 μm) during imaging.

The potentiometric GC mode relies on an ISE as the probe. The main advantage of the potentiometric GC mode is that the probe is a passive sensor of the target ion activity and does not cause the depletion or feedback effects noted before. In addition, the probe can be used to detect ions such as Ca^{2+} or NH_4^+ that would be difficult or inconvenient to detect voltametrically. The selectivity of the tip electrode for the ion of interest is an advantage, but this requires separate probes for each species. The response of an ISE is, in general, a voltage proportional to the logarithm of the ion's activity, but special consideration must be made for potentiometric GC to provide a rapid response when using microscopically small tip electrodes (76).

INSTRUMENTATION

The scanning electrochemical microscope (also abbreviated SECM) requires a probe positioning system, instrumentation to generate and measure the probe signal, and a data recorder. Many investigators have built their own SECM instruments; however, a commercial instrument designed specifically for SECM is now available from CH Instruments (Austin, TX). A schematic overview of an SECM is shown in Fig. 6. The probe tip is attached to a three-axis positioner. The type of positioner used in the SECM is determined mainly by the size of the probe used. Positioners for probes whose tips are smaller than about 1 μm are generally based on designs used

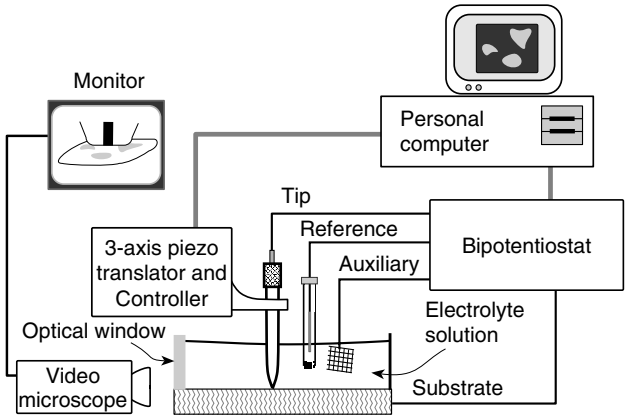


Figure 6. Schematic diagram of the SECM instrument.

in scanning tunneling or force microscopes, which use piezoelectric tube actuators. These positioners allow precise movements on the nanometer scale and are small, so that they can be placed in small, stiff mounts. This, in turn, makes these positioners less sensitive to positioning errors due to mechanical vibration and thermal drift. Most SECM experiments, however, have used probes whose tip electrodes are larger than 1 μm, and, piezotube positioners are unsuitable for these probes. A primary difficulty is that the larger size and mass of the probe is not suited to the piezotube positioner. In addition, the scan size available when using a piezotube positioner is not larger than 100 μm. Stepper-motor positioners are used for larger probes. The positioners in many of the homebuilt instruments and in the commercial example are based on a unique piezoelectric positioner, the Inchworm from Burleigh Instruments (77). These motors can produce probe scan rates greater than 100 μm/s across a 2.5-cm range of motion at a minimum resolution of less than 100 nm. Thus, these motors can act as both coarse and fine positioners. Many positioning motors operate in an open-loop configuration, and operators rely on frequent calibrations for accurate positioning. Some homebuilt instruments use closed-loop positioners that rely on integrated position encoders to provide superior positioning accuracy (at a higher price).

A potentiostat is required for feedback and amperometric GC imaging, to control the probe electrode potential and to amplify the probe current. The use of a bipotentiostat, which allows simultaneous control of two working electrodes versus common reference and auxiliary electrodes, is convenient when controlling the sample potential is also required. Alternately, a second potentiostat can be used to control a sample electrode. The probe (and, optional, sample) potentiostat should be sufficiently sensitive to allow measuring of the low current flow at an UME, which may be in the pA range. Many older potentiostat instruments are not suitable without some user modification to decrease noise and increase sensitivity. A Faraday cage around the SECM cell and probe to reduce environmental noise is advisable. For potentiometric SECM operation, an electrometer is required to buffer the high impedance of the microscopic tip electrodes.

A personal computer is used to record and display images and to control the probe positioner. Homebuilt instruments rely on custom software for programming the probe movement and displaying the probe signals in realtime (4,78). A video microscope system is often useful to assist in positioning the probe at the surface or at sample location. When a video microscope is used, the sample cell incorporates a window to allow direct observation of the tip–sample approach.

Probes

It is not an exaggeration to say that the most critical component of the SECM instrument is the probe. Two common probe types, the microdisk and etched tip electrode, are shown in Fig. 3. Microdisk electrodes are prepared by sealing a microscopic metal wire or carbon fiber in glass or epoxy and polishing or cutting to expose a disk embedded in insulator (3). These electrodes are *tapered* by carefully grinding away the surrounding insulator to produce a tip whose insulator diameter is no larger than about 10 electrode diameters. Alternately, cone-shaped tips are made by electrochemically etching metal wire or carbon fibers and then coating them with Apiezon wax or other polymers (49,54,56). The surface tension of the coating prevents coating the sharp tip apex. Conical tip electrodes are also available commercially from electrophysiology suppliers such as FHC (Brunswick, ME) and World Precision Instruments (Sarasota, FL).

Many potentiometric probes have been proposed for GC SECM. A well-established method is based on micropipets filled with neutral carrier membranes or ionophores. These tips can be made in submicrometer dimensions (76). The literature contains descriptions of probes and experiments used for measuring of pH (43–45,74), NH_4^+, K^+ (45), and Zn^{2+} (45,79). A pH-sensitive probe based on an antimony electrode (80) has the advantage that it can be converted back and forth *in situ* from a voltammetric electrode to a potentiometric electrode by oxidizing antimony to antimony oxide. The pH-sensitive antimony oxide surface allows operating in a potentiometric GC mode, and the antimony surface allows precision positioning in the feedback mode.

Micropipets can also configured as *voltammetric* ion-selective electrodes (ISEs) (81,82). In this configuration, an organic solvent plus electrolyte fills the pipet. Application of a potential transfers ions in or out of the organic phase (pipet) and into an aqueous phase that contains the substrate. The parallels between ion transfer at a voltammetric ISE and electron transfer at a metallic electrode are strong, and much of the theoretical results obtained at metal SECM tips can be applied to voltammetric ISE probes. Thus, feedback imaging is possible using electroinactive ions as mediators. The advantage is that micropipets can be made reproducibly small without significant expense. The organic phase contents necessarily vary depending on the ion used as mediator. For example, K^+ ion electrodes whose radii are less than 1 μm were produced by filling borosilicate capillaries with 1,2-dichloroethane containing 10 mM valinomycin and 10 mM of the ionophore ETH 500 (83). Alternately, crown ethers can be used to facilitate ion transfer (46,57).

Constant-Separation Imaging

Most SECM images are acquired in a constant-height mode in which the probe is scanned in a reference plane above the sample surface. Because the feedback mode requires probe tip–sample spacing of about one tip electrode diameter for good resolution, care must be taken to prevent probe crashes, when examining samples that have high relief or a tilt, particularly for small probes. Constant-separation (or constant-distance) imaging uses a servomechanical mechanism to keep the tip–substrate distance constant while imaging, thus ensuring uniform resolution throughout an image and preventing probe crashes. Although relatively straightforward in scanning tunneling or force microscopy where the magnitude of the probe signal is related to the separation, constant-separation imaging in SECM is more challenging. In GC mode, the probe signal does not indicate the tip–substrate separation. In the feedback mode, the signal depends on separation and also on the sample conductivity. For surfaces known to be completely insulating or conducting, implementing of a constant-current mode is straightforward and similar in methodology to STM (scanning tunneling microscopy) studies (1). For example, a piezoelectric element constantly repositions the probe to minimize the deviation of the probe current from an established reference current. However, when a surface has both insulating and conducting regions, this method cannot work. Any general method for controlling tip–sample separation requires more information than supplied by a probe signal.

Although no completely satisfactory method is available for constant-separation imaging, a number of methods have had limited success. One method is to impose a small-amplitude (100 nm or less), high-frequency vertical modulation of the probe position (tip position modulation, TPM) during feedback imaging (84). The low-frequency *normal* signal is unaffected by the modulation, but the high-frequency ac current has a 180° phase shift when the probe is moved from an insulating to a conducting surface. This additional information is sufficient to set the proper probe motion and reference current level on the servomechanism. A second method, the *picking mode*, monitors a transient probe current caused by repeated, rapid up and down approaches over a large probe–substrate separation (85,86). The transient current is affected by tip–substrate separation but not by substrate conductivity and, thus, can be used to position the probe automatically. Perhaps the most general method is based on the shear-force mode developed initially for the near-field scanning probe microscope (NSOM). The shear-force method uses a small-amplitude lateral probe vibration. As the probe tip approaches the sample, hydrodynamic forces act to damp the vibratory motion. Using a laser (87) or a tuning-fork sensor (59,88), the vibration intensity can be monitored and used to maintain constant separation. Because a specific electrochemical response is not required, the shear-force method can be suitable for either GC or feedback imaging. The relative

newness of the method, the requirement for special probe sizes, and the exacting experimental technique have so far, limited widespread adoption of the shear-force mode.

APPLICATIONS

Corrosion Science

SECM is increasingly valuable in investigating corrosion processes on metallic samples. Voltammetric or amperometric GC SECM can produce concentration map images of electroactive species produced or depleted on corroding surfaces. Examples from the literature include Fe ions (25); O_2, Co, and Cr ions (70); S (24); and Cu and Ni ions (89). Alternatively, the use of potentiometric GC SECM images of corroding surfaces can yield information about local pH values. A pH-sensitive probe that employed a neutral-charge-carrier ionophore imaged pH changes that had a sensitivity of less than 0.1 pH units above $AlFe_3$ and MnS inclusions in aluminum and steel alloys, respectively (44,90,91).

SECM images of passive metal surfaces can be used to preidentify microscopic sites of future corrosion. Corrosion precursor sites on titanium were detected by their ability to oxidize Br^- ions to Br_2 in solution. These microscopic active sites were identified by SECM maps of Br_2 concentration near the titanium surface (28,92,93). Pit formation, it was subsequently found, coincided with Br_2 forming regions. The power of the SECM method to examine metal oxide growth is evident in Fig. 7 (94). An SECM image of a tantalum electrode in the presence of I^- is featureless at low potentials; however, raising the electrode potential to produce an oxide film generates a current that corresponds to I^- and tantalum oxidation. The numbered peaks that appear in the SECM image are regions of high I_3^- concentration, produced by I^- oxidation at oxide-free (or thin oxide) locations. The SECM images at different tantalum electrode potentials show that oxide formation is not uniform and that active spots form and passivate at different spatial locations and electrode potentials. This behavior is not predicted from the voltammetric behavior of the electrode (21,94). The SECM can also be used to investigate corrosion by using the probe to generate aggressive conditions in a small region of solution near a metal surface. Electrochemical generation of Cl^- ion by the SECM probe can generate localized pitting corrosion on iron, steel, and aluminum metal (25,95,96).

In a related application, controlled dissolution of material is an area in which the ability of the SECM to produce a defined chemical environment near the material surface provides a significant advancement. Rates of dissolution of even very soluble materials can be obtained by using the electrolysis at the probe tip electrode to produce a local undersaturation of the dissolving material. Quantitative details about the rate and mechanism of the dissolution reaction are available from the probe current as a function of time and distance to the sample (97). Examples include the dissolution of the (010) surface of potassium ferrocyanide trihydrate (58), AgCl (98), Cu (99), and Au (100) crystallites.

Biological and Polymeric Materials

The two broad interests in SECM of biological and polymeric materials are investigation of transport of molecules or ions across interfaces and observation of chemical and biological reactions on a substrate. The activity of enzymes immobilized on solid or gel supports is readily observed by using the GC mode to image the products of the enzymatic reaction (33,101), and efforts have been made to develop SECM-immunoassay techniques (102). Living tissue has also been examined by

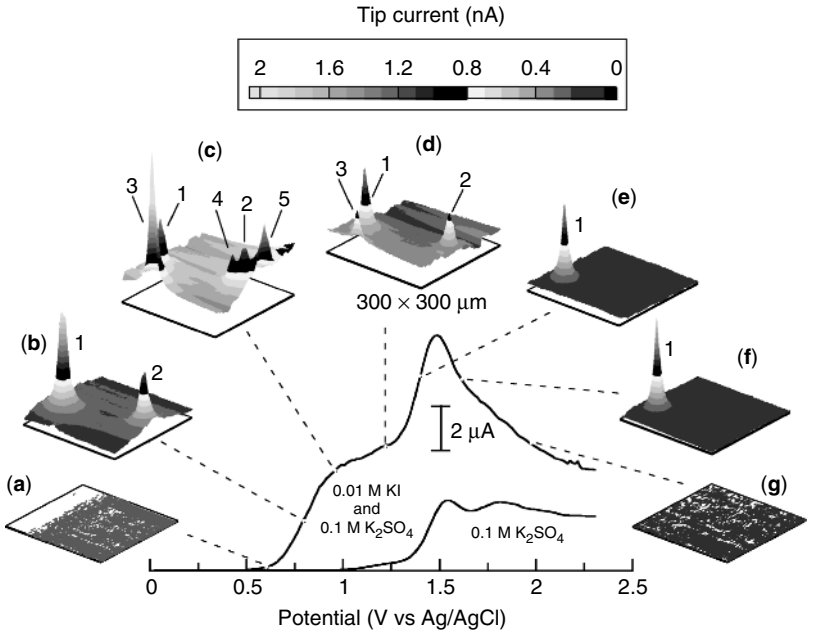


Figure 7. SECM images of a 300 × 300-μm region of a Ta/Ta_2O_5 oxide electrode in 10 mM KI/0.1 M K_2SO_4 at different electrode potentials. The probe is an 8-μm diameter carbon electrode held at a potential of 0.0 V scanned at 10 μm/s and the tip–sample separation was about 5 μm. The numbered peaks in the SECM images correspond to active regions for oxidation of I^- on the Ta/Ta_2O_5 surface. The voltammetric curves for Ta oxidation with and without KI are shown for comparison. (Reprinted with permission from *Anal. Chem.*, **71**, 3,166–3,170 (1999). Copyright 1999 American Chemical Society.)

imaging enzyme activity in mitochondria (22) and changes in O_2 concentration or pH during metabolic activity or photosynthesis (29,32,103,104).

Ion transport through pores or channels in biological membranes can be directly imaged with SECM. In most examples, quantitative models of the local transport properties are available from the data. A modeled system is the transport of ions and molecules through hairless mouse skin that has applications in transdermal drug delivery (38,71,105). Similarly, SECM was used to examine fluid flow through the dentine tubules of teeth (37,98).

SECM is often used to examine the transport properties of ions at polymeric material interfaces, for example, ion ejection or uptake during oxidation and reduction of conducting polymer films (34,106–108). Figure 8 is a graphic illustration of an SECM image in a fundamental study of size-selective permeation of ions through a patterned, electropolymerized Fe(5-amino-1,10-phenanthroline)$_3^{2+}$ film (35). Images 8A, B, and C are GC images of a region over four indium tin oxide (ITO) electrodes using three redox probes: Ru(NH$_3$)$_6^{3+}$, Fe(phenanthroline)$_3^{2+}$, and Fe(bathophenanthroline(SO$_3$)$_2$)$^{4-}$ whose diameters are 5.5, 13, and 24 Å, respectively. Images 8A', B', C' are for the same ions but use the ITO electrodes covered with a ~12-nm layer of electropolymerized Fe(5-amino-1,10-phenanthroline)$_3^{2+}$. Size-selective permeation is evident from the reduced signal of Ru(NH$_3$)$_6^{3+}$ and nearly complete loss of the signal for the larger ions Fe(phenanthroline)$_3^{2+}$ and Fe(bathophenanthroline(SO$_3$)$_2$)$^{4-}$. The spatial uniformity of the polymer can also be assessed from the SECM images, and the effects of confounding phenomena such as pinholes or cracks in the film can be dismissed.

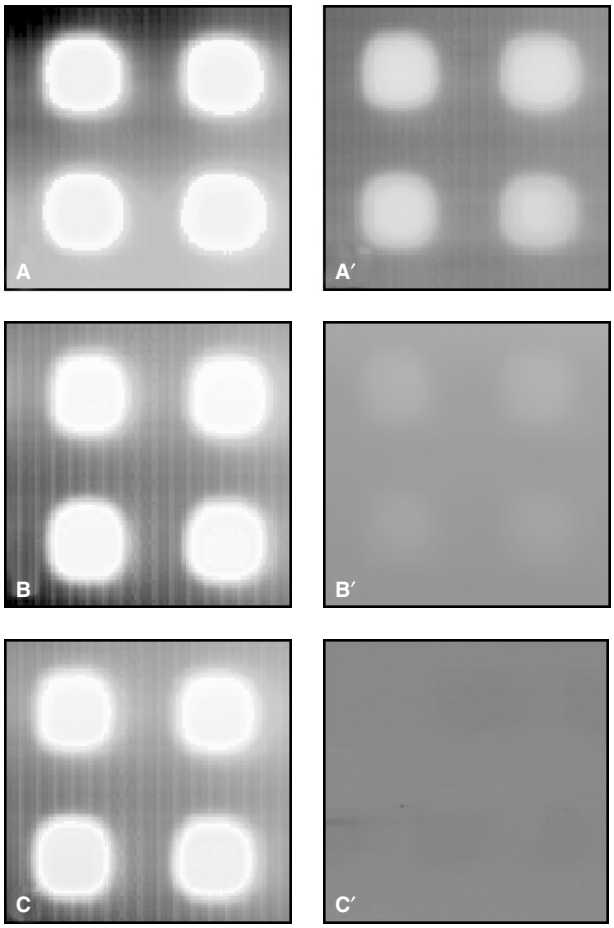


Figure 8. SECM images of a patterned array of 50-μm diameter ITO electrodes. The probe is an 8-μm diameter carbon-fiber electrode held at a potential of 0 V (vs. Ag/AgCl) scanning at a rate of 30 μm/s and had a 2-μm tip-sample separation. Left images (A, B, C) are scans of bare ITO. Right images (A', B', C') are scans of the ITO covered by a ~12-nm film of electropolymerized Fe(5-amino-1,10-phenanthroline)$_3^{2+}$. Conditions: (A, A') 4.8 mM Ru(NH$_3$)$_6^{3+}$, $E_{sub} = -0.35$ V, $S = 4.3$ nA; (B, B') 5.1 mM Fe(phenanthroline)$_3^{2+}$, $E_{sub} = +0.5$ V, $S = 7.1$ nA; (C, C') 4.3 mM Fe(bathophenanthroline(SO$_3$)$_2$)$^{4-}$, $E_{sub} = +1.0$ V, $S = 1.2$ nA. Reprinted with permission from *Anal. Chem.* **72**: 3,122–3,128 (2000). Copyright 2000 American Chemical Society.)

An important application of SECM technology involves fundamental studies of charge or molecular transport across the interface of two phases such as two immiscible electrolyte solutions (ITIES) (45,81,109–112) or gas–liquid interfaces (113–115). These studies allow all electrodes to be present in one phase, simplifying the experiment. The important aspect of these experiments is that the SECM instrument can place the SECM probe very close to the interface. The SECM probe electrode can be used to access a wide range of mass-transfer coefficients (see earlier) in a small region of the interface without the complications that appear in conventional measurements such as high solution resistance and large interfacial capacitance (81,83,111,115).

Non Imaging Applications

The utility of the SECM extends beyond imaging applications. The electrochemical probe lends itself to many nonimaging experiments that, in many cases, prove synergistic in combination with SECM imaging. A common application is to use the SECM microscope to modify surfaces locally.

The microreagent mode uses the SECM tip electrode to produce, via electrolysis, a local chemical environment that is different from the bulk environment. Changes in pH, redox activity, ion activity, and ionic strength are possible. The area of the substrate exposed to the local chemical environment is only slightly larger than the tip electrode diameter, and the volume of solution between the probe and the substrate can be quite small. For example, a volume of only 25 fL exists in the gap formed by a 5-μm radius disk-shaped tip electrode separated from the substrate by 1 μm. Consequently, changing the local chemical environment within this small volume is possible on a millisecond time scale. Rapid diffusion of the probe-generated reagents outside the vicinity of the tip electrode ensures that the reaction remains limited to the tip electrode area.

The other substrate modification method is the *direct* mode. In contrast to normal procedure in electrochemistry, which employs a large area auxiliary remotely positioned with respect to the working electrode, the direct mode of SECM employs the tip electrode as a microscopic

auxiliary electrode held at a micrometer distance from the sample electrode. Thus, faradaic current that would normally be distributed across the electrode is confined to a small region near the microauxiliary electrode (116). When the tip is positioned close to the substrate, any faradaic reaction at the substrate is limited to the region near the tip. Any electrochemical reaction that can be produced at the bulk surface can be performed, at least in principle, in a local region of the surface by using the direct mode. For example, metal etching, metal deposition, and polymer deposition are all possible in microscopic regions. Figure 5 illustrates direct mode oxidation of a carbon electrode. To form a pattern of oxidized spots and lines, a 10-μm diameter Pt tip is positioned close to the carbon surface, and 0.46-μA a current is passed between the tip (cathode) and carbon (anode) in distilled water. Reaction-rate imaging is then used to verify the presence of the oxidized region (see earlier).

Acknowledgments

The author acknowledges the National Science Foundation, the State of Mississippi, and the Mississippi EPSCOR program for support by grants DBI-9987028 and EPS-9874669. The assistance of Mario Alpuche Aviles, Robert Tenent, and Emad El-Deen El-Giar is gratefully acknowledged.

ABBREVIATIONS AND ACRONYMS

A	area
a	electrode radius
Ag/AgCl	silver/silver chloride reference electrode
C	concentration of redox species
CV	cyclic voltammetry
D	diffusion coefficient of redox species
d	distance between tip and substrate in SECM
ET	electron transfer
F	faraday constant
GC	generation/collection mode of the SECM
Hg/Hg_2SO_4	mercury/mercury(II)sulfate reference electrode
ISE	ion-selective electrode
i_T	steady-state limiting current at an SECM tip electrode
I_T	normalized SECM tip current, $i_T/i_{T,\infty}$
$I_T(L)$	normalized SECM tip current at the normalized distance L
$i_{T,\infty}$	steady-state limiting current at an SECM tip electrode at infinite tip-substrate separation
$I_{T,ins}(L)$	normalized SECM tip current over an insulator at the normalized distance L
ITO	indium tin oxide
k	heterogeneous rate constant
L	the normalized distance from tip to substrate, d/a in SECM
n	number of electrons transferred in a redox event
NSOM	near-field scanning optical microscope
SCE	saturated calomel electrode
SECM	scanning electrochemical microscope or scanning electrochemical microscopy
SPM	scanning probe microscopy
STM	scanning tunneling microscopy
TPM	tip position modulation
UME	ultramicroelectrode

BIBLIOGRAPHY

1. G. Binnig and H. Rohrer, *Helv. Phys. Acta* **55**, 726–735 (1982).
2. R. M. Wightman, *Science* **240**, 415–420 (1988).
3. R. M. Wightman and D. O. Wipf, in A. J. Bard, ed., *Electroanalytical Chemistry*, vol. 15, Marcel Dekker, NY, 1989, pp. 267–353.
4. A. J. Bard, F. -R. Fan, and M. V. Mirkin, in A. J. Bard, ed., *Electroanalytical Chemistry*, vol. 18, Marcel Dekker, NY, 1994, pp. 244–370.
5. A. J. Bard and M. V. Mirkin, eds., *Scanning Electrochemical Microscopy*, Wiley & Sons, NY, (2001).
6. M. V. Mirkin and B. R. Horrocks, *Anal. Chim. Acta* **406**, 119–146 (2000).
7. A. J. Bard, F. -R. F. Fan, D. T. Pierce, P. R. Unwin, D. O. Wipf, and F. M. Zhou, *Science* **254**, 68–74 (1991).
8. J. Kwak and A. J. Bard, *Anal. Chem.* **61**, 1,221–1,227 (1989).
9. G. Wittstock, H. Emons, M. Kummer, J. R. Kirchhoff, and W. R. Heineman, *Fresnius' J. Anal. Chem.* **348**, 712–718 (1994).
10. C. M. Lee, C. J. Miller, and A. J. Bard, *Anal. Chem.* **63**, 78–83 (1990).
11. A. J. Bard, F. -R. F. Fan, J. Kwak, and O. Lev, *Anal. Chem.* **61**, 132–138 (1989).
12. J. Y. Kwak, C. M. Lee, and A. J. Bard, *J. Electrochem. Soc.* **137**, 1,481–1,484 (1990).
13. C. M. Lee and A. J. Bard, *Anal. Chem.* **62**, 1,906–1,913 (1990).
14. I. C. Jeon and F. C. Anson, *Anal. Chem.* **64**, 2,021–2,028 (1992).
15. K. Borgwarth, C. Ricken, D. G. Ebling, and J. Heinze, *Fresnius' J. Anal. Chem.* **356**, 288–294 (1996).
16. G. Wittstock, T. Asmus, and T. Wilhelm, *Fresnius' J. Anal. Chem.* **367**:346–351–(2000).
17. A. L. Barker, P. R. Unwin, S. Amemiya, J. Zhou, and A. J. Bard, *J. Phys. Chem. B* **103**, 7,260–7,269 (1999).
18. D. O. Wipf and A. J. Bard, *J. Electrochem. Soc.* **138**, L4–L6 (1991).
19. D. O. Wipf and A. J. Bard, *J. Electrochem. Soc.* **138**, 469–474 (1991).
20. S. B. Basame and H. S. White, *J. Phys. Chem. B* **102**, 9,812–9,819 (1998).
21. S. B. Basame and H. S. White, *Langmuir* **15**, 819–825 (1999).
22. D. T. Pierce and A. J. Bard, *Anal. Chem.* **65**, 3,598–3,604 (1993).
23. R. C. Engstrom, M. Weber, D. J. Wunder, R. Burgess, and S. Winguist, *Anal. Chem.* **58**, 844–848 (1986).
24. C. H. Paik, S. B. Basame, H. S. White, and R. C. Alkire, *Proc. Electrochem. Soc.* **99–28**, 122–130 (2000).
25. D. O. Wipf, *Colloid Surf. A* **93**, 251–261 (1994).
26. Y. Y. Zhu and D. E. Williams, *J. Electrochem. Soc.* **144**, L43–L45 (1997).
27. S. B. Basame and H. S. White, *J. Electrochem. Soc.* **147**, 1,376–1,381 (2000).
28. N. Casillas, S. J. Charlebois, W. H. Smyrl, and H. S. White, *J. Electrochem. Soc.* **140**, L142–L145 (1993).

29. M. Tsionsky, Z. G. Cardon, A. J. Bard, and R. B. Jackson, *Plant Physiol.* **113**, 895–901 (1997).
30. J. Wang, L. -H. Wu, and R. Li, *J. Electroanal. Chem.* **272**, 285–292 (1989).
31. C. M. Lee, J. Y. Kwak, and A. J. Bard, *Proc. Natl. Acad. Sci. U.S.A.* **87**, 1,740–1,743 (1990).
32. T. Yasukawa, Y. Kondo, I. Uchida, and T. Matsue, *Chem. Lett.* 767–768 (1998).
33. T. Yasukawa, N. Kanaya, D. Mandler, and T. Matsue, *Chem. Lett.* 458–459 (2000).
34. M. Arca, M. V. Mirkin, and A. J. Bard, *J. Phys. Chem.* **99**, 5,040–5,050 (1995).
35. M. E. Williams, K. J. Stevenson, A. M. Massari, and J. T. Hupp, *Anal. Chem.* **72**, 3,122–3,128 (2000).
36. B. D. Bath, R. D. Lee, H. S. White, and E. R. Scott, *Anal. Chem.* **70**, 1,047–1,058 (1998).
37. J. V. Macpherson, M. A. Beeston, P. R. Unwin, N. P. Hughes, and D. Littlewood, *Langmuir* **11**, 3,959–3,963 (1995).
38. E. R. Scott, A. I. Laplaza, H. S. White, and J. B. Phipps, *Pharmaceut. Res.* **10**, 1,699–1,709 (1993).
39. E. R. Scott, H. S. White, and J. B. Phipps, *Solid State Ionics* **53–56**, 176–183 (1992).
40. K. C. Melikov and I. A. Ershler, *Biol. Membrany* **13**, 423–429 (1996).
41. N. P. Hughes, D. Littlewood, P. R. Unwin, J. V. Macpherson, and M. A. Beeston, in M.Shimono, T. Maeda, H. Suda, and K. Takahashi, eds., *Dentine/Pulp Complex*, Quintessence, Carol Stream IL, 1996, pp. 333–335.
42. S. Nugues and G. Denuault, *J. Electroanal. Chem.* **408**, 125–140 (1996).
43. E. Klusmann and J. W. Schultze, *Electrochim. Acta* **42**, 3,123–3,136 (1997).
44. J. O. Park, C. H. Paik, and R. C. Alkire, *J. Electrochem. Soc.* **143**, L174–L176 (1996).
45. C. Wei, A. J. Bard, G. Nagy, and K. Toth, *Anal. Chem.* **67**, 1,346–1,356 (1995).
46. N. J. Evans, M. Gonsalves, N. J. Gray, A. L. Barker, J. V. Macpherson, and P. R. Unwin, *Electrochem. Commun.* **2**, 201–206 (2000).
47. A. J. Bard and L. R. Faulkner, *Electrochemical Methods: Fundamentals and Applications*, J Wiley, NY, 2000.
48. J. M. Davis, F. -R. F. Fan, and A. J. Bard, *J. Electroanal. Chem.* **238**, 9–31 (1987).
49. M. V. Mirkin, F. R. F. Fan, and A. J. Bard, *J. Electroanal. Chem.* **328**, 47–62 (1992).
50. Q. Fulian, A. C. Fisher, and G. Denuault, *J. Phys. Chem. B* **103**, 4,387–4,392 (1999).
51. J. Galceran, J. Cecília, E. Companys, J. Salvador, and J. Puy, *J. Phys. Chem. B* **104**, 7,993–8,000 (2000).
52. J. L. Amphlett and G. Denuault, *J. Phys. Chem. B* **102**, 9,946–9,951 (1998).
53. J. L. Amphlett and G. Denuault, *J. Phys. Chem. B* **104**, 7,993–8,000 (2000).
54. L. A. Nagahara, T. Thundat, and S. M. Lindsey, *Rev. Sci. Instrum.* **60**, 3,128–3,130 (1989).
55. B. D. Penner, M. J. Heben, and N. S. Lewis, *Anal. Chem.* **61**, 1,630 (1989).
56. C. J. Slevin, N. J. Gray, J. V. Macpherson, M. A. Webb, and P. R. Unwin, *Electrochem Commun* **1**, 282–288 (1999).
57. Y. H. Shao, M. V. Mirkin, G. Fish, S. Kokotov, D. Palanker, and A. Lewis, *Anal. Chem.* **69**, 1,627–1,634 (1997).
58. J. V. Macpherson and P. R. Unwin, *J. Phys. Chem.* **99**, 3,338–3,351 (1995).
59. M. Buchler, S. C. Kelley, and W. H. Smyrl, *Electrochem. Solid-State Lett.* **3**, 35–38 (2000).
60. D. O. Wipf and A. J. Bard, *Anal. Chem.* **64**, 1,362–1,367 (1992).
61. C. M. Lee, D. O. Wipf, A. J. Bard, K. Bartels, and A. C. Bovik, *Anal. Chem.* **63**, 2,442–2,447 (1991).
62. K. A. Ellis, M. D. Pritzker, and T. Z. Fahidy, *Anal. Chem.* **67**, 4,500–4,507 (1995).
63. F. R. F. Fan and A. J. Bard, *Proc. Natl. Acad. Sci. U.S.A.* **96**, 14 222–14 227 (1999).
64. L. B. Anderson and C. N. Reilley, *J. Electroanal. Chem.* **10**, 295–305 (1965).
65. L. B. Anderson, B. McDuffie, and C. N. Reilley, *J. Electroanal. Chem.* **12**, 477–494 (1966).
66. M. V. Mirkin, T. C. Richards, and A. J. Bard, *J. Phys. Chem.* **97**, 7,672–7,677 (1993).
67. M. V. Mirkin and A. J. Bard, *J. Electroanal. Chem.* **323**, 29–51 (1992).
68. A. J. Bard, M. V. Mirkin, P. R. Unwin, and D. O. Wipf, *J. Phys. Chem.* **96**, 1,861–1,868 (1992).
69. A. J. Bard, F. -R. Fan, and M. Mirkin, in I. Rubenstein, ed., *Physical Electrochemistry: Principles, Methods, and Applications*, Marcel Dekker, NY, 1995, pp. 209–242.
70. J. L. Gilbert, S. M. Smith, and E. P. Lautenschlager, *J. Biomed. Mater. Res.* **27**, 1,357–1,366 (1993).
71. E. R. Scott, H. S. White, and J. B. Phipps, *Anal. Chem.* **65**, 1,537–1,545 (1993).
72. E. R. Scott, J. B. Phipps, and H. S. White, *J. Invest. Dermatol.* **104**, 142–145 (1995).
73. M. Gonsalves, A. L. Barker, J. V. Macpherson, P. R. Unwin, D. O'Hare, and C. P. Winlove, *Biophys J.* **78**, 1,578–1,588 (2000).
74. D. O. Wipf, F. Ge, T. W. Spaine, and J. E. Baur, *Anal. Chem.* **72**, 4,921–4,927 (2000).
75. K. Borgwarth and J. Heinze, *J. Electrochem. Soc.* **146**, 3,285–3,289 (1999).
76. W. E. Morf and N. F. Derooij, *Sensor Actuator A Phys.* **51**, 89–95 (1995).
77. *The Piezo Book*, Burleigh Instruments, Inc., Fishers, NY.
78. G. Wittstock, H. Emons, T. H. Ridgway, E. A. Blubaugh, and W. R. Heineman, *Anal. Chim. Acta* **298**, 285–302 (1994).
79. K. Toth, G. Nagy, C. Wei, and A. J. Bard, *Electroanalysis* **7**, 801–810 (1995).
80. B. R. Horrocks, M. V. Mirkin, D. T. Pierce, A. J. Bard, G. Nagy, and K. Toth, *Anal. Chem.* **65**, 1,213–1,224 (1993).
81. Y. H. Shao and M. V. Mirkin, *Anal. Chem.* **70**, 3,155–3,161 (1998).
82. T. Solomon and A. J. Bard, *J. Phys. Chem.* **99**, 17,487–17,489 (1995).
83. S. Amemiya, Z. Ding, J. Zhou, and A. J. Bard, *J. Electroanal. Chem.* **483**, 7–17 (2000).
84. D. O. Wipf, A. J. Bard, and D. E. Tallman, *Anal. Chem.* **65**, 1,373–1,377 (1993).
85. K. Borgwarth, D. G. Ebling, and J. Heinze, *Ber. Bunsenges. Phys. Chem.* **98**, 1,317–1,321 (1994).
86. K. Borgwarth, D. Ebling, and J. Heinze, *Electrochim. Acta* **40**, 1,455–1,460 (1995).
87. M. Ludwig, C. Kranz, W. Schuhmann, and H. E. Gaub, *Rev. Sci. Instrum.* **66**, 2,857–2,860 (1995).

88. P. I. James, L. F. Garfias-Mesias, P. J. Moyer, and W. H. Smyrl, *J. Electrochem. Soc.* **145**, L64–L66 (1998).
89. M. Küpper and J. W. Schultze, *Electrochim. Acta* **42**, 3,023–3,032 (1997).
90. J. O. Park, C. H. Paik, Y. H. Huang, and R. C. Alkire, *J. Electrochem. Soc.* **146**, 517–523 (1999).
91. J. O. Park and H. Böhni, *Electrochem. Solid-State Let.* **3**, 416–417 (2000).
92. N. Casillas, S. Charlebois, W. H. Smyrl, and H. S. White, *J. Electrochem. Soc.* **141**, 636–642 (1994).
93. L. F. Garfias-Mesias, M. Alodan, P. I. James, and W. H. Smyrl, *J. Electrochem Soc.* **145**, 2,005–2,010 (1998).
94. S. B. Basame and H. S. White, *Anal. Chem.* **71**, 3,166–3,170 (1999).
95. J. W. Still and D. O. Wipf, *J. Electrochem. Soc.* **144**, 2,657–2,665 (1997).
96. K. Fushimi and M. Seo, *Proc. Electrochem. Soc.* **99–28**, 362–369 (2000).
97. P. R. Unwin and J. V. Macpherson, *Chem. Soc. Rev.* **24**, 109–119 (1995).
98. J. V. Macpherson, M. A. Beeston, P. R. Unwin, N. P. Hughes, and D. Littlewood, *J. Chem. Soc. Faraday Trans.* **91**, 1,407–1,410 (1995).
99. J. V. Macpherson and P. R. Unwin, *J. Phys. Chem.* **98**, 1,1764–1,1770 (1994).
100. B. W. Mao, B. Ren, X. W. Cai, and L. H. Xiong, *J. Electroanal. Chem.* **394**, 155–160 (1995).
101. J. Zaumseil, G. Wittstock, S. Bahrs, and P. Steinrucke, *Fresnius' J. Anal. Chem.* **367**, 352–355 (2000).
102. S. Kasai, A. Yokota, H. Zhou, M. Nishizawa, K. Niwa, T. Onouchi, and T. Matsue, *Anal. Chem.* **72**, 5,761–5,765 (2000).
103. R. B. Jackson, M. Tsionsky, Z. G. Cardon, and A. J. Bard, *Plant Physiol.* **111**, 354–354 (1996).
104. T. Yasukawa, T. Kaya, and T. Matsue, *Chem. Lett.* 975–976 (1999).
105. B. D. Bath, H. S. White, and E. R. Scott, *Pharm. Res.* **17**, 471–475 (2000).
106. G. Denuault, M. H. T. Frank, and L. M. Peter, *Faraday Discuss. Chem. Soc.* 23–35 (1992).
107. M. H. T. Frank and G. Denuault, *J. Electroanal. Chem.* **354**, 331–339 (1993).
108. M. H. T. Frank and G. Denuault, *J. Electroanal. Chem.* **379**, 399–406 (1994).
109. Y. H. Shao, M. V. Mirkin, and J. F. Rusling, *J. Phys. Chem. B* **101**, 3,202–3,208 (1997).
110. J. Zhang and P. R. Unwin, *J. Phys. Chem. B* **104**, 2,341–2,347 (2000).
111. J. Zhang, A. L. Barker, and P. R. Unwin, *J. Electroanal. Chem.* **483**, 95–107 (2000).
112. A. L. Barker and P. R. Unwin, *J. Phys. Chem. B* **104**, 2,330–2,340 (2000).
113. C. J. Slevin, S. Ryley, D. J. Walton, and P. R. Unwin, *Langmuir* **14**, 5,331–5,334 (1998).
114. C. J. Slevin, J. V. Macpherson, and P. R. Unwin, *J. Phys. Chem. B* **101**, 1,0851–1,0859 (1997).
115. A. L. Barker, M. Gonsalves, J. V. Macpherson, C. J. Slevin, and P. R. Unwin, *Anal. Chim. Acta* **385**, 223–240 (1999).
116. O. E. Hüsser, D. H. Craston, and A. J. Bard, *J. Electrochem. Soc.* **136**, 3,222–3,229 (1989).

SILVER HALIDE DETECTOR TECHNOLOGY

RICHARD HAILSTONE
Rochester Institute of Technology
Rochester, NY

INTRODUCTION

Detectors based on silver halides are among the oldest light detectors available — both the daguerreotype and the calotype date to 1839 (1). The invention of the fixing process in 1841 provided a means of stabilizing the image (2), and the applications of photography exploded. It is quite interesting that the same basic materials being used for silver halide detectors in the twenty-first century are identical to those discovered in 1839. That is not to say that the technology has not been improved over the intervening years. There have been tremendous improvements in sensitivity, image quality, image permanence, the ability to record colors, and convenience. But, because the same basic material is still being used despite numerous attempts to discover "better" materials, one must conclude that silver halides have some very unique properties — many of which will be discussed in this article.

This article focuses on the basic aspects of the silver halide detector that are common to all of its commercial applications; specific applications are covered in other articles. A "top-down" approach is used. First, the detector and methods for characterizing it are described. Next, the preparation of the imaging elements in the detector are discussed. Then, a detailed description of the imaging elements is given, including their physical properties, the mechanism of latent-image formation, factors that affect the efficiency of image recording and the sensitivity of the detector, techniques for improving efficiency, and methods of adjusting its spectral response. A discussion follows of the signal amplification step required to achieve the final

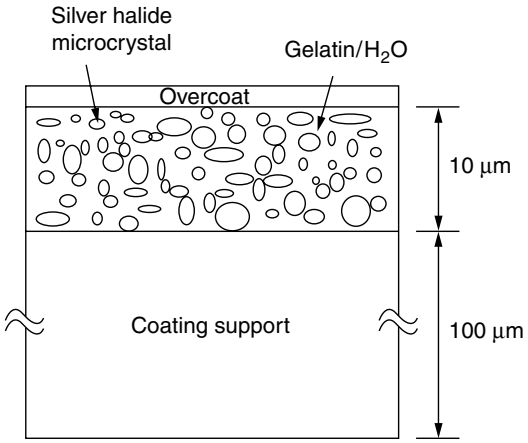


Figure 1. Schematic depiction of the cross section of a simple silver halide detector. The overcoat layer provides abrasion protection, as well as a means of introducing chemicals into the coating. Often, reagents that lead to chemical cross-linking of the gelatin molecules are introduced here. The "hardened" film is more robust, particularly to high-temperature image processing.

image. Finally, a brief discussion of design considerations is included that stresses the optimization of sensitivity, noise, and sharpness. A short discussion of imaging systems is also included. Several books and articles can be recommended to those who would like a more detailed discussion of silver halide detector technology, including applications (3–7). Several review articles focusing on different aspects are also available (8–11).

DETECTOR STRUCTURE AND FUNCTION

The detector consists of silver halide microcrystals coated in a gelatin/water mixture on either a transparent or reflective support, as exemplified by the schematic in Fig. 1. A latent image is formed on the microcrystal via a photochemical change induced by light absorption. The visible image is created through an image processing step in which those microcrystals that have received sufficient light "blacken" as they are transformed from silver halide to metallic silver. In color photography, the image processing stage produces a dye image, rather than a silver image.

The detector described in Fig. 1 is very simple. For some applications designed for black-and-white pictorial images, this structure might suffice. Usually, however, commercial products have a more complicated structure that involves several layers; some contain microcrystals, and some contain various chemicals to enhance the final image. This is particularly true for color films where upward of 15 layers are present. Manufacture of such complicated structures requires sophisticated coating technology to produce products consistently whose sensitivity and image quality are constant.

DETECTOR CHARACTERIZATION

Basic Concepts

The study of the characteristic response function of an imaging device or material is called "sensitometry" (12,13). The response function can be described schematically, as in Fig. 2. Here E is the exposure which is defined as the product of irradiance I and the exposure time t. A log scale is used for exposure because many photographic systems are designed to have dynamic ranges of two, three, or even four decades of exposure to capture images faithfully and conveniently. The response curve describes the input–output relationship in an imaging system and is an integral part of system design.

The response can be measured in many ways, but in silver halide systems, it is almost always the density of the developed image. There are two modes of measuring density: (1) transmission, which is used for transparent materials; and (2) reflection, which is used for reflective materials like prints. As shown in Fig. 3, transmission density D_t uses the material's transmittance T, whereas reflection density D_r uses the material's reflectance R. Note that I_o and I are related to the light used to view the image, not the light used to make the image. Density is used rather than transmittance or reflectance because the brightness perceived by the human visual system varies approximately linearly with the logarithm of the luminance of the object or image being viewed (14).

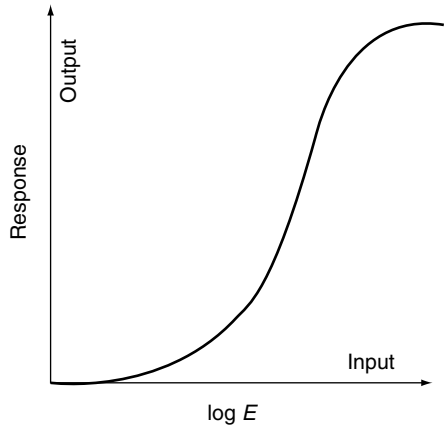


Figure 2. Input–output relationship for an imaging material.

The Characteristic Curve

The response or characteristic curve is often sigmoidal, as shown in Fig. 4. Also, as indicated, there are conventional labels for the different regions of the characteristic curve. The characteristic curve can have either a negative character, as shown at the top of the figure, which means that density increases as exposure increases, or a positive character as shown at the bottom of the figure, which means that density decreases as exposure increases. Systems that respond negatively are used as image capture materials, in which the image will often be subsequently transferred to another negative-type material for image display in a reflection mode (prints). Systems that respond positively are also used for image capture, but the intent is that the capture

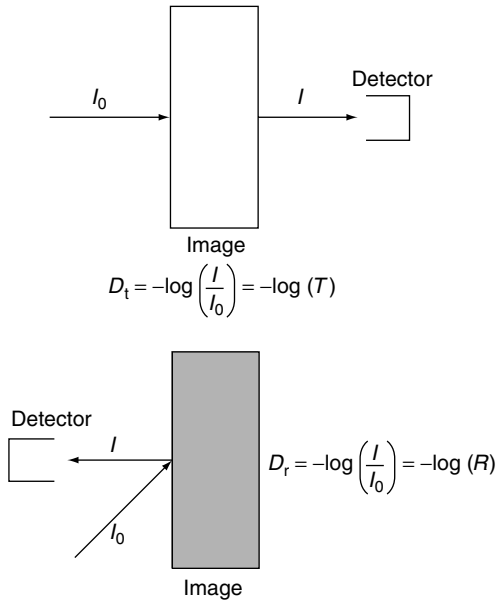


Figure 3. Schematic defining transmission (top) and reflection (bottom) density. The input irradiance is I_o, and the transmitted or reflected irradiance is I.

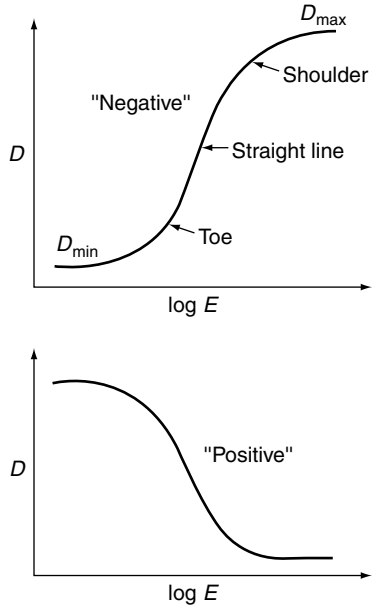


Figure 4. $D-\log E$ curves for negative- and positive-working materials. Curve labels indicate customary designations of the regions of the $D-\log E$ curve. Identical labels apply to the positive curve.

material becomes the display material in transmission mode (slides). The negative or positive nature of the characteristic curve of a given system is usually determined by the details of the image processing stage.

Sensitivity

Often, it is desired to summarize the behavior of a given imaging system in terms of a single number or index so that different systems can be more easily compared. The indexes that are most common are "sensitivity" or "speed" and "contrast." However, these indexes can be defined in many different ways. In some cases the speed is simply defined as the $\log E$ value required to achieve a certain density. But, in this method, a higher numerical value means that the sensitivity is lower, which can be confusing. Sensitivity defined in this way also depends on the reference density used, so that the relative speed of two systems whose characteristic curves have quite different shapes will be very different in the toe versus the shoulder of the curve.

Photographic manufacturers have long recognized these problems in defining sensitivity and, through the International Standards Organization (ISO), collectively designed a system called "ISO speed" for uniformly measuring sensitivity (15,16). In this system, sensitivity indexes for films are based on a reference density in the toe of the $D-\log E$ curve, as indicated in Fig. 5. In this case the ISO speed for noncolor materials is defined as

$$S_{\mathrm{ISO}} = \frac{0.8}{E_m} \tag{1}$$

where E_m is defined in Fig. 5. Notice that the $D-\log E$ curve in Fig. 5 does not go to zero density at very low

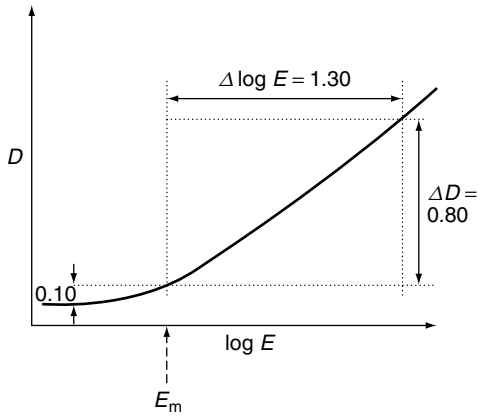


Figure 5. Definition of ISO speed for a black-and-white photographic material. Film must be processed to match the specified ΔD and $\Delta \log E$. The exposure E_{m} is used in calculating the ISO speed [Eq. (1)].

exposure, but asymptotically approaches some minimum density D_{min}, sometimes called "fog." This is often true in imaging systems, and it means that the system can produce some response, even in the absence of exposure. Fog is often an undesirable but unavoidable artifact of many systems. However, some fog can be tolerated in the negative because it can be compensated for in making the final print. But display materials such as prints must have very low fog levels.

Contrast

The contrast of an imaging system is related to how quickly or slowly the response, or density, changes with changes in the input, $\log E$. Because the characteristic curve is nonlinear, again we have several ways of measuring contrast or "gamma," as it is sometimes called. The instantaneous slope of a $D-\log E$ curve is defined as

$$\gamma = \frac{dD}{d(\log E)} \tag{2}$$

Obviously, then, the value of γ is a function of where it is measured on the $D-\log E$ curve, as illustrated in Fig. 6.

Often, one wants to define a "gamma index," or a "contrast index," to describe the overall steepness of the

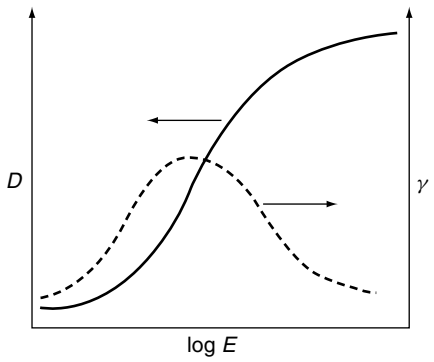


Figure 6. Variation of contrast γ with exposure.

curve. Just as for sensitivity, there are many ways to define the contrast index. One method is to use the maximum slope to define the system gamma. This is merely the maximum gamma in Fig. 6.

"Dynamic range" is sometimes used to indicate the contrast characteristics of a system. Dynamic range, also called "latitude," is loosely defined as the range of log E over which the density changes. Clearly, dynamic range and gamma are inversely related. A narrow dynamic range means that the user must be very careful with exposure, otherwise the detector will be under- or overexposed. Camera negative films are designed to have a wide dynamic range, or long latitude, so that the user will have a much easier time fitting the entire range of scene luminances onto the $D-\log E$ curve.

In many situations, we are interested in the way the silver halide responds at different wavelengths. It is common to condense this information into plots of sensitivity and gamma as a function of wavelength, $S(\lambda)$ and $\gamma(\lambda)$. The $S(\lambda)$ versus λ plot is called a "spectral sensitivity" curve. It is common to correct the relative sensitivities for wavelength-dependent irradiance differences at the exposure plane, so that perceived sensitivity differences are truly related to the material and are not influenced by the particular light source used.

Experimental Aspects of Sensitometry

Sensitometry is studied by using devices called "sensitometers." These devices must be capable of controlling the exposure. This may be done by varying both I and t, as in flash lamps, by varying t and holding I constant, or by varying I and holding t constant. The latter is the most common technique. One often uses a "step tablet" or "step wedge" placed over the imaging material. The step tablet has a series of steps of differing density that absorb (or reflect) differing amounts of the incident light. The step tablet is designed to decrement the irradiance by equal factors, that is, a geometric not an arithmetic progression.

MICROCRYSTAL PREPARATION

The microcrystals, also called "grains," are prepared in a variety of shapes and size distributions to accommodate different applications. Strictly speaking, the resulting suspension of silver halide microcrystals in aqueous gelatin is classified as a dispersion, but early in the development of photography, it was called an "emulsion." We will continue this practice in this article. Across the range of applications, the microcrystal size varies from a few tenths of a micron to several microns in diameter. Sensitivity, or speed, increases with increasing grain size, so higher speed films must necessarily use larger grains. Materials intended for image capture need to produce a distinguishable output over a wide range of inputs, and so a polydisperse size distribution is used. On the other hand, materials for image display, such as prints, need to respond only over a relatively narrow range of inputs, so a monodisperse size distribution is often used. Figure 7 contains electron micrographs of typical grains used as imaging elements in silver halide detectors. In some cases, the grains are relatively flat ("tabular grains"); in other cases, the grains are more three-dimensional.

Silver halide microcrystals are prepared via a precipitation reaction involving a soluble silver salt and a soluble halide salt to produce a silver halide precipitate. There are various ways by which the reactants can be introduced into the system; two basic schemes are shown in Fig. 8. The single-jet scheme has historical significance, but the double-jet scheme has largely replaced it in current precipitation modes. Note that in both schemes, as in all precipitation modes, gelatin is present. In the absence of gelatin or some suitable synthetic polymer, coagulation of the growing silver halide microcrystals would occur. As will be described later, the technology permits remarkable control over microcrystal shape and its size distribution, which is required for the various applications of silver halide detectors.

Crystal Structure and Shape

Silver chloride (AgCl), silver bromide (AgBr), and silver iodide (AgI) exist as crystals composed of silver (Ag^+) and halide (Cl^-, Br^-, or I^-) ions. Pure AgI is not used in photographic materials, so it will not be discussed further. The crystal structure of AgCl and AgBr is face-centered cubic (fcc) which is the "rock salt" structure of sodium chloride, as shown in Fig. 9. However, unlike sodium chloride where the crystal's cohesive energy is entirely due to ionic bonding, the cohesive energy in silver halides is partly due to covalent bonding in addition to substantial amounts of ionic bonding (18). However, an alternative view favors van der Waals forces rather than covalent bonding (19).

Looking at the fcc lattice structure, one can distinguish many different planes of ions, and, in principle, any of these planes could be used to terminate the crystal to form a wide variety of crystal shapes. However, only two morphologies, cubic and octahedral, are sufficiently stable for practical use. The Miller index for the lattice plane leading to cubes is (100), where both ion types are present in an alternating pattern, leading to octahedra is (111), where all of the ions are of one type. Other crystal shapes would require surface planes that have higher Miller indexes. But such planes would show significant deviations from planarity, leading to high surface energies. As a result, they would reconstruct themselves to achieve a lower Miller index and thereby lower the surface energy.

Particle Stability

One critical feature of silver halides is their extreme insolubility in water, which makes it possible to grow microcrystals that are stable in an aqueous medium. However, silver halides in contact with water do release some ions into solution and this is symbolized for AgBr by the following equilibrium:

$$AgBr \rightleftharpoons Ag^+ + Br^- \tag{3}$$

Expressing this equilibrium mathematically,

$$K_{sp} = \frac{[Ag^+][Br^-]}{[AgBr_{solid}]} = [Ag^+][Br^-] \tag{4}$$

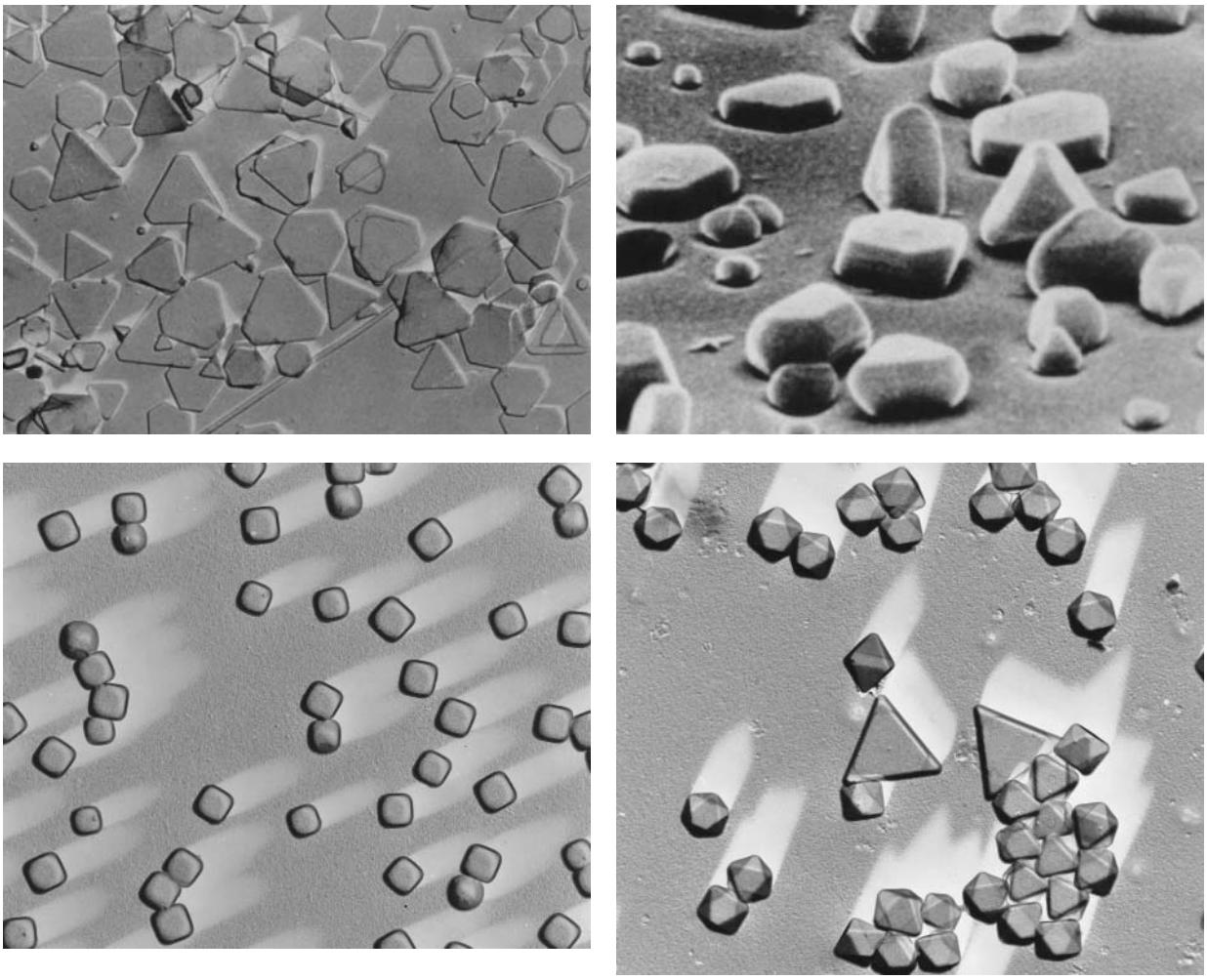


Figure 7. Electron micrographs of carbon replicas of emulsion grains. The "shadows" are obtained from metal evaporation at an angle to the plane of the sample and can be used to assess the thickness of the grain.

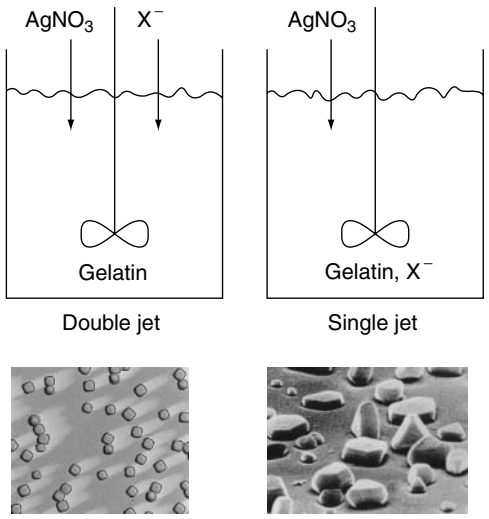


Figure 8. Basic schemes for precipitating silver halide microcrystals and a view of the particle distribution produced. X^- represents a halide ion source, typically the potassium or sodium salt.

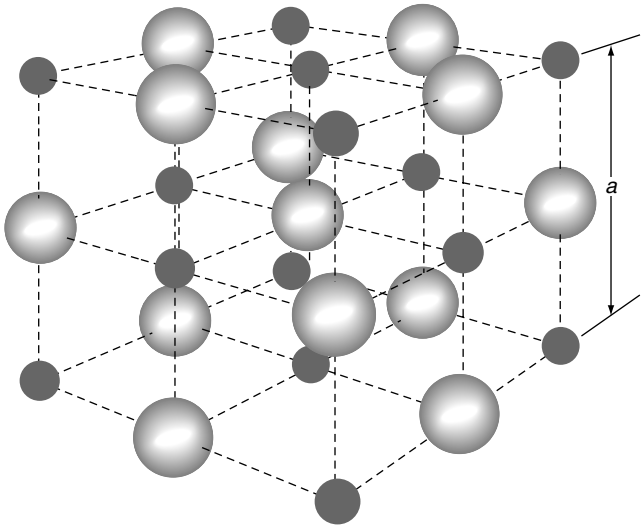


Figure 9. Lattice ion arrangement in a face-centered cubic crystal. Small filled circles are Ag^+ and large open circles are Br^-. a is the dimension of the unit cell (from Ref. 17).

where the simplification on the right-hand side recognizes that a molar concentration cannot be specified for a solid material, so $[AgBr_{solid}] = 1$. K_{sp} is called the "solubility product." Some typical values are given in Table 1. These K_{sp} values indicate that the equilibrium in reaction (3) lies far to the left. For example, a saturated AgCl solution at room temperature has silver and chloride ion concentrations of approximately 10 μM.

The silver ion concentration plays a major role in material preparation and performance. This is expressed in compact notation analogous to pH as

$$pAg = -\log[Ag^+] \tag{5}$$

Similar equations exist for pBr and pK_{sp} so that Eq. (4) can be transformed into

$$pK_{sp} = pAg + pBr \tag{6}$$

The ion concentration can be electrochemically determined by using a silver/silver chloride electrode and a reference electrode. Using the Nernst equation (20), one can derive the following equation for room temperature measurements:

$$E_{Ag} = E^o_{Ag} + 0.059pAg \tag{7}$$

where E_{Ag} is the measured potential and E^o_{Ag} is the standard potential for the reaction $Ag^+ + e^- \rightarrow Ag$. Thus, the pAg can be calculated from the measured potential by using the standard potential.

From reaction (3), we expect that when we place AgBr in water, the only ionic species in solution, besides the usual H^+ and OH^-, should be Ag^+ and Br^-. Likewise, these two ions should be at the same concentration or "equivalence point," which can be defined as $pAg = pBr = (pK_{sp})/2$. If we add NaBr to this solution, we expect the Ag^+ concentration to decrease and the Br^- concentration to increase in accordance with the equilibrium description in reaction (3). This is only partially true because Ag^+ can form soluble complexes with Br^- described by the following equilibrium:

$$Ag^+ + mBr^- \rightleftharpoons AgBr_m^{-(m-1)} \tag{8}$$

Before adding the NaBr, the system was at the equivalence point where equal numbers of Ag^+ and Br^- are present. After adding the NaBr the system is now said to be "on the excess bromide side." A similar situation would occur if we had added $AgNO_3$, leading to the excess silver side. However, because the latter can lead to the chemical formation of silver clusters (fog), silver halide materials for photographic use are always prepared, handled, stored, and coated on the excess bromide side.

Table 1. Solubility Products of Silver Halides

Temperature (°C)	AgCl	AgBr	AgI
25	1.77×10^{-10}	4.85×10^{-13}	8.28×10^{-17}
50	1.30×10^{-9}	6.45×10^{-12}	2.49×10^{-15}

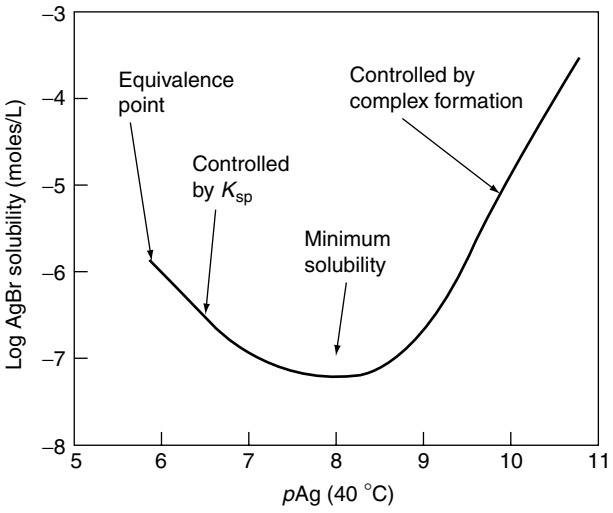


Figure 10. Solubility vs pAg relationship for AgBr at 40 °C.

The existence of soluble complexes leads to nonlinear behavior in the dependence of solubility on silver ion concentration. Solubility is defined as the total concentration of silver-containing species in solution. Its dependence on pAg for AgBr is shown in Fig. 10. Starting at the equivalence point (pAg 5.8), the solubility first decreases in accordance with the equilibrium expressed in reaction (3) because the Br^- concentration is increasing when pAg is increasing [see Eq. (6)]. Then, a minimum solubility point is reached. Finally, the solubility increases as soluble silver complexes begin to form [reaction (8)]. Because the equilibria expressed by reactions (3) and (8) are shifted to the right by increasing temperature, the solubility also increases with temperature, and the minimum solubility point shifts toward lower pAg, and increases in value (21).

To stabilize the silver halide microcrystals in aqueous solution and to prevent coalescence, all silver halide materials must be prepared in the presence of a stabilizing polymer. This role is most often filled by gelatin. Gelatin is a biopolymer prepared from animal collagen. It is a linear polymer composed of covalently linked amino acids that provide amide and carboxylic acid side groups which confer a pH-dependent charge on the polymer. The side groups also provide sites for cross-linking that lead to a three-dimensional polymer of much greater mechanical strength. This latter process, termed "hardening," is carried out during the coating operation in all commercial photographic materials, usually by introducing a suitable cross-linking reagent into the overcoat (see Fig. 1).

Imaging systems based on silver halide technology must possess a high degree of stability. The actual exposure may occur months to years after the system has been coated, and once exposed, the image processing stage may not occur for months to years later. Thus, both the silver halide and the photoproduced silver clusters must be stable for very long time periods. To this end, stabilizing agents are incorporated into the coating (22). The common feature

of these stabilizers is that they form complexes with silver ions, often with K_{sp}'s smaller than that achieved with halide ions. Thus, the silver ion activity is reduced in the coating, minimizing unwanted chemical changes during storage. These stabilizers can also adsorb on the photoproduced silver clusters and increase their stability.

The Gibbs–Thomson effect (23), which deals with nucleation of new phases, for example, formation of rain drops in clouds, is an important phenomenon in emulsion precipitation. This effect describes how the particle radius and the supersaturation affect the system's ability to change its size distribution. A thermodynamic description uses the change in the excess free energy ΔG expressed as

$$\Delta G = 4\pi r^2 \sigma_s - \frac{4\pi r^3}{3V} kT \ln S \tag{9}$$

where σ_s is the interfacial free energy, V is the atomic volume, k is Boltzmann's constant, and T is the absolute temperature. The supersaturation is defined as

$$S = \frac{C_s}{C_e} \tag{10}$$

where C_s is the concentration of silver containing species in solution and C_e is the concentration of silver-containing species that would prevail at equilibrium. When ΔG is positive, there is a strong driving force for the system to move to a more stable state by changing the particle radius until $\Delta G \leq 0$.

In emulsion precipitation, the most important variable is supersaturation, as schematically illustrated in Fig. 11. This figure shows how the excess free energy of the system varies with particle radius and supersaturation. Particles whose radius would place them to the left of the maximum on a given curve will be unstable because by becoming smaller, their excess free energy will be reduced. Particles whose radii place them to the right of the maximum will grow in size because this will also reduce the excess free energy. As shown in the figure, the position of the curve's maximum with respect to the critical particle size r_c varies inversely with the degree of supersaturation (23).

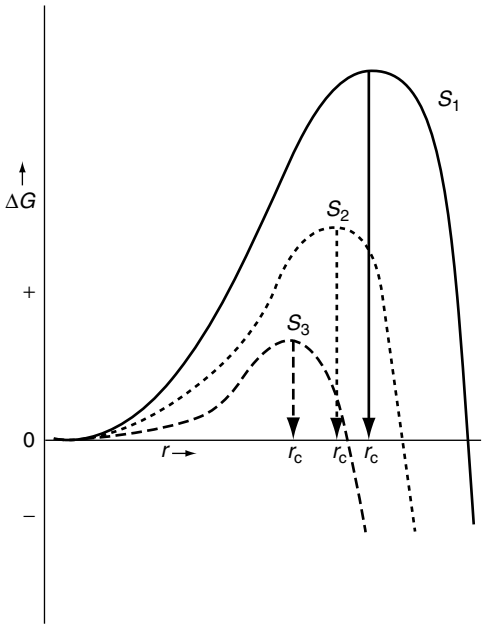


Figure 11. Relationship between the system excess free energy ΔG and the particle size r for different supersaturated conditions S. $S_3 > S_2 > S_1$. r_c is the critical radius for particle stability.

Ostwald ripening (24), which follows directly from the Gibbs–Thomson effect, is a phenomenon that plays an important role in emulsion precipitation. In Ostwald ripening, small particles dissolve, and the material released is deposited on larger particles. The rate of Ostwald ripening is a function of the size difference between particles, growth and dissolution rates, and the transport properties through the solution. Dissolution rates are controlled by the temperature, the concentration of excess halide, and the presence of any silver ion complexing agents. Emulsions whose particle size distribution is narrow will broaden their size distribution when placed in an environment that favors Ostwald ripening (24).

Phases in Crystal Formation

Emulsion precipitation can be partitioned into a nucleation and a growth phase. Nucleation is the process whereby new crystals of extremely small size are formed. The driving force is a high level of supersaturation that encourages the formation of a new solid phase from the soluble species. Just how small a stable crystal is formed has been a question for many years and undoubtedly will be a function of the precipitation conditions. Earlier studies have concluded that the diameter of particles after just the first few seconds of precipitation is on the order of several hundred angstroms (25–28). More recently, multichannel UV spectrophotometry has been used to characterize particle size (29). This technique has millisecond resolution as opposed to the second resolution of earlier methods, and the results indicate that three to four $AgBr_m^{-(m-1)}$ molecules take part in forming ultrafine particles. Depending on the pAg, m has integral values between 2 and 4. Ostwald ripening of these ultrafine particles occurred on the time scale of seconds to minutes. Thus, most of the ultrafine particles will dissolve and grow onto the few that survive.

Growth is the addition of silver and halide ions to a growing crystal. The usual observation is that the population of grains grows at more or less the same rate and maintains the same width in their size distribution during precipitation (30). There are several mechanisms by which crystals can grow when immersed in a solution of their components. For silver halide precipitation, the most common mode of growth is by diffusion from solution and this becomes the rate-determining step when new material is rapidly incorporated into the growing crystal's surface (28,31). When the diffusion layer thickness is much greater than the crystal size, the growth rate G is given by

$$G = \frac{4D}{\rho L}(C - C_i) \tag{11}$$

where D is the diffusivity, ρ is the crystal density, L is the crystal diameter, C is the bulk solute concentration, and

C_i is the solute concentration at the surface. Because of its rapid incorporation into the crystal, the concentration of new material at the surface is always less than that in the bulk solution and provides a concentration gradient that acts as a driving force for growth.

Precipitation Scheme

Figure 12 shows the basic arrangement of a double-jet precipitation. The reactants are typically at 1 to 4 M concentrations, so that the supersaturation is very high just at the point of emergence from the supply tubes near the mixing head. For example, at 70 °C and *p*Ag 7.5, the equilibrium solubility of silver ion species is 10^{-6} M. If the $AgNO_3$ concentration in the supply tube is 1 M, then the supersaturation is 10^6, which provides a large driving force for crystal formation. Thus, transient nuclei are continuously formed in this highly supersaturated region of the reactor and then are fed into the bulk solution by the mixing head, where their survival will depend on the prevailing level of supersaturation (24). In the early stage (less than one minute), the bulk supersaturation is high, and the nuclei survive. These initial nuclei grow rapidly and reduce the bulk supersaturation so that additional nuclei cannot survive. Newly generated nuclei from the mixing head region will dissolve by Ostwald ripening and provide growth material for the crystals initially formed. Nucleation ceases when the rate of Ostwald ripening equals the rate of reactant addition.

A model has been developed to understand quantitatively the parameters that control the outcome of the precipitation (32). Figure 13 shows a qualitative view of the precipitation. In the quasi-steady-state region a dynamic mass balance between material added and the total mass of the microcrystals formed leads to expressions for the number of stable growing microcrystals. In the usual diffusion-controlled growth mode, the number of

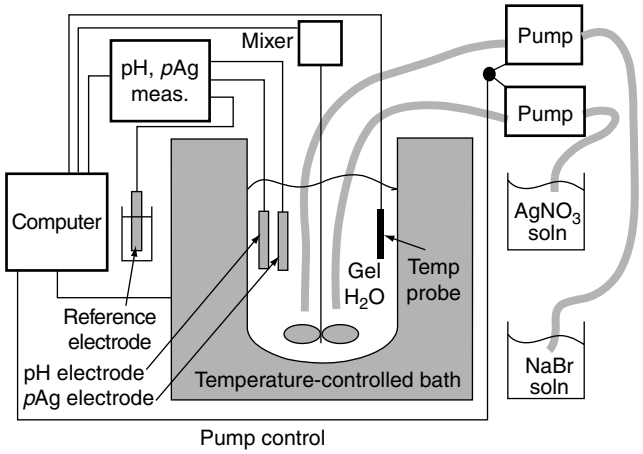


Figure 12. Diagram of an emulsion precipitation apparatus based on the double-jet scheme. The computer-controlled pumps allow the reactants to enter the reaction zone near the mixing head at the proper rate for microcrystal size control. The *p*Ag electrode signal is fed to the computer and is part of a feedback loop for adjusting the bromide pump speed to maintain the desired *p*Ag.

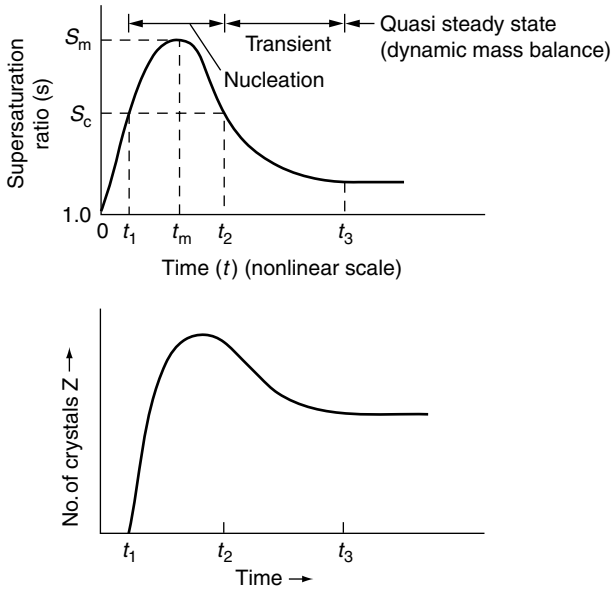


Figure 13. Qualitative representation of supersaturation ratio [Eq. (10)] and concentration of microcrystals vs. precipitation time. Axes are nonlinear. (From Ref. 32. Reprinted with permission of the American Chemical Society.)

growing microcrystals Z is given by

$$Z = \frac{ART}{C_s\left(\frac{\bar{r}}{r_c} - 1\right)} \tag{12}$$

where R is the reactant addition rate, $\bar{r}$ is the number average crystal radius, r_c is the critical radius below which the crystal will tend to dissolve due to the Gibbs–Thomson effect, and A is a constant. This equation accounts for the observed effects of reactant addition rate, *p*Ag, solubility, and temperature on the number of microcrystals formed.

In many applications of silver halide technology, the grains contain more than one type of halide ion. In some cases, the second halide ion is added at a selected point during the precipitation, or even after all the first halide is added. In such situations, recrystallization will occur, and the system will tend toward a uniform composition, but cannot usually achieve it due to kinetic factors (24). The driving force for recrystallization is increased entropy, which is fundamentally different from that of Ostwald ripening.

Twinning

In earlier applications of silver halide technology, the grains had a more or less three-dimensional structure. But in more recent applications, tabular grains have been used. To form tabular grains, a phenomenon called "twinning" must occur.

To understand twinning, we must first discuss how ions are placed within an individual layer during crystal growth. Consider a crystal growing by successive (111) layers — one layer of silver ions, the next bromide ions, the next layer silver ions, etc. Figure 14 shows the position of the ions in the first layer A. All of the ions in the next layer

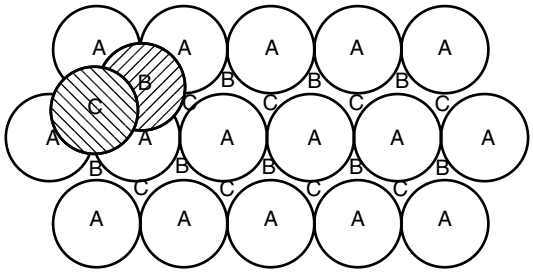


Figure 14. Layer arrangement for successive (111) planes. The bottom layer is denoted by A. The second layer is placed at the B position, the third layer at the C position, and the fourth layer would adopt the A position to start the sequence again.

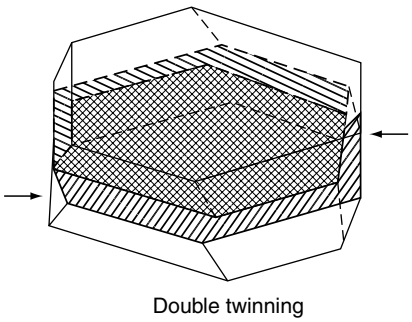


Figure 16. A view of an hexagonal tabular grain that shows four of the troughs (arrows) created at the edges by double twinning. The other two troughs are hidden in the rear of the grain.

can go in either position B or position C. Let us assume it is B. Then, the ions in the third layer would have to go into position C. The fourth layer would repeat the pattern where the ions go directly over those in position A in the first layer. Because there are two different ions in a growing AgBr crystal, we can represent one type of ion by an upper case letter, and the other type by a lower case letter. Then, the stacking pattern described above could be expressed as AbCaBcAbC...

Twinning consists of reversal of the usual stacking sequence within the crystal, as shown in Fig. 15 for twinning on a (111) plane, which is the most common application in silver halide technology. The twinned portion may be thought of as the mirror image of the parent crystal. In practice, twinning often occurs when the precipitation is carried out at $p\mathrm{Ag} > 10$. A possible explanation of twinning suggests that occasionally a Br^- (111) layer is added in the wrong (twin) position due to the predominance of complex ions such as $AgBr_3^{2-}$, (33). Other mechanisms that involve supersaturation-induced dominance of a surface integration growth mechanism (34) or coalescence of colloidal nuclei (35) have also been suggested to explain twinning.

Tabular morphology requires that at least two twinning events occur in a microcrystal. In a hexagonal tabular grain, double twinning would result in six troughs, as shown in Fig. 16. It is thought that addition of new material to these trough regions occurs much more readily than addition to the large flat faces and results in much higher growth rates in the lateral direction (36). As a result, grains with a high aspect ratio (diameter divided by the thickness) are formed. The photographic advantages of tabular morphology will be discussed later.

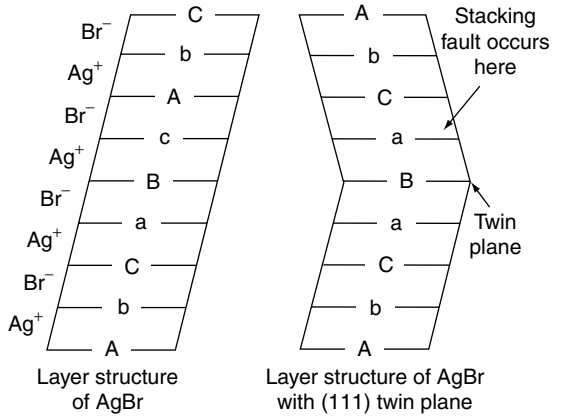


Figure 15. Normal stacking sequence of (111) planes (left) and stacking fault (right) in AgBr crystals. The stacking fault leads to the formation of a twin plane.

Controlling Crystal Shape and Size

As already mentioned, a variety of crystal shapes are used in silver halide imaging. The shape is controlled primarily by $p\mathrm{Ag}$ during precipitation (37). Cubic grains are formed at very low $[Br^-]$ (10^{-3} to 10^{-4} M), octahedra are formed at somewhat higher concentrations (10^{-2} M), and tabular grains are formed at very high concentrations (10^{-1} M). The explanation for the $p\mathrm{Ag}$ effect on shape is based on the silver-containing species in solution at the prevailing $p\mathrm{Ag}$. It is known that Br^- is more strongly adsorbed to (111) faces than to (100) faces (38,39). At relatively high $p\mathrm{Ag}$, say 9, the predominant form of soluble silver is $AgBr_3^{2-}$. As more Br^- is adsorbed to the (111) faces, these faces will repel the negatively charged complexes and therefore will grow more slowly than the (100) faces. The (100) faces will grow themselves out of existence, leaving only (111) faces (octahedra). At even higher $p\mathrm{Ag}$, twining will occur, and tabular grains are formed as discussed before.

Other crystal shapes are possible for the face-centered cubic system (40) but, as discussed before, they are unstable and rapidly convert to cubes, octahedra, or cubooctahedra, depending on the $p\mathrm{Ag}$. Growth modifiers can be used to stabilize these other shapes (40), but such compounds are not practical because they interfere with subsequent sensitization steps. There are two basic classes of growth modifiers. The first class is known as "restrainers" because they slow crystal growth, sometimes preferentially on a particular face, as just mentioned. They adsorb on the crystal surface and typically form complexes with silver ions that have very small K_{sp} values.

The second class of growth modifier is known as "ripeners." These form soluble complexes with silver ions and are primarily used to increase grain size. We see from Eq. (12) that if the solubility of the system is increased, then the number of growing crystals is decreased. For a fixed mass of silver halide added during the precipitation, each grain will be larger if the solubility is increased by

using a ripener. Studies of ripeners have shown that they, indeed, decrease the number of growing crystals (41).

In addition to ripeners, other parameters of the emulsion precipitation can be used to adjust grain size. We see by referring again to Eq. (12), that increasing the reactant addition rate will increase the number of crystals and thereby decrease grain size for a fixed mass of silver halide added. Factors that affect solubility such as *p*Ag or halide type will also affect the number of crystals and ultimately the final grain size. Many of these variables will also affect crystal shape (42,43).

One obvious method for increasing grain size is increasing the amount of material added. However, in a scheme of constant reactant flow rate, the grain volume increases linearly with time, so that the edge length increases as the cube root of time. Therefore, doubling the edge length will require an increase in precipitation time by a factor of 8. Under practical conditions, an upper limit to grain diameter is about 0.4 μm. Larger grain sizes can be achieved in several ways. One is to use a seeding technique in which the reactor is charged with small seeds that are then grown to a desired size. Another method is to use ripeners, as mentioned before. A third technique is to use accelerated flow rates. For example, if the flow rates are changed continuously, so that a constant flow rate per grain surface area is maintained, grain diameter will vary linearly with time.

Emulsion Washing

At the end of the precipitation, the reactor contains the desired emulsion grains but also a large concentration of counterions, usually NO_3^- and Na^+. If these counterions are not removed, the emulsion will gradually dissolve. To stabilize the emulsion, it is necessary to remove these counterions. Several techniques are possible. One technique is called "flocculation washing." In this procedure, modified gelatin is added to the emulsion, making the solubility pH-dependent. By lowering the pH, the grains and associated gelatin flocculate and settle to the bottom of the reactor, whereupon the counterions in the supernatant can be readily removed. After two or three cycles, the emulsion is essentially free of counterions. A variant of this technique uses inert salts instead of modified gelatin to increase the ionic strength until flocculation occurs, and the counterions can be removed by decanting.

Another method common in industrial operations is ultrafiltration. In this technique, the emulsion is pumped through a membrane cell. The membrane pore size is such that the counterions can be transmitted, but the grains and gelatin cannot. Circulation of the emulsion is stopped when the desired conductivity is reached, indicating that the ion concentration has been reduced to the desired level.

Emulsion Characterization

Once the precipitation is complete and the counterions are removed, it is necessary to characterize the emulsion to determine if the desired properties were achieved. To determine the morphology of the grains, they can be examined by a scanning electron microscope or an optical microscope if the grains are large enough. Next, the grain size must be determined. Grains in calibrated micrographs can be measured. There are also a variety of particle size distribution methods that can be employed for this purpose. It is also possible to use electrolytic reduction of the grains and, by counting the number of electrons required to reduce the grains, the volume can be calculated (44,45).

It is also useful to know the thickness of tabular grains. Optical micrographs can sometimes be used for this purpose qualitatively because the grains are thin enough to cause interference effects, so that different grain thicknesses will have different colors. Quantitative measurements can be made by obtaining SEM micrographs from tilted images. Another possibility is using shadowed micrographs from a transmission electron microscope (see Fig. 7). Knowing the length of the shadow and the angle at which the shadowing was done, it is possible to calculate the thickness. However, for very thin grains, the thickness of any adsorbed gelatin layer can obscure the true thickness.

The composition of the grains is another property of interest. There are several bulk methods for characterizing the grain as a whole such as neutron activation analysis, X-ray powder diffraction, spark source mass spectrometry, atomic absorption, and plasma emission. Each technique has its own limitations in detection limits and sample preparation. In some cases, it is useful to do compositional analysis in different parts of the grains to test for homogeneity or more usually to see if a desired heterogeneity has been achieved. Techniques here include analytical electron microscopy coupled with cross-sectioning techniques (46), luminescence microscopy at low temperatures (47,48), and secondary ion mass spectrometry (49,50).

PHYSICS OF THE SILVER HALIDE MICROCRYSTAL

Crystal Defects

Similar to other crystalline materials, silver halides can possess crystal defects. One defect of importance photographically is the dislocation. An edge dislocation occurs when a partial plane of ions is missing from the crystal, creating a line imperfection (17). Dislocations can be introduced during crystal growth or can be induced by applying pressure to a perfect crystal. The defect induces inefficiencies in the image formation stage of the photographic process because it leads to internal latent-image formation, which is inaccessible to the developer during image processing. Thus, great care is taken to avoid this defect.

Other crystal defects are known as point defects because they involve just one of the lattice ions, rather than a plane of ions. An example is the formation of an interstitial silver ion, as shown in Fig. 17. Note that a complementary species, a vacancy, is also formed and is assigned a negative charge to take into account the surrounding six Br^- ions. The process is reversible and allows interstitials to recombine with vacancies to re-form lattice silver ions. Thus, an equilibrium is set up between lattice silver ions

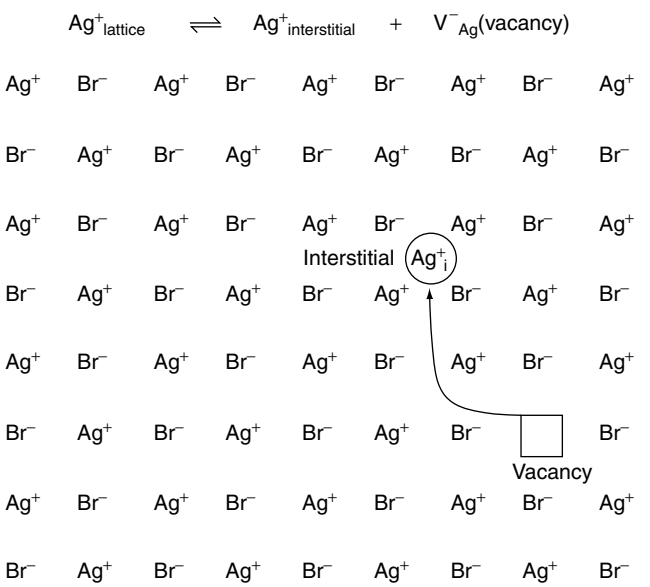


Figure 17. Two-dimensional view of a silver halide crystal showing the formation of an interstitial silver ion and a vacancy. Also shown is the equilibrium between lattice silver ions and interstitials and vacancies. Once created, the interstitial migrates throughout the crystal, driven by thermal energy.

and interstitial silver ions and vacancies. Such a defect is also known as a Frenkel disorder, after one of the pioneers in studying them. The energy requirement for such a process is very low, so it is quite favorable in AgBr, but less so in AgCl.

As will be discussed in detail later, much of the image formation and detection occurs at the surface of the silver halide microcrystals making the structure of the surface very important. A smooth surface plane would not be realizable under practical crystal growth and storage conditions. Rather, a more realistic view of the (100) crystal surface consists of steps and terraces and jogs along the steps. An example of these jogs, known as "kink" sites, is shown in Fig. 18. Because of the ionic nature of silver halides, the kink site has a partial charge associated with it because it is missing three of its normal six neighbors of opposite sign. The kink site illustrated in Fig. 18 has a formal charge of $+0.5\,e$, where e is the electrostatic unit of charge, and is known as a "positive-kink site." By interchanging the positive and negative charges in Fig. 18, one obtains a negative-kink site whose formal charge is $-0.5\,e$. However, the ions around these sites of excess charge can move to some extent to accommodate these excess charges and thereby reduce the charge to less than the indicated formal charge. This movement by the lattice ions is called "lattice relaxation." These sites of partial charge will take on importance when we discuss image formation later.

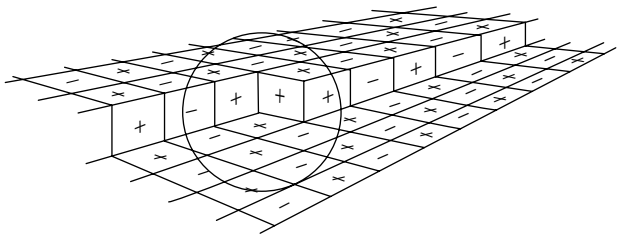


Figure 18. Illustration of a positive-kink site on a AgBr (100) surface.

A microcrystal bounded by (111) planes would be very unstable. It would have a large excess of charge because all of the surface ions would be the same; they carry uncompensated charge because, like kink sites on (100) surfaces, they are missing three of their normal oppositely charged neighbors. As a result, the crystal surface reconstructs to achieve a lower energy configuration. The exact nature of the reconstructed surface is not known. Models of the surface suggest a partial outermost layer of the ions of the same type where ions form either a hexagonal (51) or alternate row (52–54) pattern, as illustrated in Fig. 19. Some of these surface ions may be missing, thus forming "kink-like" sites.

Ionic Properties

The interstitial silver ion discussed previously is not a static entity but can repeatedly move from one interstitial position to another because the free energy requirement for such a process is very low. On the other hand, the complementary vacancy formed with the interstitial ion has a large free energy requirement for motion that reduces its mobility at room temperature by several orders of magnitude relative to the interstitial silver ion. The movement of interstitial silver ions and vacancies through silver halide crystals corresponds to a movement of charge and imparts electrical conductivity to the material. Because silver halides are electrical insulators in the dark, measurements of conductivity under such conditions relate to the ionic properties of the material.

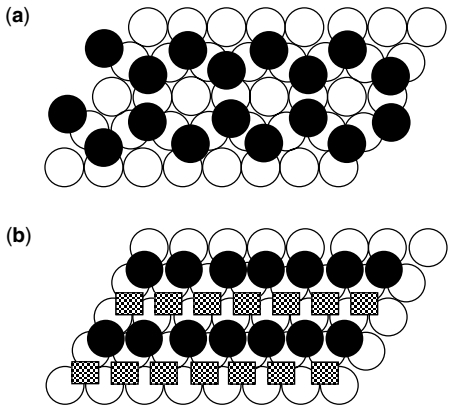


Figure 19. Two possible models of a reconstructed AgBr (111) surface. (**a**) hexagonal model. (**b**) alternate-row model. Shaded symbols are the top layer; open symbols are the second layer. The shaded circles can be either Ag^+ or Br^-, and the crosshatched squares are ion vacancies. (From Ref. 54. Reprinted with permission of The Society for Imaging Science and Technology, sole copyright owners of the *Journal of Imaging Science and Technology*.)

Ionic conductivity σ is determined by the product of mobility μ and concentration n and is expressed as

$$\sigma = e(\mu_i n_i + \mu_v n_v) \tag{13}$$

where the subscripts i and v refer to the interstitial and vacancy, respectively. At equilibrium,

$$K_F = n_i n_v = 2N^2 \exp\left(\frac{\Delta G_F}{kT}\right) \tag{14}$$

where K_F is the Frenkel equilibrium constant, N is the number of lattice sites, ΔG_F is the free energy of formation of the Frenkel pair, k is Boltzmann's constant, and T is the temperature in absolute units. Table 2 summarizes the thermodynamic data related to Frenkel pair formation in AgCl and AgBr which can be used to calculate concentrations of defects, also given in Table 2. The mobility part of the ionic conductivity can be calculated as

$$\mu = \frac{el^2 v}{6kT} \exp\left(\frac{\Delta G_\mu}{kT}\right) \tag{15}$$

where l is the jump distance, ν is the vibrational frequency of the defect, and ΔG_μ is the free energy of the jump. Table 3 summarizes the thermodynamics of the vacancy and interstitial jump, as well as the calculated mobilities.

A model consistent with experimental data has been developed that describes the mechanism by which the interstitial and vacancy move (55–57). Figure 20 illustrates that both the interstitial silver ion and the lattice ion that it will eventually replace move in a concerted process. The lattice silver ion moves into an interstitial position at the same time that the interstitial silver ion is moving into the lattice ion position. The energetics involved in the movement of both ions is the same. Key to the operation of this mechanism is the ability of the interstitial silver ion to squeeze through the triad of bromide ions (crosshatched ions in Fig. 20). The actual effective radii of the bromide ions are much larger than those illustrated in Fig. 20 to the point that they are almost touching each other. The remaining space for the silver ion to squeeze through the bromides is smaller than the free silver ion. What saves the day is that the silver ion can rather easily undergo quadrupolar deformation to form an ellipsoidal ion that can squeeze through the available space between the bromide triad.

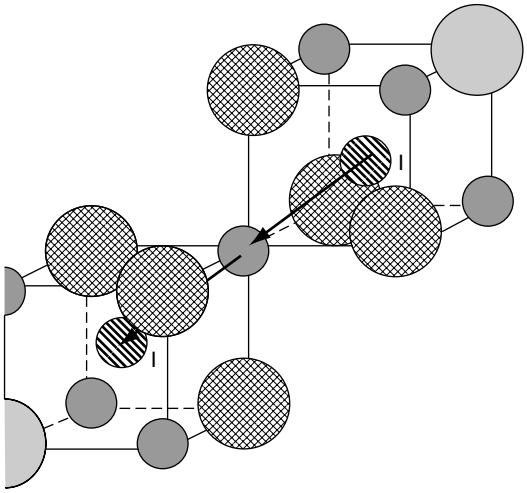


Figure 20. Schematic of the movement of an interstitial silver ion in a AgBr crystal. The ions are not drawn to scale. The interstitial silver ion I displaces a lattice silver ion into another interstitial position, taking its place in the lattice. The silver ion must undergo quadrupolar deformation to squeeze through the triad of halide ions (crosshatched).

The mechanism for vacancy motion is illustrated in Fig. 21. In the simple diagonal movement at the bottom of the figure, the silver ion moves from its lattice site toward the vacancy, so that when the jump is completed, the lattice silver ion and the vacancy have merely switched places. In this case, even after the quadrupolar deformation of the silver ion, there is significant spatial overlap between the deformed ion and the two bromide ions, leading to the higher enthalpies of motion indicated in Table 3. However, Fig. 21 also shows an alternative, presumably lower energy pathway that involves two sequential noncollinear jumps, where the normal interstitial position acts as an intermediate state. But such a low energy pathway is not consistent with experimental data, and the reasons for this discrepancy have been discussed (56,58).

Impurities can be incorporated within the grain, and this is sometimes intentionally done as in doping. Whether impurities or dopants, these species are most often incorporated in their ionic form, and when they are polyvalent ions, the Frenkel equilibrium can be altered due to the requirement for overall electrical neutrality. So, for example, if an impurity in a +2 valence

Table 2. Thermodynamic Constants for Frenkel Defect Formation in Silver Halides[a]

Property	AgCl	AgBr
Formation enthalpy (eV)	1.49 ± 0.02	1.16 ± 0.02
Formation entropy (e.u.)	11.13 ± 0.20	7.28 ± 0.58
Mole fraction defects	4.92×10^{-11}	4.71×10^{-9}
Defect concentration (cm^{-3})	1.15×10^{12}	9.80×10^{13}

[a]Ref. 7.

Table 3. Thermodynamic Constants for Frenkel Defect Motion in Silver Halides[a]

Property	AgCl	AgBr
	Interstitial	
Enthalpy (eV)	0.018 ± 0.008	0.042 ± 0.011
Entropy (e.u.)	-3.83 ± 0.12	-3.34 ± 0.28
Mobility, $cm^2V^{-1}s^{-1}$	3.3×10^{-3}	7.7×10^{-4}
	Vacancy	
Enthalpy (eV)	0.306 ± 0.008	0.325 ± 0.011
Entropy (e.u.)	-0.65 ± 0.12	1.01 ± 0.28
Mobility, $cm^2V^{-1}s^{-1}$	6.1×10^{-7}	1.7×10^{-6}

[a]Ref. 7.

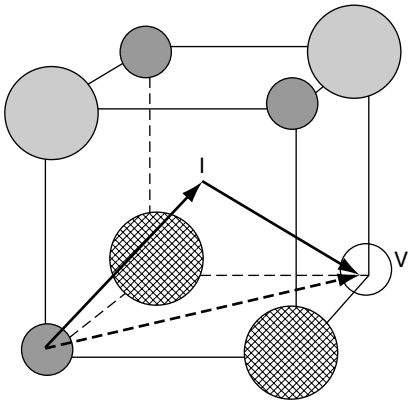


Figure 21. Schematic illustrating the movement of a vacancy V in a AgBr crystal. The lattice silver ion can move diagonally to a new lattice position (dashed arrow), thereby exchanging places with the vacancy. The ion must undergo quadrupolar deformation to squeeze through the halide dyad (crosshatched). Also shown is a presumably lower energy pathway that involves two sequential noncollinear jumps through the normal interstitial position (solid arrows).

state is incorporated, a vacancy must also be present to compensate for the extra charge on the impurity. Depending on the impurity, the vacancy may be bound to it or may dissociate from the impurity to migrate elsewhere in the crystal.

Ionic conductivity is an important property that affects both the storage and image capture properties of the silver halide detector. Thus, it is important to be able to measure this property in photographic materials. A contactless technique is required because of the small size of the microcrystal. This is done by measuring the dielectric properties of thick emulsion layers and then using a theory based on ionic relaxation of conducting particles in an insulating medium (59–68).

In this technique, the imaginary part of the dielectric constant of the coating is measured at different frequencies of an ac field. The theory predicts that the frequency at which the peak occurs is proportional to the ionic conductivity. An example of the data obtained by this technique is shown in Fig. 22 which pertains to AgCl grains doped with $CdCl_2$ (61). As the Cd^{2+} concentration increases, the peak first shifts to a lower frequency that indicates lower ionic conductivity. This can be attributed to a lowering of the interstitial silver ion concentration and a corresponding increase in the less mobile vacancy concentration. However, a further increase in the Cd^{2+} concentration shifts the peak to higher frequencies which indicates an increase in ionic conductivity. In this region, the Frenkel equilibrium is dominated by vacancies and, even though they are less mobile than interstitials, their concentration is high enough that conductivity increases.

Problems with this theory and associated theories (69) have been noted when multiple peaks are observed in the dielectric loss spectra of silver halide (70,71). Nevertheless, the reasonable correlation between observed and expected trends makes the technique at least qualitatively useful.

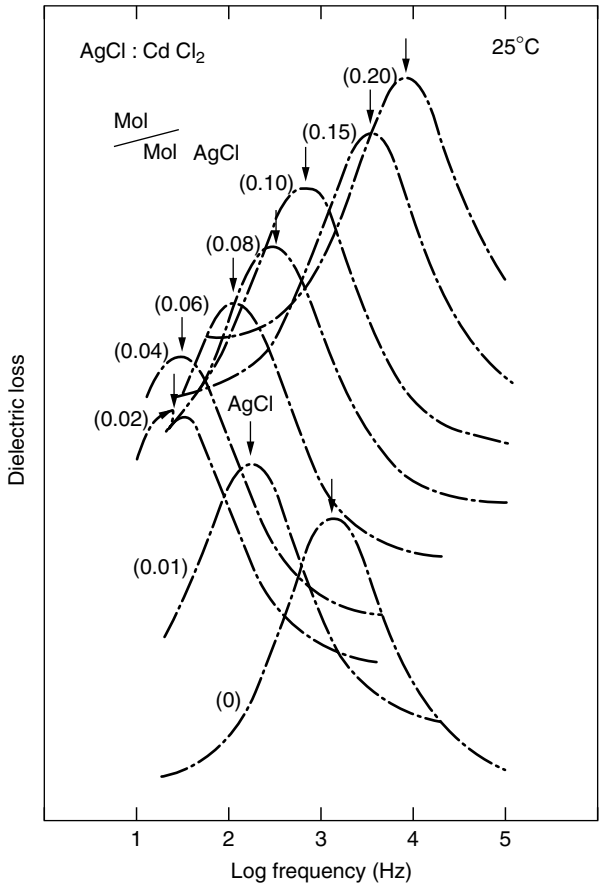


Figure 22. Dielectric loss spectra for $CdCl_2$ doped AgCl. Curve labels indicate the mole fraction of dopant. Curves are displaced vertically for clarity. (From Ref. 61. Reprinted with permission of The Society for Imaging Science and Technology, sole copyright owners of the *Journal of Imaging Science and Technology*.)

Table 4. Ionic Conductivity of Silver Halides at Room Temperature[a]

	Ionic conductivity, $\Omega^{-1}cm^{-1}$	
	Microcrystal	Bulk crystal
AgCl	0.8×10^{-8}	2×10^{-9}
AgBr	1×10^{-6}	2×10^{-8}

[a] Ref. 61.

It is of interest to compare those results obtained by techniques for measuring ionic conductivity in emulsion grains with measurements on large crystals using more conventional techniques. This comparison, given in Table 4, shows that microcrystals have 10 to 100 times higher ionic conductivity than large crystals. These results are explained by the large surface-area-to-volume ratio in microcrystals and the realization that interstitial silver ions can be created independently of vacancies at the surface via positive-kink sites, as illustrated in Fig. 23. This mechanism lowers the free energy requirements relative to the bulk process (7,72) and leads to higher concentrations of interstitial silver ions. The vast majority

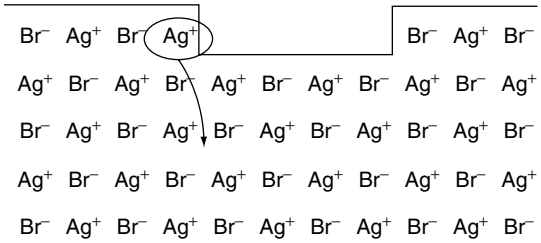


Figure 23. An edge-on view of the formation of an interstitial silver ion from a surface positive-kink site.

of the interstitial ions in emulsion grains are created by this surface generation process and the following equilibrium is appropriate:

$$Ag^{+}_{kink} \rightleftharpoons Ag^{+}_{interstitial} \tag{16}$$

The increase in the density of negative kinks at the surface is another important feature associated with the surface generation process of interstitials. In fact, there is an increase of one negative kink for every interstitial silver ion created by the surface generation process and a corresponding decrease of one positive kink. Thus, an excess negative charge is built up on the surface, which is balanced by a subsurface region of interstitial silver ions. This spatially inhomogeneous charge distribution is known as the "ionic space charge region." The concentration gradients are illustrated in Fig. 24. Note that the concentration of interstitials and vacancies is quite different near the surface, but they gradually approach equal concentrations as one moves toward the center of the crystal where the normal bulk process dominates. The potential profile that arises from the ionic space charge region is also shown in Fig. 24.

Theories that treat the ionic space region have been developed (73–77), and they predict how the concentration of interstitials and, based on knowledge of mobilities (Table 3), ionic conductivity will vary with grain size (78). These calculations agree reasonably with experimental data and show that ionic conductivity decreases with increasing microcrystal diameter. Using space charge theory, it is possible to calculate the number of interstitials per grain. A 1-μm diameter spherical AgBr grain would contain about 5,000 interstitial silver ions (78). Because of the higher energy requirements in AgCl (see Table 2), the number of interstitials in the same size AgCl grain would be two to three orders of magnitude lower.

Electronic Properties

The electronic properties of a material are controlled by the energy levels accessible to its electrons. The relevant energy levels in molecules are those of the highest filled and lowest vacant molecular orbitals. In crystalline solids, these energy levels broaden into bands called the "conduction band" and "valence band." These bands arise because the Pauli exclusion principle dictates that all of the electrons in the material be distinguishable (i.e., have a set of four quantum numbers that is unique). These bands in AgBr are derived from the Ag^{+} 4*d* and 5*s* and the Br^{-} 4*p* and 5*s* atomic orbitals (79,80), as shown schematically in Fig. 25. Of special interest is the energy gap between the two bands, called the "forbidden gap" or the "band gap." Electrons are quantum mechanically forbidden to have energies encompassed by this gap.

Insulators like AgBr and AgCl are characterized by a filled valence band and an empty conduction band. Light absorption occurs when a photon transfers its energy

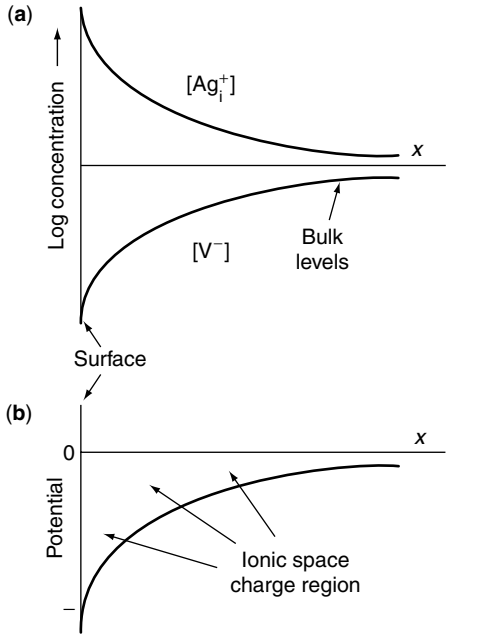


Figure 24. Concentration profiles (**a**) of interstitials and vacancies and electric potential (**b**) as a function of distance *x* from the surface. The space charge region is indicated. "Bulk levels" are the concentrations expected in the absence of the ionic space charge effect.

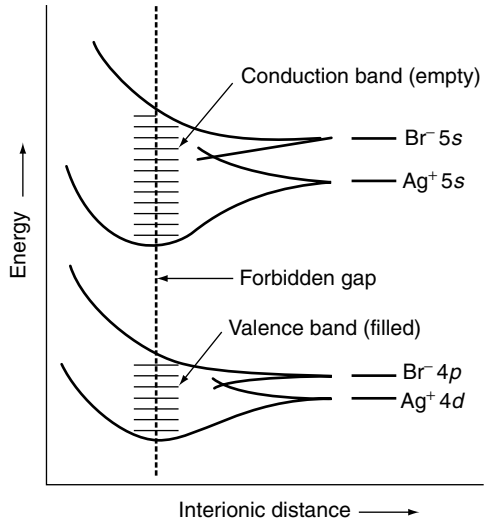


Figure 25. Energy vs. interionic distance schematic illustrating the formation of the conduction and valence bands in AgBr. On the right are the orbital energies of the isolated ions. These orbitals mix together as the interionic distance decreases to form the indicated bands.

to an electron in the valence band and promotes it to the conduction band. Once in the conduction band, the electron is free to migrate throughout the crystal, whereas those electrons in the valence band are confined to a very localized region. The energy of the band gap defines the lowest photon energy that can cause light absorption.

Based on photographic data, the effective band gap is about 3 eV (410 nm) for AgCl and about 2.6 eV (475 nm) for AgBr. Mixed halide crystals are often used in practical applications. In AgClBr crystals, the photon threshold is intermediate between that for AgCl and AgBr, whereas the threshold wavelength in AgIBr crystals extends to somewhat longer wavelengths than those in AgBr. Typical absorption spectra of large crystals shown in Fig. 26 (81) indicate that absorption extends to a slightly longer wavelength than indicated by the photographic data. But this absorption is too weak to be photographically detected.

The absorption spectra shown in Fig. 26 indicate a rather significant problem in recording color images. The AgBr detector responds only with sufficient sensitivity in the UV and blue regions of the spectrum, and incorporation of iodide extends the absorption only into the short green region. This deficiency is resolved by adsorbing spectral sensitizing dyes on the silver halide surface, which extends their absorption range through the red and even into the near IR. We will discuss this feature in a later section.

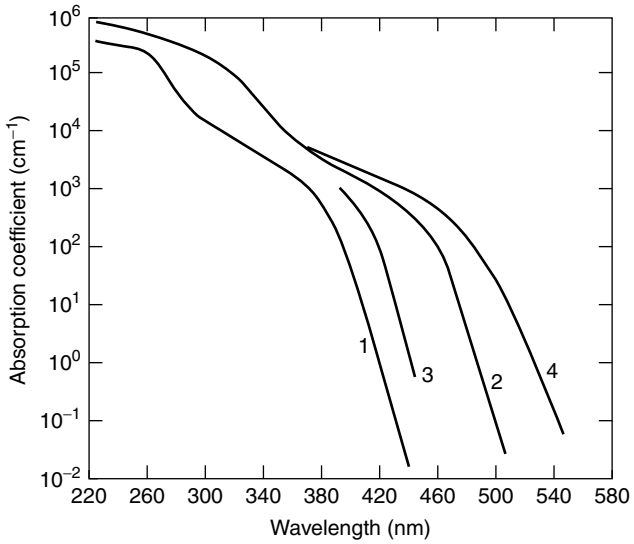


Figure 26. Absorption spectra of pure and mixed halides at room temperature. (1) AgCl; (2) AgBr; (3) AgCl containing 10 mole% AgBr; (4) AgBr containing 3 mole% AgI. (From Ref. 81.)

Light absorption also leads to an empty electron position in the valence band which is referred to as a "hole." Neighboring electrons can move to occupy the empty electron position, but this creates a vacant position elsewhere; the hole appears to move. Although the electrons in the valence band are the actual charge carriers, it is mathematically valid and simpler to focus on the hole as the moving entity. The hole has a characteristic mobility and mass, as well as a positive charge to complement the negative charge on the electron.

Electrons and holes are confined to the energy bands, but localized regions in the crystal lattice, known as "traps," can exist at energies within the forbidden gap. These traps are sites that may localize the electron or hole, as illustrated in Fig. 27. An electron is localized within a trap by losing enough energy, so that its energy is now coincident with that of the trapping level. The excess energy is dissipated to the surrounding lattice ions as thermal energy. An escape from the trap requires the trapped electron to acquire sufficient thermal energy so that its energy is now coincident with the bottom of the conduction band. Note that the holes in this diagram are trapped "upward" because the energy direction refers to that of the electron. Bear in mind that when a hole moves either spatially or energetically in one direction, it can be viewed as an electron moving in the opposite direction.

Traps in silver halide materials are impurities (such as dopants) and lattice defects (such as dislocations and kink sites). Lattice defects generally have an associated uncompensated charge and interact with charged carriers (electron or hole) via Coulomb's law. The situation is analogous to that of a hydrogen atom with a central charge and an orbiting particle of opposite charge. An equation to calculate the binding energy of the carrier to the trapping site comes from the Bohr treatment of the H atom, modified for the effective mass m^* of the carrier and the dielectric constant ε of the material (82,83):

$$E = \frac{q_t^2 e^2 m^*}{2\overline{h}^2 \varepsilon^2} \tag{17}$$

where q_t is the trap charge and $\overline{h}$ is Planck's constant divided by 2π. The effective mass arises because the carriers within the crystal appear to move with a mass different from their real mass. This effect is induced by the carrier interaction with the surrounding lattice ions as it moves through the crystal. An analogous equation is

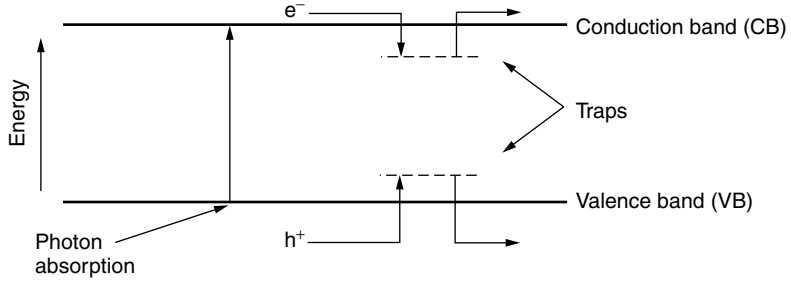


Figure 27. Illustration of electron and hole trapping levels in a crystal. The energy arrow applies only to electrons; an oppositely directed one exists for holes, but is not shown

available for calculating orbital radius r:

$$r = \frac{e\overline{h}^2}{q_t e m^*} \tag{18}$$

Typical binding energies and orbital radii are given in Table 5. The binding energies for the electron are small (kT or less), and the radii are correspondingly large (10's of lattice spacings). Because of its much higher effective mass, the hole has a much higher binding energy and smaller orbital radii than the electron. However, the trapped carrier in both cases is in a delocalized state characteristic of coulombic traps. Impurity traps are derived from the atomic energy levels of the specific impurity but are modified by their environment within the crystal lattice. Examples of such traps are I^- and Cu^+. These are isoelectronic with the lattice ions that they replace and therefore provide no long-range coulombic effects. However, their atomic energy levels are different from the surrounding lattice ions and provide the requisite trapping level. In contrast with the coulombic traps, these impurity traps provide high localization of the trapped carrier. In addition to these two types of trapping sites in the silver halide lattice, a third possibility is derived by combining these two; an example is Ir^{3+}.

The probability of trapping by a particular trap is determined by its effective trapping radius or cross-sectional area (84). The traps present a "sphere of influence" to a carrier diffusing through the crystal, and trapping occurs if the carrier passes through this sphere. The effective trapping radius is determined by the amount of energy that must be dissipated as a carrier falls from the free band to the trapping level. If the amount of energy to be dissipated is only on the order of a few kT or less, then the transition has unit probability of occurring, and the trapping radius is determined solely by the charge on the trap. This is typical for coulombic traps such as lattice defects. Then, the capture cross section of these so-called "shallow traps" is determined by equating the binding energy to the average thermal energy available (kT), and the resulting trap cross section for a trap with charge 0.5 e is of the order of 10^{-13} cm^2. Although obviously a small number, this cross section is about 100 times larger than that calculated for an uncharged site, which would have a trapping radius of the order of a lattice spacing.

Table 5. Trap Characteristics of Charged Centers in Silver Halides

	AgBr		AgCl
	e^-	h^+	e^-
m^*/m	0.29	1.71	0.43
ε	13.1	13.1	12.3
$q_t = 1$			
E(meV)	23	140	39
r(Å)	24	4	15
$q_t = 0.5$			
E(meV)	6	45	10
r(Å)	48	8	30

When the amount of energy to be dissipated is larger than a few kT (>0.05 eV), a multistep process is required in which the energy is transferred to the surrounding lattice in small increments (85). This process requires the presence of intermediate energy levels that act like rungs on a ladder. These so-called "deep traps" are usually also associated with a shallow trapping coulombic level. For example, a silver atom located at a positive-kink site would be characterized by a shallow trapping level due to the coulombic nature of the kink site and by a deep trapping level associated with the 5s atomic energy level of the silver atom. This is depicted schematically in Fig. 28. The effective trapping cross section σ_t is given by (86)

$$\sigma_t = \sigma_c \eta = \sigma_c \left(\frac{t}{t+i} \right) \tag{19}$$

where σ_c is the trapping cross section of the coulombic trap and η is a "sticking coefficient" that describes the probability that the carrier will transit to the deep level t, as opposed to returning to the conduction band i.

Trapping of the electron (or hole) causes a lattice distortion (lattice relaxation) around the localized carrier because the charge density has been altered (87). Basically, the ions rearrange themselves to accommodate the higher (lower) charge, thereby lowering the energy and deepening the trap relative to the case in which there is no lattice relaxation. These processes are represented as configuration coordinate diagrams. An example of such a diagram is given in Fig. 29 for electron trapping at a deep trap such as a silver atom, which also has an associated coulombic level. The three-dimensional nature of the crystal is collapsed into a one-dimensional lattice distortion Q, where $Q = 0$ is the equilibrium configuration in the absence of the trapped carrier. Once the electron is trapped in the coulombic level, it can move to the relaxed state, provided that the energy barrier E_a is less than the coulombic trap depth E_c. Transition to the deep level occurs where the energy curves of the relaxed coulombic level and the deep level intersect (point C in Fig. 29). As drawn, this transition is favored, but, if the crossing point C lies above the barrier for thermal escape (point B), then detrapping is favored.

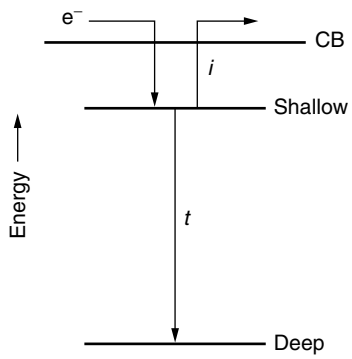


Figure 28. Schematic showing the capture of an electron at a deep trap. Transition t represents the movement of the electron to the deep level, whereas transition i represents thermal release to the conduction band from the shallow trapping level.

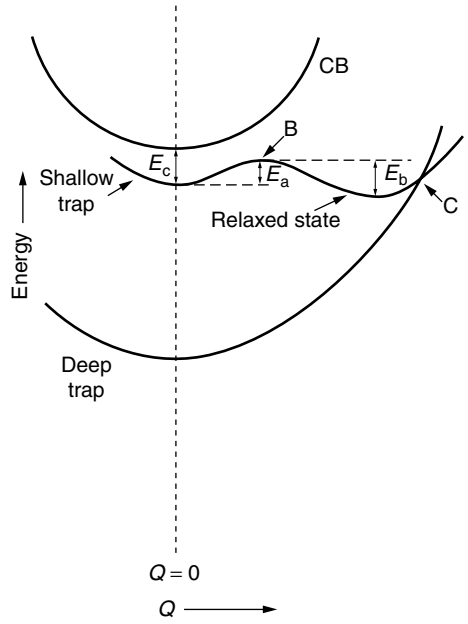


Figure 29. Configuration coordinate diagram for trapping at a deep trap. Following capture of the electron at the shallow trap (coulombic level), a thermally assisted process allows formation of the relaxed state. From this state, the transition to the deep level occurs at crossing point C, also requiring thermal assistance. E_c is the coulombic trap depth of the shallow trap, E_a is the energy barrier at point B to relaxation, and E_b is the thermal trap depth of the relaxed state. Q is a one-dimensional simplification of the three-dimensional rearrangement of lattice ions as they accommodate the extra charge of the electron.

Escape of a carrier from a trap involves transferring thermal energy from the surrounding lattice ions to the carrier. The rate of transfer depends on the lattice vibrational frequency which is about 10^{12} s^{-1} in AgBr. The probability that a certain amount of energy will be available is given by the usual Boltzmann factor, so that the overall rate of escape of a carrier is given by

$$\text{rate}_{\text{escape}} = \nu \exp\left(\frac{-\Delta E}{kT}\right) \tag{20}$$

where ν is the attempt-to-escape frequency, synonymous with the lattice vibrational frequency, and ΔE is the trap depth, which is measured from the edge of the free band to the trap energy level. The inverse of the release rate gives the time for escape of the carrier from the trap, so that

$$\tau = \nu^{-1} \exp\left(\frac{\Delta E}{kT}\right) \tag{21}$$

Table 6 summarizes some typical escape times for a range of trap depths and indicates the kinds of entities that are associated with these trap depths.

A major loss process in the photochemistry of silver halides is recombination between electrons and holes. Quantum mechanical selection rules prevent recombination between free electrons and free holes (7). Nevertheless, recombination does occur, provided that one of the carriers is trapped, as illustrated in Fig. 30.

Table 6. Trap Escape Times

Trap depth (eV)	Escape time[a]	Comments
0.01	0.4 ps	Intrinsic defect traps
0.05	1.8 ps	Intrinsic defect traps
0.10	13.6 ps	Intrinsic defect traps
0.20	0.75 ns	Chemical sensitization traps
0.30	41 ns	Chemical sensitization traps
1.0	67 h	Latent image

[a] $\nu = 10^{-12}$ s^{-1}

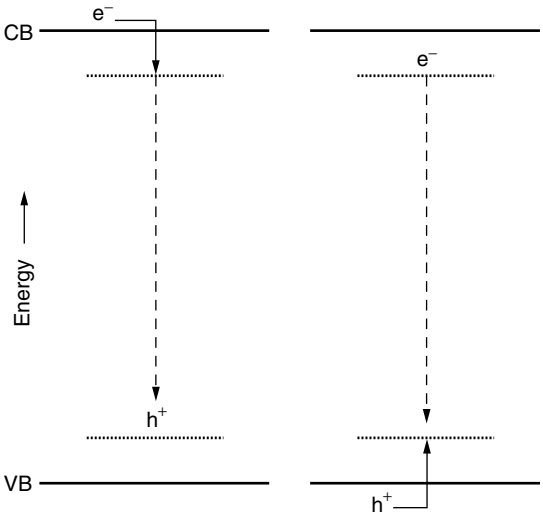


Figure 30. Electron-hole recombination pathways in AgBr. Dotted lines depict trapping levels for carriers. Dashed lines indicate the recombination step. In the right-hand diagram, the hole is undergoing trapping at a filled level of the center that is providing the electron trapping level, after which the recombination occurs.

Under most conditions, the pathway on the left is favored because holes have a higher effective mass than electrons and therefore are more deeply trapped and because the electron's mobility is much higher than that of the hole. The recombination often leads to luminescence at low temperatures (88,89), but this process is quenched at room temperature where nonradiative processes dominate.

The photochemistry of silver halides involves the formation of atomic species — silver and halogen. The silver atom is formed when an interstitial silver ion migrates to the site of a trapped electron. The energy level scheme associated with this transformation is shown in Fig. 31. The electron is initially trapped in a shallow coulombic trap that is characteristic of a positive kink. Following this initial trapping, the lattice surrounding the trapped electron relaxes to accommodate this new added charge, thus deepening the trap. Then, the interstitial silver ion is attracted to this site which now bears a partial negative charge. Finally, the electron is localized at the deep level corresponding to the 5*s* level of the silver atom. There is a corresponding process for the hole, except that now a lattice silver ion must leave the site

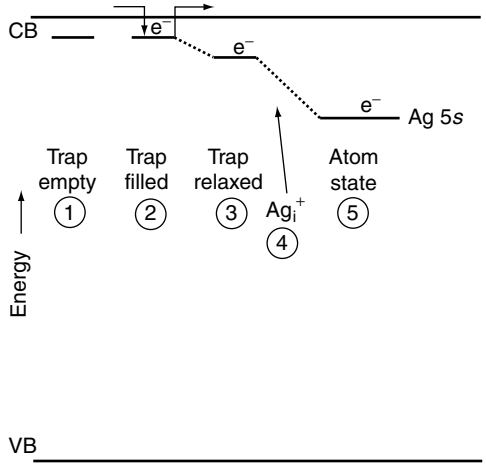


Figure 31. Sequence of events leading to the formation of a silver atom. Circled numbers indicate the sequence of events. Arrow depicts the ionic step in which a silver ion interstitial is captured at the relaxed trap.

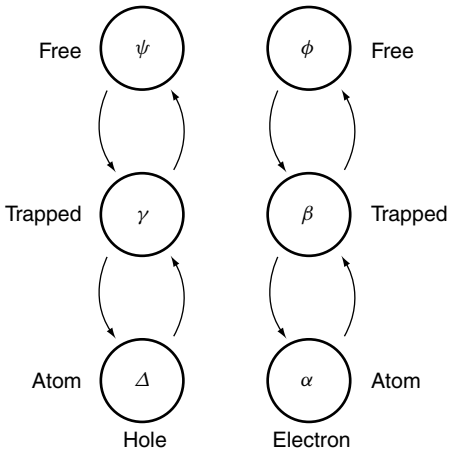


Figure 32. Schematic representation of the three-way equilibria between the free, trapped, and atom states for both the hole and electron.

adjacent to a trapped hole at a negative-kink site and move into an interstitial position, leaving behind a halogen atom (90).

Both the silver and halogen atoms are unstable species. Most of the evidence for the instability of the silver atom, will be discussed in later chapters. Now, we can merely realize that for the silver cluster formation process to be efficient, which it often is, there must be a pathway for collecting all of the photoproduced silver atoms in one or a few sites on the grain surface. Although diffusion of silver atoms on the grain surface is one way of accomplishing this, the general consensus is that silver clusters form by alternate addition of electrons and interstitial silver ions, as discussed in the next section. Thus, the silver atoms must decay and re-form at a preferred growth site. There is no direct way to measure the lifetime of the silver atom, but indirect methods provide an upper limit of about 1 s (91–96). In the case of halogen atoms, experiments in which large silver halide crystals are placed in a chamber of halogen gas have shown increased conductivity of the crystals; this could happen only if the adsorbed halogen gas injected a hole into the valence band, resulting in the formation of a halide ion (97).

In this section we have described three effective states for the electron — free, trapped, and atom. Similar states exist for the hole. These states are all transient; it is useful to use the schematic in Fig. 32 to describe the transitions between the different states. The fraction of time spent in each state is symbolized by Greek letters, so that $\phi + \beta + \alpha = 1$ and similarly for the three hole states. Because of the high interstitial silver ion concentration in silver bromide, the electron spends most of its time in the atom state. The low mobility of the hole coupled with its higher effective mass results in a deeply trapped state which means that the hole spends most of its time in the trapped state. We will return to this partitioning of the electron and hole into transitory states in the next section.

LATENT-IMAGE FORMATION

Nucleation-and-Growth Model

A number of models have been developed to explain latent-image formation in silver halides (7,98–114). The most successful of them is based on ideas first proposed by Gurney and Mott in 1940 (99,100), shortly after the new ideas of quantum mechanics had been applied to crystalline solids. The Gurney–Mott model is based on the following assumptions:

- Silver atoms form when photoelectrons combine with preexisting mobile interstitial silverions.
- Silver atoms are unstable, and their diffusion to give silver clusters is not possible.
- Aggregation of silver atoms proceeds by the addition of photoelectrons and mobile interstitial silver ions to the initial trapping site.
- Trapping of a number of electrons occurs at a given site in the microcrystal and then a like number of interstitial silver ions is added.

It was recognized rather quickly that the Gurney–Mott picture had a serious flaw. After trapping the first electron, a trapping site would acquire a negative charge that would repel subsequent electrons. To avoid this problem, it was suggested that each trapping of an electron should be followed by the addition of a silver ion (115). Thus, the charge should alternate between 0 and −e. Later, it was pointed out that the sites where the silver clusters form are most likely to be physical defects such as kink sites, that have a partial electrostatic charge (116). The charge on the site should alternate between partial negative and partial positive as electrons and silver ions are added alternately. These ideas have become the basis of a model first proposed by Hamilton and Bayer (117) and subsequently referred to as the nucleation-and-growth model (7,106) because it includes a two-stage concept of nucleation and growth (118–123). It also considers that the trapping of electrons at trapping sites and atom

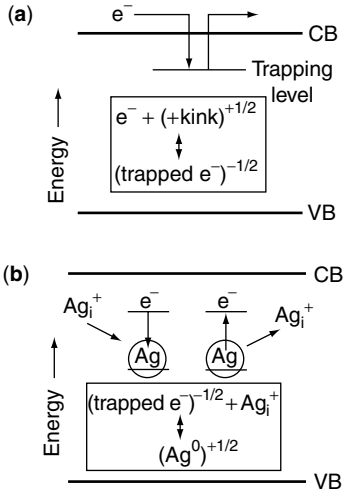


Figure 33. The initial electronic (**a**) and ionic (**b**) events in the formation of the latent image. Figures give a band diagram view, whereas the inset shows the chemical equilibrium between the two states.

formation are reversible, as shown in Fig. 33. These two reversible events form the basis for the three-way cycle diagrams seen earlier in Fig. 32.

Following the formation of a silver atom, it is possible for a second electron to be trapped at this site followed by the addition of a second silver ion. These two events constitute the nucleation stage and result in a two-atom center. Unlike the silver atom, the two-atom center is considered a permanently stable species during latent-image formation. Because nucleation cannot occur without a silver atom, nucleation requires at least two electrons in the grain. Now, we can add nucleation to the other events depicted in Fig. 31 to get Fig. 34. Growth is the continuation of this scheme of alternate addition of electrons and interstitial silver ions at the trapping site. The growth stage continues as long as both reactants are available.

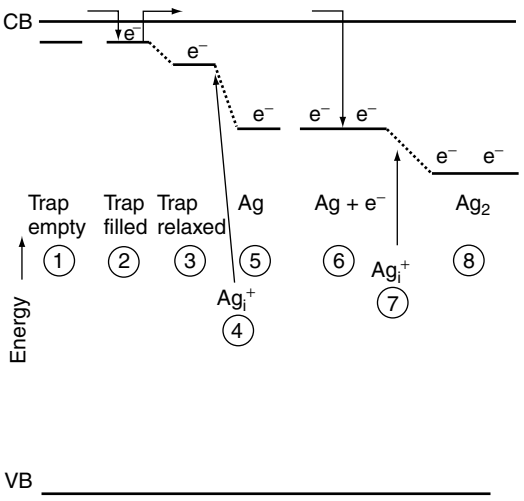


Figure 34. A continuation of Fig. 31 showing the formation of Ag_2.

Interstitial silver ions are continually replenished via the Frenkel equilibrium (Fig. 17), so the terminating step of the growth sequence is exhaustion of the electron supply.

A basic premise of the model is that the various states alternate in charge and provide an electrostatic driving force for each step (7), as shown in Fig. 35. As the state of the grain moves from left to right along this diagram, larger and larger silver clusters are produced. As we will see shortly, a certain size cluster is required to promote the development of the grain, so that growth need only achieve this threshold size for the grain to be developable and thereby contribute to the macroscopic image.

From what has been said so far, one might conclude that, for example, if a five-atom silver cluster is required to initiate the development reaction, then the grain need absorb only five photons to produce such an occurrence. However, we have not yet included the major inefficiency in the model that will dramatically increase the number of absorbed photons required for the grain to achieve developability. This inefficiency involves recombination between electrons and holes, and, as discussed earlier, there are two possible pathways. For free electrons recombining with trapped holes, we recognize that a negative-kink site would be a possible hole trap that would then act as a recombination center whose formal charge is $+0.5\,e$. Likewise, for free holes recombining with trapped electrons, we recognize that an electron trapped at a positive-kink site would be a recombination center whose charge is $-0.5\,e$. Thus, both pathways have an electrostatic driving force that allows them to compete readily with nucleation and growth. As discussed earlier, the dominant pathway is recombination between free electrons and trapped holes.

Figure 36 summarizes all of the events in latent-image formation that we have discussed. On the left are the two three-way cycles that characterize the transient states of the electron and hole. Also included within this region are the two recombination pathways. To the right are the irreversible events of nucleation and growth. Note that nucleation involves a free electron combining with an electron in an atom state to produce an electron-at-an-atom state. Then, moving diagonally from this state represents the addition of a silver ion to form a two-atom center. Similar processes occur to produce larger

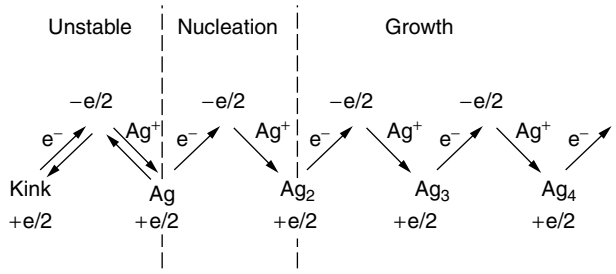


Figure 35. The sequence of events that lead to the formation of a developable silver cluster. Note the alternation of charge on the site which is given at the top and bottom of the diagram. (From Ref. 106. Reprinted with permission from World Scientific Publishing Co. Pte. Ltd.)

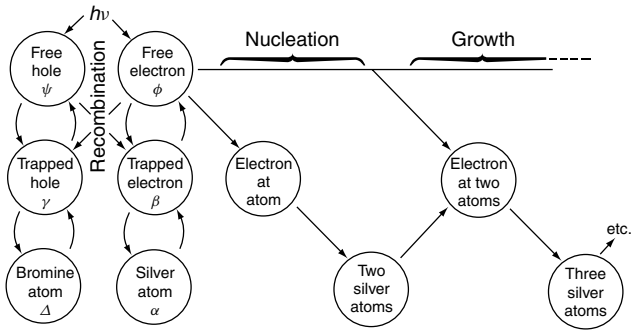


Figure 36. A schematic representation of the essential events in the nucleation-and-growth model. Light absorption is symbolized by $h\nu$. Additional events may be needed to treat specific cases. (From Ref. 124. Reprinted with permission of The Society for Imaging Science and Technology, sole copyright owners of the *Journal of Imaging Science and Technology*.)

silver clusters. The scheme depicted in Fig. 36 is the minimum number of events needed to explain latent-image formation. The sensitometric behavior of silver halides can be explained largely by understanding how the forward moving processes of nucleation and growth compete with the backward moving recombination process, as well as how nucleation and growth compete with each other. There will be many examples in the remaining sections to illustrate these concepts. In some situations, additional events may be needed to explain latent-image formation.

The ideas expressed before concerning latent-image formation can be put into equations that can be used as a basis for a simulation program to predict the photographic consequences of various factors (117,125–127). The three-way cycles can be treated with a steady-state approximation. Equations for the irreversible events can be derived by assuming that the ionic event — the addition of an interstitial silver ion — is a fast process and electron or hole capture is rate-limiting. Generically, the rates λ_{ct} of the irreversible processes can be expressed as

$$\lambda_{ct} = n_c n_t v_c \sigma_t V^{-1} \tag{22}$$

and

$$= n_c n_t \Gamma_{ct} \tag{23}$$

where n_c is the number of carriers, n_t is the number of traps, v_c is the thermal velocity of the carrier, σ_t is the capture cross section of the trap, Γ_{ct} is the rate constant, and V is the grain volume which is used to put the rates in per-grain terms.

If we let e be the number of electrons partitioned among the free, trapped, and atom states, and let h be the number of holes similarly partitioned, then, using the template expressed by Eqs. (22,23), the rate of nucleation can be expressed as

$$\lambda_n = [\text{free electrons}][\text{silver stoms}](\text{rate constant}), \tag{24}$$

$$= (\phi e)\alpha(e-1)v_e\sigma_{\mathrm{Ag}}V^{-1} \tag{25}$$

and

$$= (e)(e-1)\Gamma_n \tag{26}$$

where

$$\Gamma_n = \phi\alpha v_e \sigma_{\mathrm{Ag}} V^{-1} \tag{27}$$

Note how the concentration of silver atoms is expressed. The $(e-1)$ factor is used so that the nucleation rate cannot be greater than zero unless there are two or more electrons in the three-way electron cycle. Physically, this factor indicates that an electron can participate with all other electrons in silver cluster formation, but not with itself. Similarly, a growth rate at an n-atom silver cluster can be expressed as

$$\lambda_{\mathrm{g(n)}} = [\text{free electrons}][\text{n-atom centers}](\text{rate constant}) \tag{28}$$

$$= (\phi e)(\mathrm{Ag}_n)v_e\sigma_{\mathrm{Ag_n}}V^{-1} \tag{29}$$

and

$$= eAg_n\Gamma_{\mathrm{g(n)}} \tag{30}$$

where

$$\Gamma_{\mathrm{g(n)}} = \phi v_e \sigma_{\mathrm{Ag_n}} V^{-1} \tag{31}$$

and Ag_n is the number of n-atom centers. Equation (30) gives the rate of growth at just one size silver cluster. The total growth rate is the sum of rate expressions for all of the silver clusters present. Note that the rate constant for growth differs from that for nucleation primarily in the capture cross section used — that for a silver atom versus that for a silver cluster. We will shortly see that this is a very fundamental difference.

We can develop similar equations for recombination. However, now there are two rates — one for free-electron/trapped-hole recombination and the other for free-hole/trapped-electron recombination. For the first pathway,

$$\lambda_{r1} = [\text{free electrons}][\text{trapped holes}](\text{rate constant}), \tag{32}$$

$$= (\phi e)(\gamma h)v_e\sigma_\gamma V^{-1} \tag{33}$$

$$= eh\Gamma_{r1} \tag{34}$$

where σ_γ is the cross section for electron capture at a trapped-hole site. For the second pathway,

$$\lambda_{r2} = [\text{free holes}][\text{trapped electrons}](\text{rate constant}), \tag{35}$$

$$= (\psi h)(\beta e)v_h\sigma_\beta V^{-1} \tag{36}$$

$$= eh\Gamma_{r2} \tag{37}$$

where σ_β is the cross section for hole capture at a trapped-electron site,

$$\Gamma_{r1} = \phi\gamma v_e\sigma_\gamma V^{-1} \tag{38}$$

and

$$\Gamma_{r2} = \psi\beta v_h\sigma_\beta V^{-1} \tag{39}$$

The rate constant for the total recombination is

$$\Gamma_r = \Gamma_{r1} + \Gamma_{r2} = [\phi\gamma v_e\sigma_\gamma + \psi\beta v_h\sigma_\beta]V^{-1} \tag{40}$$

Equation (40) can be simplified by the following definition:

$$f = \frac{v_h\sigma_\beta}{v_e\sigma_\gamma} \tag{41}$$

so that now

$$\Gamma_r = \phi\left(\gamma + \frac{f\psi\beta}{\phi}\right)v_e\sigma_\gamma V^{-1} \tag{42}$$

Finally, to make the expression for Γ_r similar to that for Γ_n and $\Gamma_{g(n)}$, we make the following definition:

$$\omega = \gamma + \frac{f\psi\beta}{\phi} \tag{43}$$

which then leads to

$$\Gamma_r = \phi\omega v_e\sigma_\gamma V^{-1} \tag{44}$$

Now, let us summarize the rates for the three irreversible events of interest:

$$\lambda_n = e(e-1)\Gamma_n \tag{45}$$

$$\lambda_{g(n)} = e\mathrm{Ag}_n\Gamma_{\mathrm{g(n)}} \tag{46}$$

$$\lambda_r = eh\Gamma_r \tag{47}$$

where

$$\Gamma_n = (\phi v_e V^{-1})\alpha\sigma_{\mathrm{Ag}} \tag{48}$$

$$\Gamma_{\mathrm{g(n)}} = (\phi v_e V^{-1})\sigma_{\mathrm{Ag_n}} \tag{49}$$

and

$$\Gamma_r = (\phi v_e V^{-1})\omega\sigma_\gamma \tag{50}$$

These rate equations can be further simplified by making the following assumptions. First we recognize that the shallow coulombic levels involved in nucleation, growth, and recombination have similar charges, so that the shallow-level trapping cross section σ_c is the same for all three events (128). Second, because charge is the predominant factor that determines the trap cross section, the size of the silver cluster has little effect on σ_c.

A further assumption is that the transition to the deep level is inefficient for the silver atom, but efficient for silver clusters. Figure 37 is a configuration coordinate diagram, similar to that shown in Fig. 29, that compares the electron transition to the deep level for these two centers. Based on molecular orbital calculations, it is postulated that there are intermediate levels for two-atom silver clusters and larger that facilitate the dissipation of energy, but no such levels exist for electron capture into the silver-atom site (106). Recall that earlier we defined the trapping cross section in terms of the shallow-level trapping cross section and the sticking coefficient η [Eq. (19)],

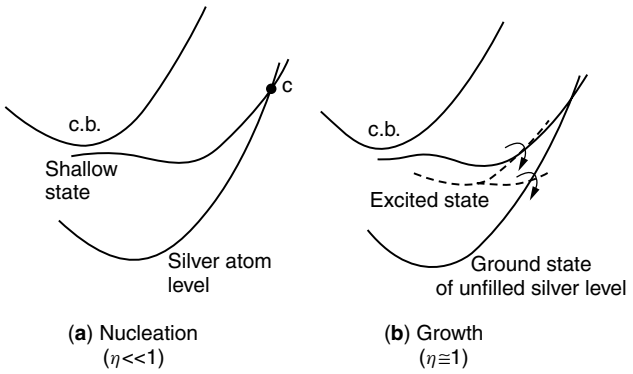


Figure 37. A configuration coordinate diagram, similar to that in Fig. 29, that illustrates the trapping of an electron at a silver atom (**a**) and a silver cluster (**b**) in an unsensitized grain. The "excited state" curves in (**b**) indicate postulated intermediate orbital energies available in a silver cluster that are not available in the silver atom. (From Ref. 106. Reprinted with permission from World Scientific Publishing Co. Pte. Ltd.)

$$\sigma_t = \eta\sigma_c \tag{51}$$

Following the preceding discussion, we can identify the sticking coefficient for nucleation (Fig. 37a) as η_n, and that for the growth (Fig. 37b) as η_g. We can also think of these sticking coefficients as efficiencies and also note that $\eta_n \ll \eta_g$ based on the previous arguments. If we assume that the transition to the deep level for both growth and recombination is highly efficient, then $\eta_g = \eta_r = 1$.

Based on the previous simplifications, we can now rewrite the rate constant expressions [Eqs. (48–50)] using

$$\sigma_{\mathrm{Ag}} = \eta_n\sigma_c \tag{52}$$

$$\sigma_{\mathrm{Ag_n}} = \eta_g\sigma_c \tag{53}$$

and

$$\sigma_r = \eta_r\sigma_c \tag{54}$$

to give

$$\Gamma_n = (\phi v_e V^{-1})\alpha\eta_n\sigma_c \tag{55}$$

$$\Gamma_g = (\phi v_e V^{-1})\eta_g\sigma_c \tag{56}$$

and

$$\Gamma_r = (\phi v_e V^{-1})\omega\eta_r\sigma_c \tag{57}$$

Furthermore, because we have already argued in the previous section that $\alpha \approx 1$ and have now assumed that $\eta_g = \eta_r = 1$, then

$$\Gamma_n = (\phi v_e V^{-1})\eta_n\sigma_c \tag{58}$$

$$\Gamma_g = (\phi v_e V^{-1})\sigma_c \tag{59}$$

and

$$\Gamma_r = (\phi v_e V^{-1})\omega\sigma_c \tag{60}$$

From these equations, we can define the nucleation efficiency η_n and the recombination index ω as

$$\eta_n = \frac{\Gamma_n}{\Gamma_g} \tag{61}$$

and

$$\omega = \frac{\Gamma_r}{\Gamma_g} \tag{62}$$

Silver Clusters and Development

We will discuss development of the latent image in much detail later, but some discussion is required at this point to show how the latent image confers developability upon the emulsion grain. Developers are multicomponent solutions, but the key ingredient pertinent to our present discussion is the developing agent itself. These are organic molecules that can transfer electrons to appropriate acceptors, here, the latent image. A typical energy-level diagram for the initiation of development is shown in Fig. 38. On the left of the diagram is an exposed grain in which the electron accepting level of the photoproduced silver cluster is within the band gap and is energetically coincident with the highest filled (HF) molecular orbital of the developer molecule. Thus, electron transfer from the developer molecule to the latent image is energetically favorable.

Also, shown in Fig. 38 on the right is the situation for an unexposed grain. Electron transfer from the developer to the grain in this case can occur only by transfer to one of the many energy levels within the conduction band. Such a transition has low probability because of the high energy barrier. However, if given enough time, electron transfer will occur and lead to the formation of a silver cluster whose energy levels are within the band gap. Now, further electron transfers become much more likely and cause the unexposed grain to develop. This grain will contribute to the fog density.

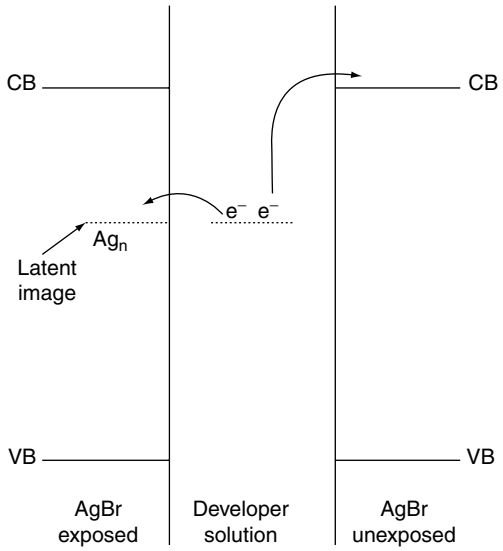


Figure 38. An energy-band diagram depicting the development stage of the photographic process for an exposed and an unexposed grain. The energy level of the highest filled molecular orbital of the developing agent is depicted in the center of the diagram.

It should be apparent from the preceding discussion that the discrimination between exposed and unexposed grains is due to the different kinetics of development of these two types of grains. The presence of the latent image increases the development rate and allows detecting the latent image before significant fog arises. This large difference in development kinetics makes the photographic process viable.

There is evidence that there is a critical silver cluster size, below which the development reaction cannot be initiated within the normal development time. Indirectly, experiments with vacuum-deposited silver clusters on SiO_2 followed by physical development have shown that development is initiated by four-atom and larger silver clusters (129). Also, light latensification experiments, in which the film is given an overall uniform exposure at low irradiance following an initial exposure through an irradiance-modulating step wedge, have led to sensitivity increases up to a factor of 10 (130,131). The uniform exposure in these experiments provides additional electrons that allow the subdevelopable silver cluster ("latent subimage") to grow to a latent image. The number of developable grains increases and results in the observed speed increase. These experiments indicate there are stable silver clusters that are not detectable under the usual development conditions.

Direct techniques have recently been devised for determining the minimum developable size which involve selecting the size of silver clusters emitted from an oven and directing them toward a target that contains silver bromide grains from which a major part of the adsorbed gelatin has been removed (132,133). Initial concerns about fragmenting the clusters have been decreased by using a buffer gas that reduces the kinetic energy of the clusters when they arrive at the grain. Results show that grains become developable when four-atom and larger clusters are used. Concerns still remain that residual gelatin is present and that the deposited clusters reside on top of the gel layer, where their catalytic activity may be different from that of clusters adsorbed directly on the grain surface.

The preceding discussion has focused on a single threshold size for catalyzing the development reaction. In fact, studies at different development times show that the sensitivity continually increases as the development time increases, until a fog limit is reached. One cannot explain such results by assuming a fixed threshold size or a threshold size that changes by one-atom jumps. Rather, it is necessary to assume that there is a distribution of developable sizes that varies continuously with increasing development time and allows detection of smaller and smaller clusters (134).

In establishing a threshold size for development, we are really saying that the energy level of the silver cluster is such that it acts as an acceptor for electrons transferred from the HF level of the developer molecule. Early attempts to explain these results hypothesized that the electron accepting levels of the silver clusters continually decreased with increasing size relative to the CB bottom until the level matched that of the developer HF level, whereupon development would be initiated (135).

Because bonding in silver clusters is expected to involve primarily the 5*s* electrons, simple molecular orbital theory can also be used to calculate occupied and unoccupied states (5). A typical result is shown in Fig. 39. Note the odd–even oscillating nature of the electronic properties. Clusters that contain an even number of atoms have either filled or empty orbitals; this results in a large difference between their electron affinity and ionization potentials. Clusters that contain an odd number of atoms have half-filled, as well as filled and empty orbitals, that result in a near coincidence of the electron affinity and ionization potentials. This alternating odd–even character has been confirmed in gas-phase metal clusters that have filled *d* shells and lone *s* electrons (136). Such behavior has also been found in photoproduced silver clusters on emulsion grains (137).

To be relevant to the photographic process, an energy-level picture of silver clusters adsorbed on AgBr is needed. This has been attempted in calculations using a model that consists of silver clusters in virtual lattice sites but allows no ionic relaxation of the lattice (138). Electronic interactions between the AgBr substrate and the silver clusters, however, were allowed, and the energy-level picture that emerges is similar to that in Fig. 39. A feature missing in these results is an indication of a critical threshold size. More recent studies using more sophisticated calculations that allow for lattice relaxation have produced similar results (139).

Inefficiencies

Recombination has already been discussed previously as a dominant inefficiency, but other inefficiencies can also be important in certain situations. Inefficiencies can be classified into three categories (140). The first category deals with the loss of light-generated electrons, which includes recombination as the dominant inefficiency. Other pathways for electron loss involve reactions with oxygen and/or water in the film coating and irreversible trapping at impurities either within the grain or adsorbed on the surface. The second category deals with loss of silver which includes reaction with oxygen and/or water, thermal dissociation, and oxidation by reaction with holes. The latter process is sometimes called "regression" because it is the reverse of the cluster-forming reaction.

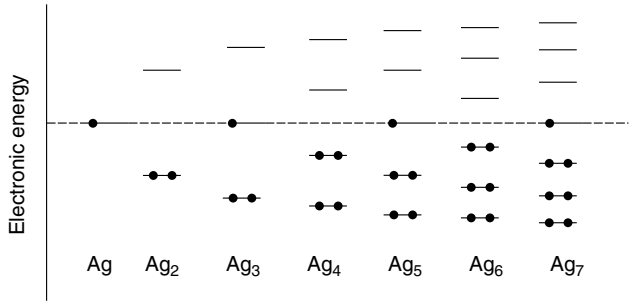


Figure 39. An energy-level diagram from simple molecular orbital theory for isolated linear silver clusters. Only those orbitals formed from 5*s* atomic orbitals are shown. Filled circles represent electrons. (Reprinted with permission from the American Institute of Physics.)

The third category focuses on the inefficient use of electrons that involve either internal latent-image formation or dispersity. In the former, competition between internal sites provided by lattice defects or perhaps impurities and surface sites leads to internal latent-image formation. All developers are formulated to detect only the latent image at the surface of the grain, or at best slightly subsurface, so some of the silver is produced in a region that is inaccessible to the developer and results in a speed loss.

When all of the surface silver is not located at the same site, this is termed a "dispersity inefficiency." Any inefficiency from the dispersity of silver may be equated specifically to the undevelopable fraction of grains, which would be developable if the silver atoms they contain could be concentrated into a single center per grain (141). Thus, the inefficiency depends strongly on the minimum developable size. Dispersity occurs when the rate of nucleation becomes comparable to that for growth. We see from Eqs. (45,46,55,56) that this would occur when e and Ag_n are large and $\eta_n \approx \eta_g$. From what was said earlier about nucleation efficiency, it seems that the latter condition is not possible, but we will see later that such a condition is approached in chemically sensitized emulsions. The discussion that follows assumes this.

If the minimum developable size were two atoms, then dispersity would be nonexistent because the initial nucleation makes the grain developable and any other nucleation would be superfluous. If the minimum developable size is three atoms, the dispersity inefficiency would correspond to the fraction of grains that has two or more two-atom clusters, but none larger. Simple probability considerations indicate that this fraction will be very small except in the shoulder region of the $D-\log E$ curve, where it would cause a loss in sensitivity of about 40%. When the minimum developable size is four or more atoms, the number of ways of configuring the silver in stable but undevelopable centers on the surface increases (142) and leads to strong dependence of the degree of dispersity inefficiency on the threshold size for development.

In an ideal situation leading to one cluster per grain, only one nucleation should occur followed by a series of growth events to produce a developable cluster. Thus, nucleation must not be competitive with growth. This is usually the case at low irradiance (exposure times of 100's of seconds or longer) where electrons are being produced very slowly and the rate of nucleation, given by Eq. (45), is zero most of the time. This ideal is also approached at intermediate irradiances (exposure times from 0.1 to 10 s). However, at high irradiance (ms or shorter exposure times), the rate of nucleation can be comparable to that of growth, provided that the nucleation efficiency is reasonably high (say, greater than about 0.2). Under these conditions and assuming a minimum developable size greater than three atoms, the dispersity inefficiency will cause a significant loss of sensitivity. This leads to high-irradiance reciprocity failure which will be discussed in a later section.

Computer Simulation

In its general form, the nucleation-and-growth model is too complicated for analytical study, but computer simulation is an alternative way to study the model's predictions. Simulation forms a bridge between the experimental and the theoretical worlds and gives us insight into the ways the physical properties of the silver halide grain determine its photographic properties.

The simulation is based on the Monte Carlo technique (117,125,126). It is assumed that events occur at random, but the sequence is weighted by the relative probabilities of the events. The rates of the various events are found from Eqs. (45–47) plus an additional one for light absorption λ_a:

$$\lambda_a = kI \tag{63}$$

where k is a constant and I is the irradiance. To derive appropriate values for the rates of the various processes, we need values for the rate constants as well, Eqs. (58–60), and the relative values of these will in turn depend on η_n and ω. Once the appropriate rates are available, they are normalized by the total rate λ_{tot} (sum of all the rates) to form relative probabilities. Then a uniform random number r_1 is generated, and the next event u to occur is determined by the following inequality:

$$\sum_{i=1}^{u-1} \frac{\lambda_i}{\lambda_{\text{tot}}} < r_1 < \sum_{i=1}^{u} \frac{\lambda_i}{\lambda_{\text{tot}}} \tag{64}$$

Then, a second random number r_2 is generated and used to determine the time between events τ using the equation

$$\tau = \frac{1}{\lambda_{\text{tot}}} \ln\left(\frac{1}{r_2}\right) \tag{65}$$

The rates are then updated by changing the values for e, h, Ag_n in accordance with the event that just occurred, and the process is repeated until the specific exposure time is reached, whereupon $I = 0$ and the process continues until $\lambda_{\text{tot}} = 0$. At this point, the simulation for one grain is complete. But, because the events are chosen at random and because the experimental data for comparison are the mean responses of a large ensemble of grains, the entire process must be repeated for each of many grains.

By surveying the outcome for this ensemble of grains, we can determine the fraction of grains F made developable for a given minimum developable size for a particular exposure level. To simulate the entire response curve, we need to increase the number of absorbed photons/grain and repeat the entire process. This yields an $F - \log P$ curve where P is the mean number of absorbed photons per grain. Finally, if we add increments to the irradiance and decrease the exposure time in steps repeatedly, while keeping the total exposure constant, we can simulate the reciprocity failure characteristics.

Quantum Sensitivity

In the absence of any other inefficiency besides recombination, a simple analytical model (143) can be used to derive an equation for the mean number of absorbed photons E_N required to produce a minimum developable size N:

$$E_N = N + \frac{2\omega}{\eta_n} + \omega \sum, \tag{66}$$

where $\sum$ is the number of holes accumulated during the growth stage, i.e.,

$$\sum = \sum_{h=3}^{N} h. \tag{67}$$

The first term in Eq. (66) represents the number of absorbed photons required to create a cluster of size N. The second and third terms of Eq. (66) represent the number of absorbed photons, and therefore electrons, that are lost to recombination during nucleation and growth, respectively. E_N is equivalent to the quantum sensitivity (QS) of the emulsion which is the mean number of absorbed photons/grain required to make 50% of the grains developable. Note that this definition means that *lower* QS values lead to *higher* efficiency. A situation which will produce the maximum efficiency is $\eta_n = \omega = 1$, and Table 7 summarizes the QS values for this case for different N values.

An assumption in deriving Eq. (66) is that the quantum yield of carrier production by light absorption is unity. Every absorbed photon generates a conduction band electron, so that any deviation of QS from the ideal value is due to recombination inefficiency. This assumption is supported by measurements on AgBr (144), which show that the quantum yield is greater than 0.9, and other work on AgCl, which indicates even higher values (145).

An interesting exercise that illustrates the role of recombination in reducing efficiency can be done assuming $\eta_n = \omega = 1$. Under these conditions, the rate constants for nucleation, growth, and recombination [Eqs. (58–60)] are all the same, and the relative rates of these processes [Eqs. (45–47)] are simply determined by the concentrations of electrons, holes, and silver clusters. It is also assumed that holes are immediately trapped upon creation. Imagine that one silver atom has been produced and that a second photon absorption has just produced a conduction-band electron and a trapped hole. Then, the situation would look like the top panel of Fig. 40, where the second hole is the companion to the electron that led to the formation of the silver atom. Note that the electron can be removed from the conduction band either by trapping at the silver atom to produce Ag_2 (after addition of an interstitial silver ion) or recombination with either of the two trapped holes. As indicated before, these three events

Table 7. Quantum Sensitivities Calculated from Eq. (66)

N	$(2\,\omega)/\eta_n$	$\omega\Sigma$	QS, absorbed photons/grain
2	2	0	4
3	2	3	8
4	2	7	13
5	2	12	19

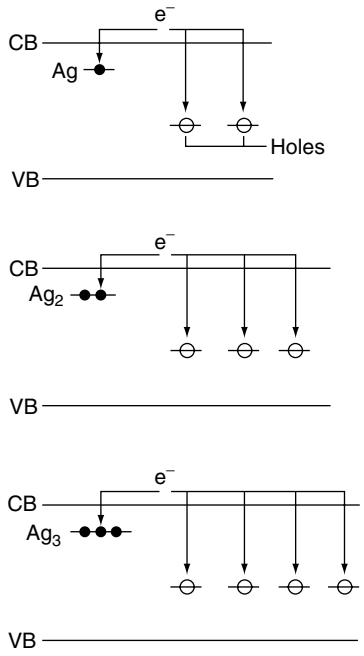


Figure 40. Schematic showing the pathways available to a conduction-band electron at various stages of latent-image formation.

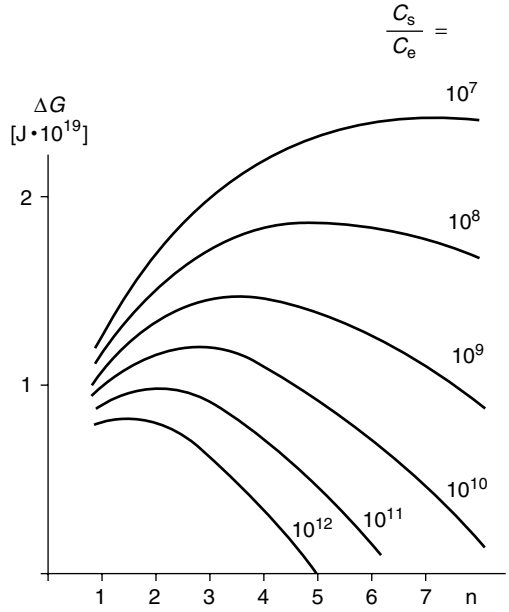


Figure 41. Free energy vs. cluster size diagram based on the phase-formation model. Labels on curves give supersaturation values. (From Ref. 112. Reprinted with permission of The Society for Imaging Science and Technology, sole copyright owners of the *Journal of Imaging Science and Technology*.)

have equal probability, and therefore recombination is twice as likely as nucleation.

After a number of absorptions, successful nucleation occurs. This brings us to the middle panel of Fig. 40, where after absorption of an additional photon, there are one Ag_2 center and three trapped holes. Now, recombination is three times as likely as growth. After a number of tries, successful growth occurs and we move to the bottom panel of Fig. 40. Here, we see that after an additional photon absorption, recombination is four times as likely as growth. By this point, the pattern has become obvious.

If we assume that development can be catalyzed by a four-atom cluster, then the total number of electrons required is four to make Ag_4, plus an average of two electrons lost to recombination during nucleation, plus an average of three electrons lost to recombination in growing from Ag_2 to Ag_3, plus an average of four electrons lost to recombination during growth from Ag_3 to Ag_4, a total of 13 electrons. This means that the QS for this case is 13 absorbed photons/grain, consistent with the results summarized in the third row of Table 7. Furthermore, the schematics in Fig. 40 clearly indicate where the electrons are being lost and that the efficiency of latent-image formation decreases as the cluster size required for development increases. This is a feature of the photographic process that is well known experimentally (134,146).

Other Models of Latent-Image Formation

Moisar and Granzer favored a phase formation model of latent-image formation (110–112). Their approach uses the Gibbs–Thomson treatment of forming a new phase under supersaturated conditions [see Eq. (9)], as shown in Fig. 41. It is assumed that a solute/solvent system applies to silver cluster formation in AgBr, where AgBr is the solvent and silver in the form of interstitial silver ions and electrons is the solute. Exposure produces a nonequilibrium concentration of electrons and holes. Electrons form silver atoms, and holes attack silver so that the system can return to equilibrium. A hole removal pathway must be available to produce a stable photoproduct. To calculate the supersaturation, it is necessary to know the number of electrons and holes in the material before exposure. This was calculated using equations from semiconductor physics which lead to a value of $4.5 \times 10^{-15}/m^3$ when the band gap of AgBr is used. If the generation rate of electrons and holes during exposure is 10^0–10^3/grain-s and the lifetime of the electron is 10^{-7} s, then the concentration is 10^{-7} to $10^{-4}/m^3$. This leads to a supersaturation of about 10^8 to 10^{11}.

As seen in Fig. 41, these supersaturation values lead to unstable small clusters (left of curve peak) and stable large clusters (right of curve peak). Furthermore, the peak position along the x axis is a function of the supersaturation, which could be correlated with the exposure irradiance. Atomically, the small clusters are attacked by holes and decrease in size, whereas the larger clusters are on the other side of the peak where they grow in size because they are not attacked by holes.

The main criticism of the phase-formation model is that an average value of supersaturation is used, but what is really needed is the instantaneous supersaturation (7). If two electrons are generated in a $10\,\mu m^3$ crystal, then the instantaneous supersaturation is about 10^{17}. A supersaturation of this magnitude would produce a curve in Fig. 41 that would simply have a maximum at $n = 1$ and decrease at larger n. Thus, all silver centers,

including the silver atom, would be stable and grow. There would be no threshold size for stability. Additionally, an equilibrium solubility of $10^{-15}/\mu m^3$ does not make sense physically because it implies that most grains will have zero electrons. The first photon absorbed will produce a supersaturation of infinity! The phase-formation model also rules out the existence of any latent subimage because it would dissolve in the host AgBr phase. Finally, the Gibbs–Thomson theory does not deal with the tendency of the silver clusters to react with holes. Thermodynamically, all cluster sizes should react.

Another model of latent-image formation has been proposed by Mitchell (101–104). Silver atoms in this model form by simultaneously trapping electrons and interstitial silver ions at a trapping site. The resulting silver atom is thermally unstable and is not a trap for electrons. Ag_2 forms by the simultaneous capture of an electron and an interstitial silver ion at the silver atom. Because Ag_2 is not stable, the Ag_3 cluster forms similarly to Ag_2, except, of course, now it is at the Ag_2 site.

Ag_3 is the cluster of minimum size that can extract a silver ion from a neighboring lattice position to form Ag_4^+, which is a stable cluster and, by virtue of its positive charge, is a good trap for electrons. Mitchell calls it a "concentration speck." It is considered that it has a very long range effect on electrons because of its charge, and provides a large trap depth. Once this stage of latent-image formation is reached, the silver cluster grows by alternate addition of electrons and lattice silver ions. The main inefficiency in this model is the attack of holes on silver atoms and dimers, the positive charge on Ag_4^+ and larger clusters preventing such a regression reaction. Recombination between electrons and holes is ignored.

There are several criticisms of the Mitchell model. For example, it assumes that no traps are present due to lattice defects, yet in emulsion grains there is a large difference between the mobility measurements of electrons that are designed to include the time spent in traps and those that are designed just to measure the mobility between trapping events (147). Such a difference is not found in large crystals. This difference is best explained by the large surface-area-to-volume ratio in emulsion grains and the inherent trapping sites associated with charged surface defects. Quantum mechanical treatments of the trapping problem also lead to the type of binding energies and trap radii discussed earlier using the hydrogenic model (148–150). The inability of Ag, Ag_2, and Ag_3 to act as traps is based on the assumption that the electron affinity of the free clusters is pertinent. But a large downward shift of energy levels occurs when the substrate effect is considered (139,148–150).

Furthermore, the large effective capture radius attributed to the concentration speck is untenable in view of the large dielectric constant of silver halides. Because the Ag, Ag_2, and Ag_3 form at uncharged sites in the Mitchell model, they should have a small cross section for capturing holes. The known inefficiency of latent-image formation is difficult to explain in these terms. Reciprocity failure, an inherent inefficiency in the photographic process, is also largely ignored by the Mitchell model.

EFFICIENCY AND SPEED

Quantum Sensitivity Measurement

In the previous section, we introduced the concept of quantum sensitivity (QS) in connection with the nucleation-and-growth model of latent-image formation. This is a property that can also be measured experimentally. To accomplish this, a transformation of the characteristic curve of the emulsion from $D-\log E$ to $F-\log P$ space must be possible. In monodisperse emulsions, this is relatively straightforward, and the $\log E$ transformation can be accomplished by measuring the appropriate physical properties of the coating and emulsion that will be described shortly. To convert the density axis into fraction of grains developable, we rely on the Nutting equation which says that density is proportional to the number of developed grains multiplied by their average projective area (151). Therefore, normalization of the D values produces the required F values. This normalization procedure is expected to work well for monodisperse grains and development in which the entire grain is converted to silver (134).

A further assumption in QS measurements is that all of the grains in the emulsion coating have equal access to the incoming photon flux. This assumption can be approximated only in practice because a finite layer thickness is required to achieve densities that can be measured accurately. Under these conditions, there is an inevitable decrease in irradiance depthwise through the coating, although light scattering by the grains (152) tends to diminish the magnitude of the effect that is predicted by the Bouguer–Lambert Law (153–155). In practice, a compromise is struck between accurate density measurement and uniform exposure by coating the emulsion to give a maximum density of about one. As an example, an emulsion composed of 0.5-μm diameter AgBr grains and exposed at 400 nm would have a decrease of light irradiance of about 25% between the top and bottom of the coating, when coated to achieve a D_{max} of one (156).

Using the preceding assumptions and approximations, the following equation (156) is used to calculate QS from experimental data:

$$\mathrm{QS} = \frac{\left(\frac{E}{E_c}\right) AITV\rho}{SW}, \tag{68}$$

where E is the exposure at the reference density point on the $D-\log E$ curve, E_c is the exposure at the clear step of the step tablet (the point of maximum exposure), A is the asorptance of the coating, I is the irradiance at the film plane, T is the exposure time, V is the grain volume, ρ is the density of silver halide, S is the silver coating weight in the emulsion layer, and W is a factor for converting silver coating weight to silver halide coating weight. A propagation of error analysis using this equation has shown that typical uncertainty (two sigma) in the measured QS is about 15% (156). An example with typical values is shown in Fig. 42.

Measuring QS in polydisperse emulsions is considerably more involved. The transformation from density to

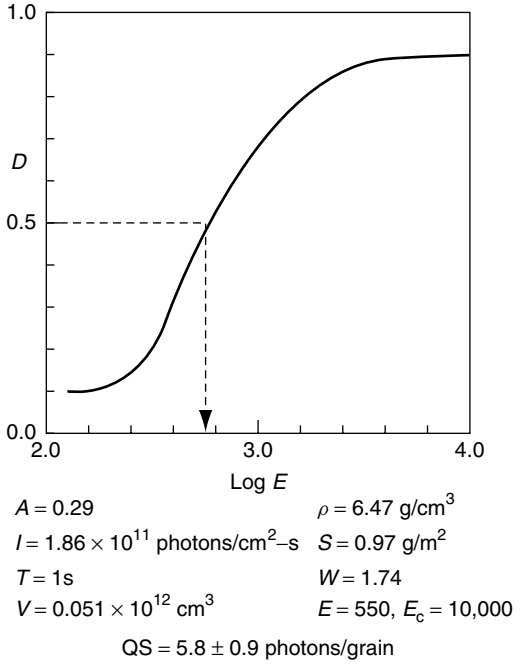


Figure 42. A sample calculation of quantum sensitivity for the indicated experimental data using Eq. (68).

fraction developable using the Nutting equation is no longer possible because different size grains make different contributions to the measured density. Likewise, transformation from log E to log P is not straightforward because different size grains have different mean absorption values. As a result, the D–log E curve is a convolution of many D–log E curves, one for each size class of grains in the emulsion.

These considerations indicate that the emulsion response must be analyzed in terms of individual grain size classes usually using some form of microscopy — either optical microscopy for large grains (>2 μm diameter) or electron microscopy for smaller grains. The fraction of grains developable can be determined directly by sizing developed and undeveloped grains or indirectly by using residual grain analysis in which only the undeveloped grains are sized and the developable fraction is inferred by reference to the unprocessed emulsion. The latter technique is more common because development often destroys any remnants of the grain before development. Once these data are available and the absorbed light is apportioned among the grains according to their size, it is a simple matter to generate the F–log P curves for each size class.

Theoretical Limits

The theoretical limits of QS were derived by Silberstein many years ago (157). Assuming no inefficiencies in latent-image formation, one latent image per grain, and a very thin coating, so that all grains have equal access to the incoming light, it is possible to derive theoretical curve shapes for various values of exposure expressed as the mean number of absorbed photons/grain for different threshold values M to induce developability by using the following equation:

$$F = 1 - e^{P}\left(1 + P + \frac{P}{2!} + \frac{P}{3!} + \cdots + \frac{P^{M-1}}{(M-1)!}\right) \qquad (69)$$

Examples of the resulting curves are shown in Fig. 43. Note that the response curve is not a step function, as might be expected if all the grains respond at the same photon threshold, because light absorption within a coating follows Poisson statistics and the abscissa in Fig. 43 has units of **mean** absorbed photons/grain.

A summary of QS data is shown in Fig. 44. Note that most of the data points fall between 10 and 30 photons/grain. The filled symbols represent special cases where the sensitization was done to eliminate all recombination. In this case, the QS is 2–3 photons/grain, which is consistent with the expected minimum developable size of three atoms for these types of emulsions (124). The remaining data show that some inefficiency still remains in commercial emulsions, indicating that there is still room for efficiency improvement in the silver halide detector.

Factors Affecting Speed

Photographic speed is a measure of the amount of light that must fall on a film to form a latent image. Speed is a function of three factors: (1) the amount of light absorbed, (2) the efficiency with which that absorbed light is used, and (3) the minimum developable size of the latent image. We have already discussed the second two factors in detail. Light absorption by AgBr grains is proportional to their size — to volume if the exposure is in the intrinsic region of absorption, to surface area if the exposure is in the region of absorption by adsorbed dye. As a result, we expect speed to increase linearly with grain size (either volume or surface area, depending on the exposure wavelength), assuming that factors (2) and (3) are independent of grain size.

The shape of the grain also can affect speed. Consider "flattening" a cubic grain into a tabular one, while maintaining its volume. The tabular grain would have a much higher surface area than the cubic grain, leading to

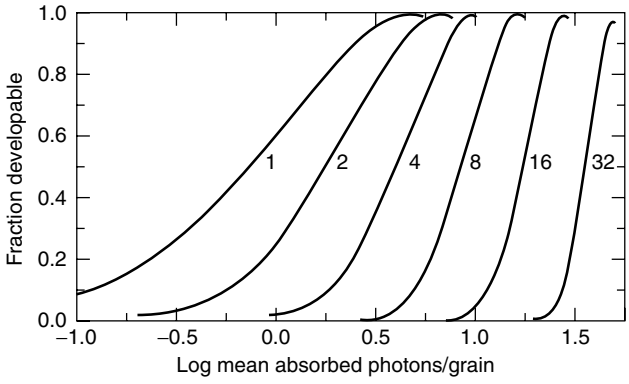


Figure 43. Theoretical response curves from the Silberstein model [Eq. (69)]. Curve labels indicate the photon threshold to reach the developable state. (Reprinted with permission of The Society for Imaging Science and Technology, sole copyright owners of the *Journal of Imaging Science and Technology*.)

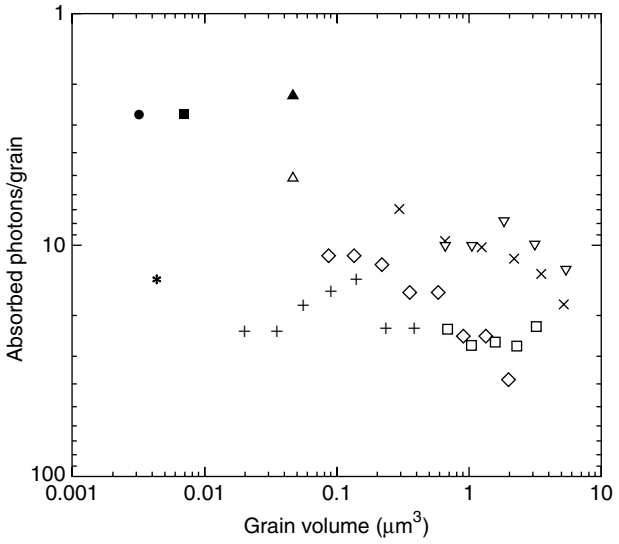


Figure 44. A summary of available QS data. The several cases with multiple data points that have the same symbol represent the different size classes in a polydisperse emulsion. Inverted triangles (high-speed X-ray emulsion); squares (high-speed negative); diamonds (medium-speed negative); and + (low-speed negative) (158); × (160); * (159); filled circle (161); filled and empty triangle (156). (Figure taken from Ref. 156. Reprinted with permission of The Society for Imaging Science and Technology, sole copyright owners of the *Journal of Imaging Science and Technology*.)

more dye adsorbed on the surface. Thus, light absorption at wavelengths longer than that absorbed by the silver halide grain would be enhanced which would translate into higher speed, assuming that the efficiency of latent-image formation is independent of shape. For this reason, tabular grains are used in most systems designed for image capture.

Experimental studies have consistently shown that speed is proportional to grain size only over a limited size range (162–165), as shown schematically in Fig. 45. At large grain sizes, the measured speed is less than that predicted by the expected relationship between speed and size based only on the absorption factor. This means that either factors (2) or (3) or both are grain-size dependent. There is no indication that (3) is grain-size dependent, so the key factor in explaining this discrepancy between theory and experiment is that the efficiency of latent-image formation decreases at large grain sizes.

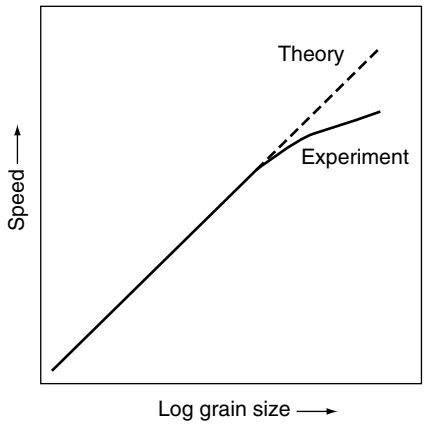


Figure 45. Speed vs. grain size relationship typical for photographic materials compared with the theoretical relationship.

There is not general agreement on what grain-size dependent inefficiency is causing this behavior. Some think that the effective range over which the electron can diffuse is limited, so that different parts of a large grain compete with each other for the available electrons (160,162–164,166,167). However, exactly what process limits the range of the electron is not known. Others have suggested that internal image formation is more likely in large grains, making it more difficult to form a surface image efficiently (168). Other factors might involve enhanced recombination because internal lattice defects, assumed to be more prevalent in larger grains, provide additional recombination centers or impurities that act as irreversible electron traps (165). Finally, because the experimental data pertain to chemically sensitized grains, a fog increase induced by sensitization reactions (to be discussed later) may prevent optimization of sensitization. Limited experimental data suggest that larger grains are more prone to fog formation during chemical sensitization (162).

RECIPROCITY FAILURE

A general law of photochemical reactions proposed by Bunsen and Roscoe indicates that the amount of photoproduct depends only on the total exposure, not on the individual components (irradiance and time) (169). Because of the reciprocal relationship between irradiance and time in determining a constant value of exposure and therefore, a photoproduct, this law is known as the reciprocity law (105). Photochemical systems that deviate from this law are said to "fail the reciprocity law," or more tersely, to exhibit "reciprocity failure."

In photography, density is proportional to the amount of photoproduct. Long time, low irradiance exposures are expected to give the same density as short time, high irradiance exposures if the reciprocity law is obeyed. Deviations from the reciprocity law are quite common in silver halides because the amount of photoproduct, that is, the fraction of developable grains, depends on the rate of arrival of photons, in addition to the total number of photons.

There are many different ways to plot experimental data to show whether the reciprocity law is obeyed. One example is shown in Fig. 46 in which speed is plotted against the logarithm of exposure time. Implicit in this plot is that irradiance varies reciprocally with exposure time, so that the total number of photons incident on the film remains constant. If the reciprocity law were obeyed, the speed should be constant with exposure time, and the plot should simply be a horizontal line. The example in Fig. 46 shows deviations from ideal behavior at both long (low-irradiance reciprocity failure, or LIRF) and short (high-irradiance reciprocity failure, or HIRF) exposure times.

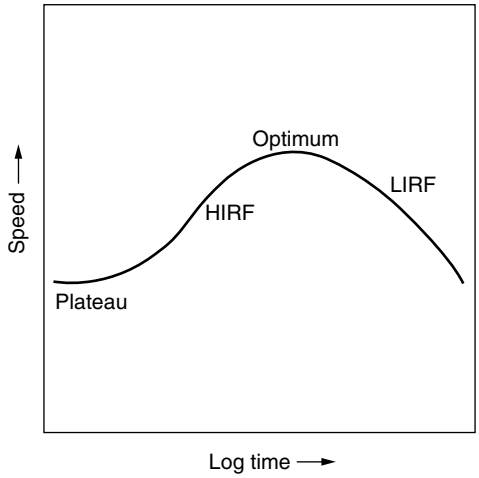


Figure 46. Reciprocity-failure plot showing LIRF, HIRF and the high-irradiance plateau. The optimum position indicates the exposure time that leads to maximum speed.

Low-Irradiance Reciprocity Failure

The appearance of LIRF at low irradiance indicates instability. The photoproduct of some initial number of absorbed photons fades unless there are at least S successful absorptions within the fading period. Analysis shows that the ultimate slope of the LIRF curve is $-(S-1)$, where S is the number of absorptions required for stability (170). Comparison with experimental reciprocity data indicates that the $S = 2$ theoretical curve gives the best fit of experimental data. One concludes from this result that the two-atom clusters are the smallest stable size and that silver atoms are unstable. These ideas have been incorporated into the nucleation-and-growth model discussed earlier.

At low irradiance (minutes to hours), the photon arrival rate will be low. As a result, the silver atom goes through many formative and decay reactions, leading to inefficiency in the nucleation stage of latent-image formation. The decay of silver atoms is not in itself a problem because the released electron could simply be trapped elsewhere in the grain and the atom re-forms. This cycle could be repeated many times until a second electron becomes available through photon absorption, making the nucleation event possible [recall Eq. (45)]. However, while the released electron is in the CB, it is also possible that it will encounter a trapped hole and undergo recombination. Emulsions that have been sensitized so as to destroy holes efficiently show virtually no LIRF (156,171–173).

High-Irradiance Reciprocity Failure

HIRF can appear at high irradiance (ms to ns exposures). High-irradiance exposures in chemically sensitized emulsions, where the nucleation efficiency is enhanced relative to unsensitized emulsions, lead to the formation of many small silver clusters on the emulsion grain. Unless the minimum developable size is small, such grains will not develop, and speed is decreased relative to intermediate irradiance conditions where the photon arrival rate is lower and only one or a few silver clusters are formed. The inefficiency at work here is the dispersity inefficiency discussed earlier. Because of the large number of electrons generated by high-irradiance exposures in a very short time, the nucleation rate is enhanced relative to the growth rate [Eq. (45) vs. (46)] and leads to multiple cluster formation and inefficient growth.

As might be expected, the degree of HIRF depends strongly on the minimum developable size detected and the nucleation efficiency of the emulsion (124). If two-atom centers can be detected by the developer, then there would be no HIRF because the smallest stable size would be detectable and any additional clusters formed would be inconsequential. If a three-atom center constitutes the minimum developable size, then there could be some competition between nucleation and growth. However, we saw earlier that the dispersity inefficiency is negligible in this case and HIRF would be undetectable. As the minimum developable size increases beyond three atoms, the degree of the dispersity inefficiency increases (141), and the degree of HIRF becomes increasingly greater.

Nucleation efficiency comes into play because this factor determines the degree of competition between nucleation and growth for high-irradiance exposures. At low nucleation efficiency, there is little chance of a second nucleation after the first at any irradiance level, and all the silver is collected in one site. As nucleation efficiency increases, nucleation and growth become more competitive, eventually leading to more than one nucleation per grain at high irradiance and dispersity. As we discussed earlier, the nucleation stage is considered inefficient, so that there really should be no dispersity. However, in the next section, we will see that chemical sensitization improves the efficiency of nucleation, which will make HIRF observable, depending on the minimum developable size N.

The connection between HIRF, the minimum size, and nucleation efficiency can be made more quantitative. When there is a high-irradiance plateau, such as in Fig. 46, the magnitude of HIRF δ can be defined as the speed difference between the high-irradiance plateau and the optimum irradiance. Figure 47 shows how δ, computed by computer simulation based on the nucleation-and-growth model, depends on N and η_n (124). As expected from the previous discussions, $\delta = 0$ for $N = 2$, and almost so for $N = 3$. But for $N > 3$, δ becomes progressively greater as both N and η_n increase.

It may be thought that recombination might be the cause of HIRF because high-irradiance exposures lead to large concentrations of electrons and holes. However, Eqs. (45) and (47) show that both nucleation and recombination have a second-order dependence on carrier density. Thus, it is generally believed that competition between nucleation and growth is the cause of HIRF (105). However, Eq. (46) shows that growth has a first-order dependence on carrier density, indicating that recombination will dominate over growth at high irradiance. So, the extent of recombination will influence the magnitude of HIRF. This can be seen in Fig. 47 where an increase in the recombination index ω produces higher δ values.

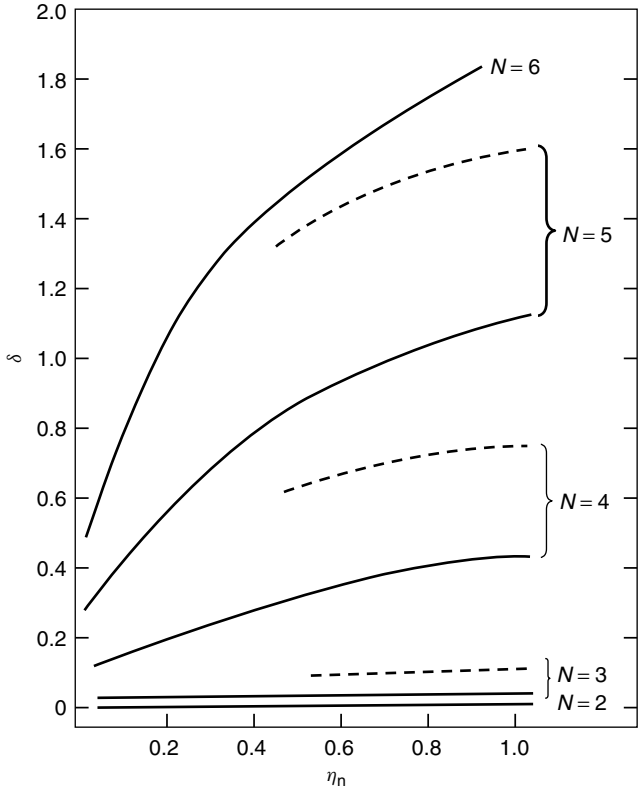


Figure 47. The dependence of the degree of HIRF δ on nucleation efficiency η_n and minimum developable size N. Solid line, $\omega = 1$; dashed line, $\omega = 4$. (From Ref. 124. Reprinted with permission of The Society for Imaging Science and Technology, sole copyright owners of the *Journal of Imaging Science and Technology*.)

Double-Exposure Experiments

Now that we know something about the mechanistic causes of LIRF and HIRF, we can discuss double-exposure experiments (118–123). These experiments, summarized in Fig. 48, involve reciprocity studies but use either a uniform pre- or postexposure in conjunction with the usual step-tablet exposure. In the absence of limitations due to exposure-induced fog, these experiments can eliminate LIRF by a short, high-irradiance preexposure and eliminate HIRF by a long, low-irradiance postexposure. In the less than ideal case when the pre- or postexposure creates fog centers, there is still a substantial reduction in the degree of LIRF and HIRF. These results indicate that there are two distinct stages in latent-image formation that have different inefficiencies in each stage and are one of the foundation experiments that led to applying the nucleation-and-growth concept to latent-image formation.

The results of the double-exposure experiments can easily be explained by the nucleation-and-growth model. At low irradiance, nucleation is inefficient relative to growth. Thus, a low-irradiance uniform exposure following a high-irradiance step-tablet exposure will primarily provide electrons that can be used to enlarge a latent subimage to a developable size, with minimal formation of new clusters (fog centers). This procedure, known as "light latensification," will eliminate HIRF. At high irradiance, nucleation is efficient, whereas growth is not. Therefore,

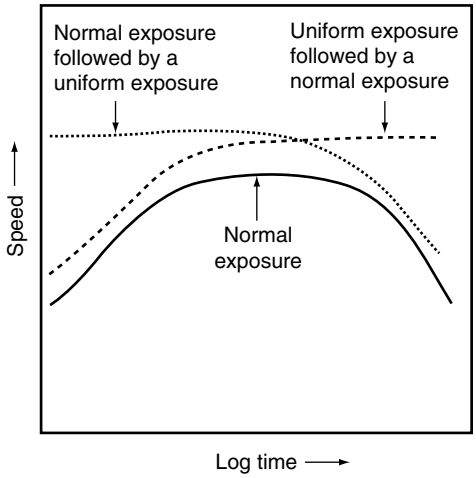


Figure 48. An illustration of the results of double-exposure experiments using a reciprocity-failure plot. "Normal exposure" means that a step tablet was used. When the uniform exposure follows the normal exposure, it has low irradiance, but when it precedes the normal exposure, it has high irradiance.

a uniform preexposure at high irradiance will produce many small subdevelopable centers. When followed by a low-irradiance step-tablet exposure, these centers will be made developable by a few growth steps. The normally inefficient nucleation at low irradiance is avoided, and LIRF is removed.

CHEMICAL SENSITIZATION

Chemical sensitization (5,7) involves treating the grain surface with low levels (100 ppm or less) of chemical reagents to increase the efficiency of latent-image formation. In some cases, chemical sensitization may have an alternative goal such as modifying the shape of the D–log E curve, reciprocity failure, or the response to changes in development time. Because chemical sensitization occurs at the grain surface, this process must necessarily occur after the emulsion has been washed but before the coating operation. In contrast, dopants, which are the subject of the next section, are placed in the emulsion reactor so that they can be incorporated into the grain during crystal growth.

There are two principal inefficiencies to be minimized in improving the efficiency of latent-image formation. The dominant one that is always present is recombination. In addition, some grains may have a tendency to form an internal latent image in the unsensitized state, but the surface modifications by chemical sensitization are intended to favor surface over internal latent-image formation.

It will be useful first to review the state of the emulsion before any chemical sensitization. Such unsensitized grains are characterized by very poor quantum sensitivity of 100 or more photons/grain (159,175). This is explained by postulating inefficient nucleation, as done in an earlier section. Using the quantum sensitivity equation [Eq. (66)], taking the recombination index $\omega = 1$, and assuming that

the minimum developable size is five atoms,

$$\mathrm{QS} = 5 + \frac{2}{\eta_n} + 12 \tag{70}$$

Now assume that η_n is 0.01, which leads to

$$\mathrm{QS} = 5 + 200 + 12 = 217 \text{ photons/grain} \tag{71}$$

Thus, the low efficiency agrees with experimental observations. Furthermore, this simple analysis shows that most of the inefficiency occurs in nucleation. This calculation assumes that recombination is the only inefficiency. Internal latent-image formation would simply increase the number of photons required.

Unsensitized emulsions display no HIRF but have significant LIRF. They have primarily one latent-image center/grain at all exposure irradiances and show little speed change when the development time is varied over a wide range. All of these effects can be explained by inefficient nucleation. Such inefficiency would prevent nucleation from competing with growth, which avoids the dispersity inefficiency and the associated HIRF. Because there is no dispersity inefficiency, there is only one silver cluster/grain, even at high irradiance. It follows that, if there is only one cluster/grain, increasing the development time will not increase speed because there are no grains that contain only subdevelopable centers that might be revealed by lengthening the development time and effectively decreasing N.

Sulfur Sensitization

Historically, sulfur sensitization is probably the first type of sensitization used for silver halides. Early gelatin materials had many impurities, including those containing sulfur. Simply heating the emulsion would produce a speed increase and a fog increase if overdone (176,177). Modern gelatin is quite pure, and sulfur sensitization is accomplished by intentionally adding reagents that contain labile sulfur atoms. Examples are thiourea derivatives and thiosulfate; we will use the latter as an example in our discussions because it is a commonly used reagent.

The schematic in Fig. 49 illustrates the practical aspects of chemical sensitization. The emulsion is in a gelled state at room temperature, so typically it is heated to 35–40 °C to produce a homogeneous liquid state. Then, the sulfur-sensitizing reagent is added after adjusting the pH and *p*Ag to the desired levels. Most sulfur-sensitizing reagents react slowly at this temperature, so the emulsion is heated to accelerate the reaction. Typical reaction temperatures are 50–70 °C. After the desired reaction time at the elevated temperature, the emulsion is cooled to 40 °C, and stabilizers are added. The emulsion is now ready for coating.

The overall reaction between thiosulfate and AgBr is shown in Fig. 50. Only the outer sulfur atom is involved in the reaction. It is thought that the reaction occurs in three stages (178–183). The first stage involves adsorption of the thiosulfate anion. The second stage involves the actual reaction to form sulfide anions

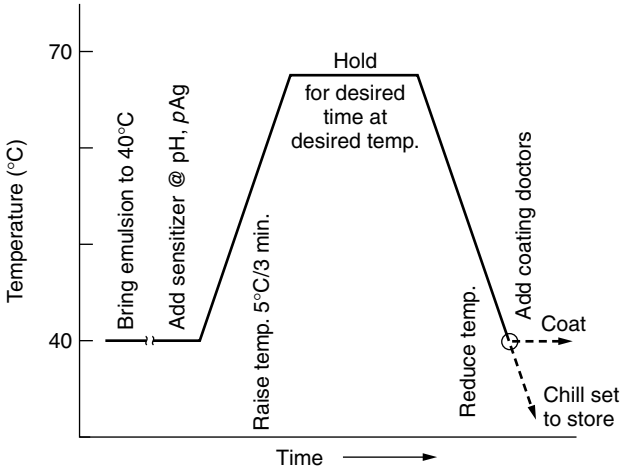


Figure 49. Schematic illustrating the practical steps involved in the chemical sensitization of an emulsion.

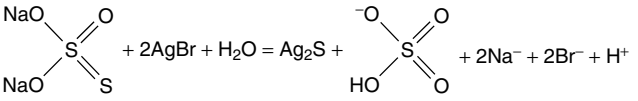


Figure 50. The chemistry involved in silver sulfide formation on a AgBr surface using sodium thiosulfate.

adsorbed on the grain surface. Formally, we denote the species Ag_2S, although the exact structure of the formed species is unknown. Experiments that limit the reaction to these first two stages result in little, if any, speed increase (178–180,184).

A third stage occurs upon normal sulfur sensitization, which involves formation of the active species. The exact nature of the active species is unknown, but it is often assigned to an aggregate of silver sulfide. That is, the site contains two or more S^{2-} which somehow improve the efficiency of nucleation and increase the sensitivity of the emulsion. Some experiments, as well as modeling, suggest that the active site contains two S^{2-} (184–186). It has been suggested that the number of these sites needed to reach maximum speed is about 1,000/μm^2 (184), based on radiotracer determination of the deposited S^{2-} and the resulting speed.

The energy levels of the sulfur-containing species on the silver halide surface obviously play a critical role in their sensitometric effects. Because bulk silver sulfide is black, we know that its band gap is much smaller than that of silver halides. We might surmise that there is a gradual decrease in the energy gap as one goes from very small centers containing just a few sulfide ions to very large centers whose structure approaches that of bulk silver sulfide. Quantum mechanical calculations on free single- and double-sulfide molecules show this effect (187). Consistent with these ideas is the knowledge that sulfur sensitization induces photographic sensitivity at longer wavelengths where silver halide does not absorb (188–191). Assuming a general pattern of decreasing energy gap as the number of sulfides increases, the electronic properties of the sulfide centers will depend on the way their vacant and filled levels align with the silver halide conduction and valence bands.

Unfortunately, direct physical measurement of these levels is not possible. Indirect measurements combining photoconductivity and dielectric loss measurements have indicated that the lowest vacant level of the sulfide centers is about 0.3 eV below the conduction band (192). Studies using luminescence modulation spectroscopy have also suggested a similar value (185,186). Photographic measurements of the temperature dependence of the previously mentioned long wavelength sensitivity also indicate trapping levels of several tenths of an eV (189–191).

Some researchers have attributed a hole trapping property to the sulfide centers (101), and the measured and calculated energy levels are consistent with this (185–187,189–191). However, it is likely that sulfide centers are formed at positive-kink or kink-like sites, and would, as such, provide a very small cross section for hole trapping.

The relationship between speed and sensitizer level or reaction time is shown schematically in Fig. 51. There is a clear optimum sensitizer level or reaction time, beyond which speed decreases. Associated with the speed decrease is an increase in fog, which in unfavorable situations begins to increase before the maximum speed is reached. These two effects are correlated. The speed loss at high sensitizer concentration or long reaction times is called "oversensitization," but we will postpone a discussion of the mechanism of this effect until after a general discussion of the sulfur-sensitization mechanism.

The source of fog upon sulfur sensitization is often attributed to overly large aggregates of Ag_2S (193,194). The proposed energy-level scheme given in Fig. 52 shows that small aggregates provide electron-trapping levels below the CB that will improve nucleation, as will be discussed shortly. However, it is hypothesized that large aggregates have even lower lying electron-accepting levels that can accept electrons from the developer, much like that of the latent image. As a result, a grain that contains such a cluster will develop without exposure and lead to an increase in fog.

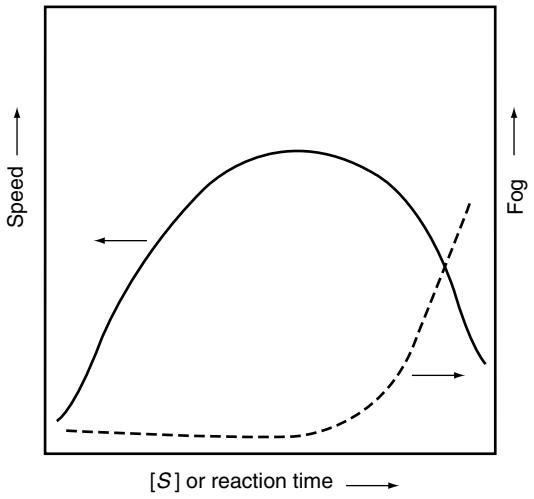


Figure 51. Schematic showing how speed and fog depend on either thiosulfate concentration at a fixed reaction time or reaction time at a fixed thiosulfate concentration.

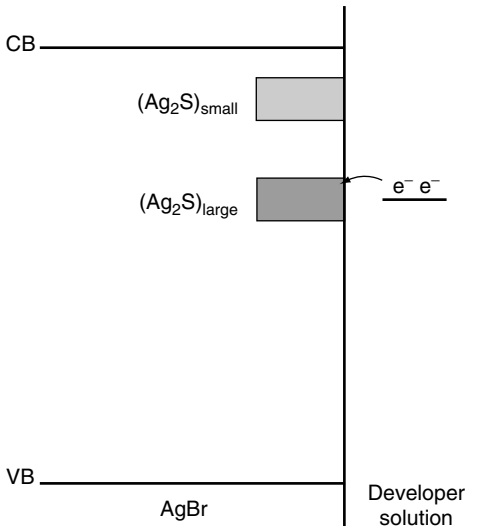


Figure 52. Energy-band diagram that shows the suggested difference between sensitization centers and fog centers from sulfur sensitization of AgBr. Shaded rectangles depict the range of electron-accepting levels. The energy level in the developer solution phase is that of the highest filled molecular orbital of the developing agent.

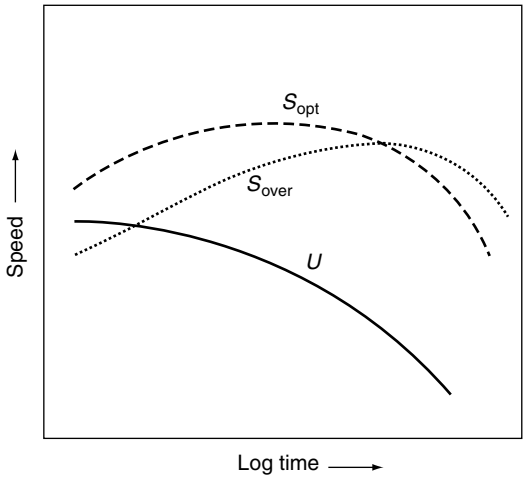


Figure 53. Changes in reciprocity failure due to sulfur sensitization. U denotes the unsensitized emulsion, S_{opt} the optimum level of sensitizing reagent, and S_{over} an excessive amount of sensitizing reagent.

Sulfur sensitization has profound effects on the reciprocity characteristics of silver halides (131,195–198). As demonstrated schematically in Fig. 53, sulfur sensitization, relative to the unsensitized emulsion, shifts the onset of LIRF to longer exposure times but also introduces HIRF. The extent of these changes increases with increasing sulfur level. The speed of sulfur-sensitized emulsions exposed for short times is very sensitive to the development conditions (time, temperature, activity of the developer) (195), as discussed earlier in connection with dispersity inefficiency.

Centers arising from sulfur sensitization direct the location of the latent-image centers (197). It is possible to

interrupt the emulsion precipitation, carry out chemical sensitization, and then to continue precipitation so as to "bury" the sulfide centers. Emulsion grains in which the sulfide centers are located internally form an internal latent image (detected with a special developer for developing an internal latent image). Emulsion grains in which the sulfide centers are located on the surface form a surface latent image. This latent-image directing property is most consistent with the idea that sulfide centers act as electron traps.

An understanding of the many sensitometric effects of sulfur sensitization can be found in the nucleation-and-growth model (127). All of the effects can be explained by enhanced electron-trapping ability of the grain. From the QS analysis of the unsensitized emulsion, we know that nucleation is very inefficient, so it is reasonable to suggest that sulfur sensitization somehow enhances this stage of latent-image formation. This enhancement is due to the creation of a deeper trapping level associated with the silver sulfide center, as shown on the right-hand side of Fig. 54 (106). Here, we see that the sulfide center increases the trap depth and thereby lowers the crossing point on the configuration coordinate diagram, so that the transition to the deep level associated with the silver atom becomes more likely. As a result, nucleation becomes more efficient. Experimental quantum sensitivity measurements yield a value of about 30 photons/grain for optimally sulfur-sensitized emulsions (131). Using Eq. (66) and a nucleation efficiency of 0.2 yields a QS of 27 photons/grain.

The effect of sulfur sensitization on reciprocity failure follows similar reasoning. As discussed earlier, the probability of the dispersity inefficiency, which is the cause of HIRF, increases with increasing nucleation efficiency. As nucleation becomes more efficient, this stage competes more favorably with growth at high irradiance and leads to a dispersed silver cluster distribution and HIRF. The effect on LIRF is understood by realizing that the onset of LIRF is proportional to the fraction of time that the electron spends free in the CB (127). Sulfur sensitization causes the electron to spend more time in the deeper traps associated with the sulfide centers and less time free in the CB, requiring a longer exposure time to observe LIRF.

When the sulfur-sensitizer concentration is increased beyond the optimum level, oversensitization results, as illustrated in Fig. 51. Recalling Eq. (43), we see that the recombination index ω depends on γ, ψ, β, and ϕ. We have already argued that the hole spends most of its time in the trapped state so $\gamma \cong 1$, which means that $\psi \ll 1$. In the unsensitized state, the electrons spend most of their time in the atom state, but the remaining time is partitioned more toward the free state than the trapped state because the intrinsic defect traps are very shallow. Thus, $\beta < \phi$ and the second term of Eq. (43) is negligible, leading to $\omega \cong 1$, as we assumed before.

As more and more traps are created by the increased sulfur-sensitizer concentration, the electron spends less and less time free in the conduction band, causing ϕ to decrease and β to increase, so that the second term in Eq. (43) eventually becomes larger than one. At this point, ω begins to increase (131,198). This latter parameter is present in both the second and third term of the quantum sensitivity equation [Eq. (66)], so we see that the efficiency decreases. This happens because now the second pathway for recombination involving free holes and trapped electrons, whose extent is determined by the second term of Eq. (43), has become important. Thus, the optimum level of sensitizer or optimum sensitization time that leads to maximum speed corresponds to the minimization of the rate of both recombination pathways.

These considerations show why an optimum sensitizer level exists. Increasing sensitizer concentration leads to more traps or deeper traps or some combination of the two. As a result, nucleation efficiency increases and quantum sensitivity improves. However, at some sensitizer concentration, the parameter ω may start to increase, leading to poorer efficiency. The possible speed versus sensitizer level behavior is shown in Fig. 55. C corresponds to the case in which ω remains at unity for any sensitizer level, whereas B corresponds to the optimum sensitization condition in which η_n and ω are both one at the sensitizer level that leads to the maximum sensitivity point. When ω increases before

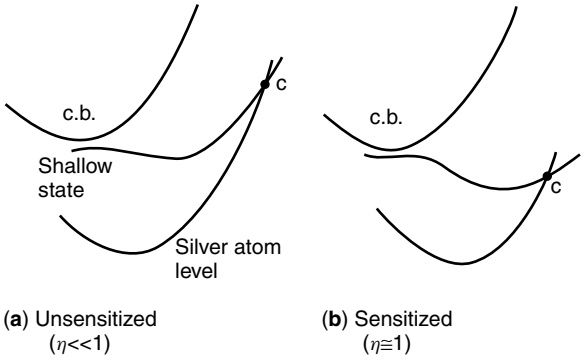


Figure 54. Configuration coordinate diagram similar to Fig. 29 that illustrates the effect of sulfur sensitization on electron trapping at a silver atom (nucleation). Note the lower crossing point C for sulfur sensitization. (From Ref. 106. Reprinted with permission from World Scientific Publishing Co. Pte. Ltd.)

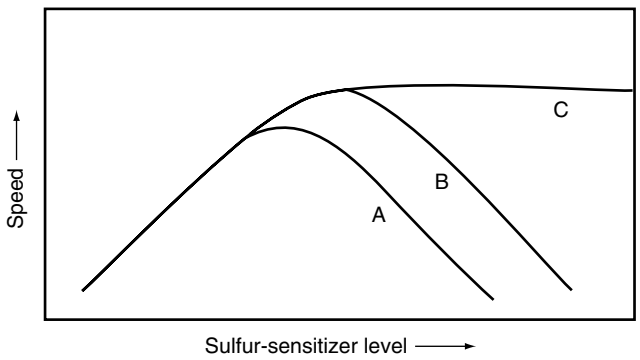


Figure 55. Hypothetical speed-sensitizer concentration relationships. See the text for an explanation of the labels. (From Ref. 198. Reprinted with permission of The Society for Imaging Science and Technology, sole copyright owners of the *Journal of Imaging Science and Technology*.)

η_n reaches one, A results. The quantum sensitivities at the highest speed values for B and C are those already discussed in Table 7. Note, however, that these quantum sensitivity values are greater than those expected in the fully efficient case considered by Silberstein in which $QS = N$. Evidently there is a residual recombination inefficiency that cannot be removed by optimizing the sulfur sensitization (131,198).

Sulfur-Plus-Gold Sensitization

Most commercial silver halide materials are chemically sensitized with gold- and sulfur-containing reagents. Typical gold reagents are $KAuCl_4$ and $Na_3Au(S_2O_3)_2$. In the first reagent, the gold is in the +3 valence state, whereas in the second, it is in the +1 valence state. Nevertheless, the gold in the sensitizer center is in the +1 oxidation state, indicating that Au^{3+} is reduced during sensitization, most likely by components of the gelatin, particularly methionine (199,200).

Studies of heavily sulfur-sensitized emulsions that have been treated with gold solutions show a change in the composition of the surface species where Au^+ replaces Ag^+ (201). In addition, the reflection spectra take on the characteristics of AgAuS, as opposed to Ag_2S. It has been suggested that the sensitizer center contains metallic gold, but the preexposure treatment of emulsion coatings with bleaches that are known to oxidize metallic gold produces very little change in photographic sensitivity (202). Chemical oxidation studies of the latent image formed in sulfur-plus-gold sensitized emulsions, however, suggest that the latent image is composed partly of gold atoms (203).

In some cases, the presence of gold leads to an antifogging effect (201,204). Apparently, overly large $(Ag_2S)_n$ that act as fog centers during development are converted to $(AgAuS)_n$ centers that do not. Two sources of fog have been suggested for sulfur-plus-gold sensitization (205). The first is due to overly large $(AgAuS)_n$, similar to the situation with Ag_2S. The second is attributed to those centers that contain metallic nuclei composed of silver, gold, or silver–gold.

Many of the sensitometric effects of sulfur-plus-gold (S + Au) sensitization are similar to those found with sulfur (S) sensitization, except that S + Au sensitization routinely results in a higher speed than S sensitization (159). Furthermore, S + Au sensitization leads to little, if any, HIRF (202). This effect is ascribed to the incorporation of metallic gold in the latent image and a corresponding change in its catalytic activity (124). Indirect evidence from deposited metal clusters on mica has shown that development is initiated by two gold atoms versus four silver atoms (129). Additionally, the ionization potentials of silver versus gold suggest that the presence of gold should lower the electron-accepting level of the latent image and make smaller clusters more developable. Finally, molecular orbital calculations focused on small clusters have also predicted a lowering of energy levels when gold is present (139,206). All of this suggests that one of the effects of gold is to reduce the minimum developable size, which can be called a "development effect."

The development effect does not fully explain the increase in speed for S + Au sensitization versus S sensitization. The development effect can be simulated by treating exposed sulfur-sensitized emulsion coatings with a gold solution, which allows incorporating metallic gold into the latent image. Large speed increases and the reduction or elimination of HIRF are observed, but the speed is still slower than S + Au sensitized emulsions (207). This suggests that gold exerts an electronic effect in addition to its development effect when it is present in the sensitizer center during exposure. In the nucleation-and-growth model, this electronic effect is attributed to an increase in nucleation efficiency over that possible by S sensitization alone.

The magnitude of the development and electronic effects can be illustrated by the sensitivity equation [Eq. (66)]. Assuming that $\omega = 1$, we arrive at the data in Table 8. In the first row, we calculate the QS of a S + Au sensitized emulsion where $\eta_n = 1$ and $N = 3$ atoms, leading to QS = 8 absorbed photons/grain. In the second row, we calculate the QS for a S-sensitized emulsion where $\eta_n = 0.2$, but $N = 3$ atoms, so as to mimic just the electronic effect of gold, leading to QS = 16 absorbed photons/grain. Finally, the last row corresponds to a S-sensitized emulsion where $\eta_n = 0.2$ and $N = 5$ atoms, leading to QS = 27 absorbed photons/grain. The difference between the first and second rows is the electronic effect, which is 0.30 log E, whereas the difference between the second and last Rows indicates the development effect, which is 0.23 log E. So, the two effects have similar magnitudes. These results are for the optimum position on the reciprocity plot (see Fig. 46). At high irradiance, the development effect is much greater than the electronic effect because of HIRF in the sulfur-sensitized emulsions.

A minimum developable size of three atoms was used in the preceding calculations because this size will eliminate HIRF according to the results from the nucleation-and-growth model given in Fig. 47. As discussed before, experimentally, S + Au sensitized emulsions show little, if any, HIRF. However, the data in Fig. 47 also show that a minimum size of two atoms eliminates HIRF, so why not choose this as the minimum size? The reason is that the two-atom cluster is the smallest stable cluster and if this were the minimum developable size, there would be no latent subimage, contrary to experimental data (130).

An atomic explanation is not available for the increase in nucleation efficiency when gold is part of the sensitization. Measurements of trap depths show that they are not deeper than when sulfur sensitization is used alone (190). Possibly, the atom state is now a gold atom that might provide unoccupied levels between the trapping

Table 8. Calculated QS for Different Sensitizations and Minimum Developable Sizes

Sensitization	η_n	N	$2/\eta_n$	Σ	QS, absorbed photons/grain
S + Au	1.0	3	2	3	8
S with $N = 3$	0.2	3	10	3	16
S	0.2	5	10	12	27

level and the gold 6*s* atomic orbital and lead to a higher probability of electron transition to the deep level during nucleation.

Reduction Sensitization

The species formed from this mode of chemical sensitization are silver clusters (208). Thus, typical reagents are reducing agents such as boranes, stannous chloride, and thiourea dioxide. The electronic properties of these silver clusters have been of keen interest (138,165,209–216). Do they act as latent-image precursors and trap electrons or do they react with holes and decrease in size? The answer to this question seems to depend on the concentration of sensitizing reagent used. At low and intermediate concentrations, the sensitizer centers produced can not direct latent-image location (217,218), unlike sulfur or sulfur-plus-gold sensitization. This behavior would suggest they have hole-trapping properties. But, at higher concentrations, at least some of the centers apparently possess latent-image directing properties (217,218) and behave as latent-image precursors (electron traps) (210,216).

Because at least some of these centers can be made developable by treatment with a gold solution, which converts the silver cluster to the more easily developed silver–gold cluster, they can be observed in the electron microscope after mild development (219,220). The number of silver clusters detected by this technique decreases upon increasing exposure, that is, they are photobleached, as would be expected from the suggested hole-trapping properties. Fog occurs when these silver clusters become large enough to trigger the development reaction, suggesting that their properties are the same as those of latent-image centers.

Reduction sensitization increases speed, particularly for low-irradiance exposures, so that LIRF is reduced (221). There is no HIRF introduced by this type of sensitization. Oversensitization is not usually possible with reduction sensitization. Rather, fog limits the amount of sensitizer that can be used (210–213). An interesting possibility arises when the extent of hole removal by the reduction sensitization centers is high enough that there is no recombination. Then, of course, we should expect the quantum sensitivities to be consistent with the theoretical limits (Fig. 43). Such behavior has, in fact, been observed in model systems (156,161).

An interesting aspect of chemically produced silver clusters is their ability to produce electrons upon hole trapping via the following reactions:

$$Ag_2 + h^+ \longrightarrow Ag_2^+ \tag{72}$$

$$Ag_2^+ \longrightarrow Ag + Ag_i^+ \tag{73}$$

$$Ag \longrightarrow Ag_i^+ + e^- \tag{74}$$

This mechanism has been verified experimentally (222–224). The theoretical quantum yield of conduction-band electrons is two per absorbed photon.

Given that sulfur-plus-gold and reduction sensitization have complementary functions, it might be expected that the best of both worlds could be obtained by using both. Sadly, this is not usually possible because significant fog arises when they are combined (156,225), most likely due to an interaction between the gold and the silver clusters that produces developable metal clusters. One way to prevent this might be to place the reduction sensitization inside the grain to avoid this interaction. However, it appears that, in at least some cases, the properties of internal chemically produced silver clusters are different from those created on the surface and they actually trap electrons and desensitize the surface image (226). There are also problems with stability of the reduction sensitization over time. All this means that reduction sensitization is usually not a viable option to the photographic film manufacturer.

DOPANTS

Dopants differ from chemical sensitizers in that they are introduced during the emulsion precipitation, so that they can be incorporated in desired regions of the grain. They are used to modify the intrinsic sensitometry such as reciprocity failure or contrast (227,228). Additionally, they are used to reduce the desensitization that is sometimes observed when spectral sensitizing dyes are used. The electronic effects of dopants are controlled by the atomic orbitals of the dopant but modified by the crystal potential, crystal field effects, Jahn–Teller distortions, and relaxation effects.

Many of the dopants used today are transition-metal ions that have positive valence states. These ions exist as complexes, most often of octahedral symmetry. Their chemistry is controlled by the *d* electrons, and because of their directionality, not all of the *d* orbitals have the same energy. In an octahedral complex of six negative ligands, the d_{xy}, d_{yz}, d_{zx} orbitals (the t_{2g} levels) have lower energy than the free ion because they point between the ligands, whereas the remaining *d* orbitals d_{z^2} and $d_{x^2-y^2}$ (the e_g levels) have higher energy than the free ion because they point toward the negative ligands.

When these ions (and their ligands) are incorporated into the silver halide lattice, they often have a charge (+2, +3, or +4) which is greater than the silver ion that they replace. To maintain electrical neutrality, a vacancy must also be incorporated for every unit of charge greater than one. These vacancies can associate with the transition-metal complex, in which case the symmetry of the octahedral field is destroyed (229,230) and causes distortion of the complex and a shift in energy levels. Because these dopants take part in one-electron trapping during exposure (or one-electron release in hole trapping), they undergo valence changes that, depending on the dopant, produce or eliminate paramagnetic species upon exposure. Thus, the photochemistry can often be followed by magnetic resonance techniques.

An example of the way these ionic effects can influence the electronic properties of the dopant is illustrated by $Pd(Cl)_6$, which contains the d^8 transition-metal ion Pd^{2+} (231). The dopant would, of course, take the place of a silver ion, and the six Cl^- ligands would occupy the positions normally occupied by the halide ions that surround the silver ion. The energy-level diagram for this dopant, shown in Fig. 56, indicates that it can trap

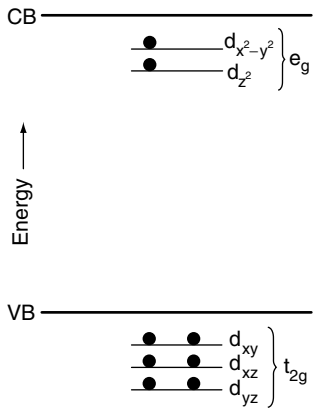


Figure 56. Hypothetical energy-level scheme for $PdCl_6$-doped AgCl without vacancy association. Braces indicate orbital energy degeneracy. Filled circles represent the orbital electrons.

electrons in one of the e_g levels. There is also no possibility of trapping holes because the t_{2g} levels are in the VB. However, because of the excess charge on the Pd ion, a vacancy may be associated with the $Pd(Cl)_6$ which would distort the complex and shift the energy levels to the new positions shown Fig. 57. Now, the dopant can no longer trap an electron, but it may trap a hole in the d_{xy} orbital. Evidence has been found for both electron-trapping and hole-trapping species that coexist in emulsion grains (231).

Ir^{3+} is a widely used dopant in photographic systems; at less than ppm concentrations it decreases both LIRF and HIRF (232,233). At higher concentrations it leads to the formation of an internal latent image located adjacent to the dopant (234). In contrast to palladium discussed before, Ir^{3+} exists as a fully vacancy-compensated center $[(IrCl_6)^{3-} \cdot 2V]^0$, where V represents a vacancy and the outer superscript indicates the overall charge on the site (227,235). During exposure, the following reactions can occur at this center:

$$[(IrCl_6)^{3-} \cdot 2V]^0 + e^- \longrightarrow [(IrCl_6)^{3-} \cdot 2V \cdot e^-]^{1-} \quad (75)$$

$$[(IrCl_6)^{3-} \cdot 2V \cdot e^-]^{1-} \longrightarrow [(IrCl_6)^{4-} \cdot 2V]^{1-} \quad (76)$$

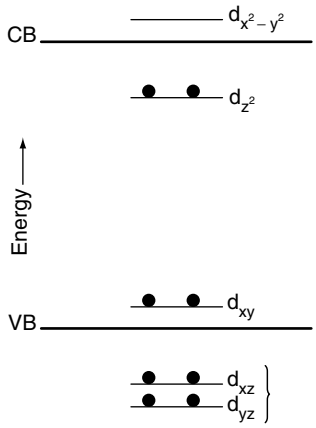


Figure 57. Effect of vacancy association on the hypothetical energy-level scheme of Fig. 56. Braces indicate orbital energy degeneracy.

$$[(IrCl_6)^{4-} \cdot 2V]^{1-} \longrightarrow [(IrCl_6)^{4-} \cdot V]^0 + [V_{free}]^{1-} \quad (77)$$

The first reaction depicts shallow electron trapping brought about by the coulombic attraction of the iridium ion, followed on a longer time scale by an electronic relaxation in which the electron becomes localized in an iridium e_g level [reaction (76)]. On an even longer time scale, ionic relaxation occurs [reaction (77)] as one of the vacancies diffuses from the complex to become a free vacancy and produces a deeply trapped electron whose a trap depth is about 0.45 eV. Atom formation occurs when an interstitial silver ion diffuses to the iridium site and the electron transfers to it. In this way, the interstitial silver ions help to reset the trap for subsequent trapping of another electron. Latent-image formation follows this initial atom formation by an alternating series of electron and interstitial silver ion captures.

$Fe(CN)_6$ which contains Fe^{2+}, a d^6 ion, is another dopant. This dopant, which can be incorporated in AgCl emulsions with all six ligands intact, is associated with one vacancy (236). The t_{2g} levels are filled and lie above the valence band, whereas the e_g levels are empty and lie within the conduction band. This is an example of a dopant that can trap a hole, after which a second vacancy can be associated with the complex. These two steps are shown in reaction (78).

$$Fe(CN)_6^{4-} \cdot V + h^+ \longrightarrow Fe(CN)_6^{3-} \cdot 2V \quad (78)$$

Although the e_g levels of this dopant are above the CB minimum and therefore it has no electron trapping levels, it is still possible to localize an electron at the dopant by virtue of its $+e$ charge with respect to the surrounding lattice. It behaves as a coulombic trap whose depth is about 65 meV (236). The CN^- ions act as ballasts to prevent diffusion of the Fe^{2+} through the lattice, as can happen with other dopants.

SPECTRAL SENSITIZATION

As discussed earlier, silver halides absorb light only in the UV or blue region of the spectrum. To enable image capture and reproduce full color scenes, it is necessary to extend the spectral sensitivity to the green and red regions of the spectrum. This is accomplished with spectral sensitization in which adsorbed dyes are used to absorb in the spectral regions where the silver halide cannot and make that photon energy available for latent-image formation (5,237). The energetics of the spectral sensitization process are illustrated in Fig. 58 for a red-absorbing dye.

Spectral sensitization is an example of photoinduced electron transfer. To be an efficient sensitizer (high quantum yield of conduction-band electrons), the energy of the lowest vacant (LV) molecular orbital of the dye must be at or above the conduction-band minimum of the silver halide, as shown in Fig. 58. Otherwise, the processes of radiative and nonradiative return to the ground state dominate. These latter processes occur on a nanosecond or

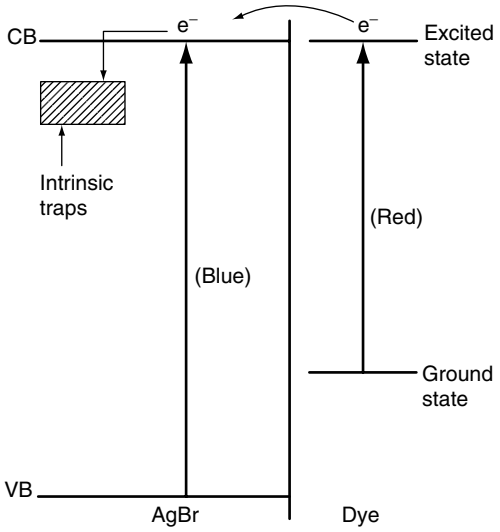


Figure 58. Illustration of the energetic requirements for spectral sensitization by a red dye. The excited state is the energy level of the lowest vacant (LV) molecular orbital of the dye, and the ground state is the energy level of the highest filled (HF) molecular orbital of the dye.

faster time scale, so electron transfer must be at least this fast to produce a useful quantum yield of electrons.

Dye Structure

Spectral sensitizing dyes are taken from the cyanine dye class (238). An example of a structure is shown in Fig. 59. The chromophore is the conjugated chain (alternating single and double bonds) between the nitrogen atoms, whose length is a major determinant of the spectral region where the dye will absorb. Cyanine dyes dominate other dye classes for spectral sensitization because (1) they have the highest extinction coefficient of almost any class of organic compounds, (2) their ionic charge promotes strong adsorption to the silver halide surface, (3) structural modifications to achieve the desired characteristics are relatively easy, and (4) they are very efficient in electron transfer to the silver halide conduction band.

Cyanine dyes are an example of molecules to which the familiar one-dimensional particle-in-a-box model from quantum mechanics can be applied. With some minor adjustments, the shift in absorption wavelength caused by increasing the length of the chromophore can be predicted by this model with surprising accuracy (239). A general rule is that each additional double bond inserted into the chromophore (making the box longer) shifts the absorption maximum about 100 nm toward a longer wavelength (238). Additional changes to the absorption maximum can be made by the choice of the ring structure on each end of the chromophore, as well as by substituents on the outer ring. Using these "knobs," it is relatively easy to design structures that absorb at any desired wavelength within the visible or near infrared spectrum.

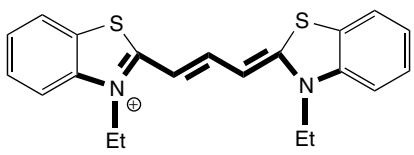


Figure 59. Typical structure of a cyanine dye. The chromophore of the dye is in bold. Hydrogen atoms not shown. Et is the $-CH_2CH_3$ group. The sulfur atom S indicates the heteroatom position in the dye structure.

The charge on the dye can be varied by modifying the alkyl groups on the nitrogen atoms with groups that have carboxylic or sulphonic acid functionalities; this leads to three possible charge states for the dye — cationic, anionic, and zwitterionic. The first two are the most commonly used. The cationic form of the dye is usually the more strongly adsorbed but is less water soluble (37). The solubility difference becomes important in materials for color photography because dyes that are less soluble in aqueous systems are more easily removed from the grain during storage by the incorporated image dye precursors needed to form a color image. The image dye precursors are organic molecules that are contained in small droplets dispersed throughout the coating. These droplets provide an organic phase where a wandering dye molecule would prefer to be. As a result, the anionic form of the dye is preferred in color imaging systems.

Dye Adsorption

The dye must be adsorbed on the silver halide surface to be effective. This property is often measured by using a technique in which the dye and emulsion are first allowed to equilibrate with each other, followed by a centrifugation step. The amount of dye in the supernatant is measured spectrophotometrically using a prior calibration of the dye in a gelatin/water solution. The amount of adsorbed dye is calculated by difference and then plotted as an adsorption isotherm, as shown in Fig. 60. Adsorption often follows the Langmuir adsorption equation, and only single monolayer adsorption is usually observed (240). Factors that can affect dye adsorption include the nature of the grain halide and morphology, the dye charge and substituents, and the environment such as pH, pAg, gelatin, and stabilizers.

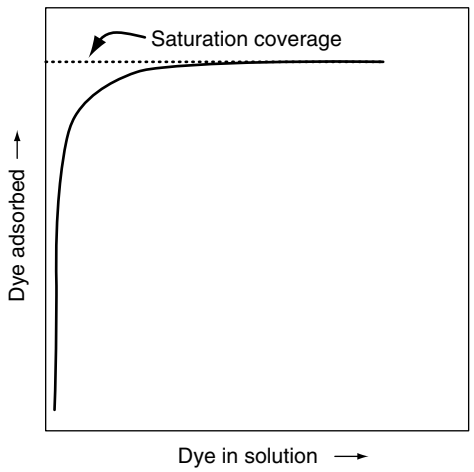


Figure 60. An adsorption isotherm typical of a cyanine dye. The saturation coverage allows determining the molecular area of the dye, provided that the surface area of the grain is known.

Driving forces for adsorption are van der Waals and coulombic forces. Dyes tend to be more strongly adsorbed on the more polarizable AgBr than on the less polarizable AgCl (241–244). Dyes whose atoms are more polarizable, particularly those in the heteroatom position (see Fig. 59), are also more strongly adsorbed. Dye adsorption strength tends to increase with increasing Br^- concentration at the surface, suggesting coulombic forces (245,246). Adsorption isotherm measurements allow determining the molecular area (see Fig. 60), and these values are considerably smaller than that expected for flat-on adsorption (245,246). The conclusion is that the dyes are adsorbed edge-on in a vertical or almost vertical orientation, where the long axis of the dye molecule is parallel to the grain surface. Furthermore, the effects of changing the heteroatom on adsorption suggest that this part of the dye molecule is near the grain surface and the N atoms of the chromophore are away from the surface. These results indicate that there is a dye–dye interaction at work that also drives adsorption (cooperative adsorption).

The effect of the crystallographic nature of the grain surface on dye adsorption is mixed (247). In some cases, the dye adsorbs equally well on both (100) and (111) surfaces. In other cases the dye has a preference for the surface on which it adsorbs. Adsorption also affects the absorption maximum of the dye and routinely shifts it about 20 nm to a longer wavelength in the absence of aggregation effects which will be discussed next. This shift is due to the refractive index change of the dye environment upon adsorption on the silver halide surface (248,249).

Dye Aggregation

Cyanine dyes also tend to form aggregates in which the individual dye molecules form a larger two-dimensional "supermolecule" (240). The structure of the aggregate can take on several different forms, as shown in Fig. 61. The ordered nature of these aggregates leads to dye–dye electronic interactions that shift the absorption maximum. These aggregates are distinguished by their so called "slip angle," as defined in Fig. 61. Variation of the slip angle changes the nature of the dye–dye interaction and leads to different spectral shifts of the absorption peak (5). Slip angles of 60° are associated with H aggregates, where the absorption maximum is shifted to shorter wavelengths, whereas slip angles of 30° or less are associated with J aggregates, in which the absorption maximum is shifted to longer wavelengths.

J aggregates are particularly interesting because their absorption band is significantly narrowed with respect to the unaggregated dye. This leads to a high extinction coefficient that makes them desirable for many photographic film applications. Factors that control the degree of J aggregation include the dye structure and the grain morphology. As might be surmised from Fig. 61, dyes whose structure leads to a high degree of planarity are more likely to aggregate (251). Similarly, dyes that have long chromophores (three or more double bonds) rarely form aggregates because such dyes have a large number of conformations that interfere with the ordered nature of the aggregate.

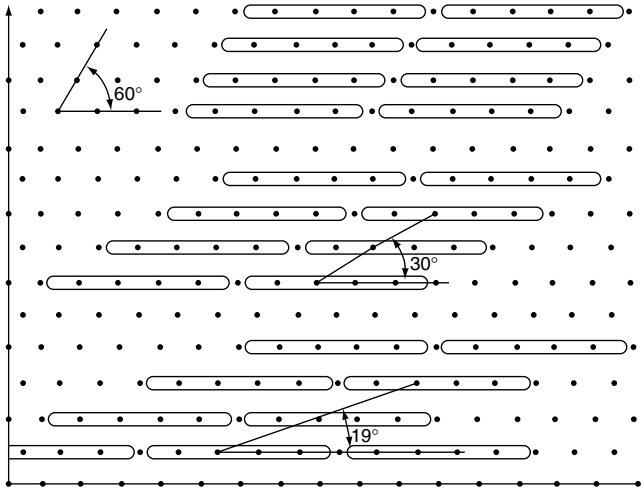


Figure 61. Arrangement of dye molecules in H and J aggregates. View is from the top of the grain, and the filled circles represent ions in the surface layer. The dye molecules are adsorbed edge-on and depicted as oblong structures. The slip angle is indicated. (From Ref. 250. Reprinted with permission of The Society for Imaging Science and Technology, sole copyright owners of the *Journal of Imaging Science and Technology*.)

Dye Energy Levels

As would be expected from Fig. 58, dye energy levels play a critical role in the efficiency of spectral sensitization. Unless the dye excited state is at or above the CB minimum, the return of the electron to the ground state dominates, and the probability of electron transfer to the CB per photon absorbed is small. Thus, it is important to be able to correlate dye energy levels with spectral sensitizing efficiency.

It is difficult to measure these energy levels directly on the adsorbed grain in a film coating, so what has been done is to measure the electrochemical properties of the dye in solution (252) and then correlate the solution redox potentials with photographic behavior. The electrochemical reduction potential E_{red} is related to the LV level of the dye, whereas the oxidation potential E_{ox} is related to its HF level. Earlier measurements of redox potentials were hampered by the irreversibility of the measurement which will influence the measured potentials in ways not easily predicted (253). More recently, a phase-selective second-harmonic ac voltammetric technique has been described which makes the measurements on a shorter time scale and avoids most of the problems of secondary reactions of the reduced or oxidized dye (253). All potentials given in this article are based on this technique, are for dyes in acetonitrile, and are measured against a Ag/AgCl reference electrode.

Spectral Sensitizing Efficiency

The efficiency of spectral sensitization can be compared in a relative sense to the efficiency of latent-image formation for intrinsic exposures, leading to what is known as the "relative quantum efficiency" (RQE) (237). The RQE Φ_r is the ratio of the number of photons absorbed by the silver halide to produce a specified reference density on the

$D-\log E$ curve to the number of photons absorbed by the dye to produce the same density. An equation symbolizing this definition is

$$\Phi_r = \frac{400 \cdot E_{400} \cdot A_{400}}{\lambda \cdot E_\lambda \cdot A_\lambda} \tag{79}$$

where E_{400} is the incident irradiance at 400 nm and A_{400} is the fraction of incident radiation at 400 nm absorbed by the grain. Analogous definitions apply at wavelength λ absorbed by the dye. Of course, absolute efficiencies can also be measured using the quantum sensitivity procedures discussed earlier, in which case the ratio of the QS at 400 nm to that at wavelength λ would give Φ_r.

Using this more quantitative measurement of photographic efficiency, we can now discuss the correlation of electrochemical reduction potential with photographic behavior. This comparison is shown in Fig. 62 (254). The RQE is very closely coupled to the solution E_{red}. The solid curve in this figure is a fit from Marcus electron-transfer theory (255). Note that the efficiency varies by a factor of 100 for a very narrow range of E_{red}, indicating how critical the positioning of the dye LV level is for efficient electron transfer. We see from these data that "good" spectral sensitizers should have an E_{red} more negative than about -1.3 V.

It is also of interest from the mechanistic point of view to consider the dye's ability to transfer a hole to a silver halide grain. This can be done by studying how dyes bleach internal fog (256). With this technique, it is necessary to include a strong electron-trapping agent to prevent the electron from contributing to latent-image formation, so that only bleaching can occur. Then, the results from these experiments can be compared with the same dye's ability to form a surface negative image by the usual electron transfer process.

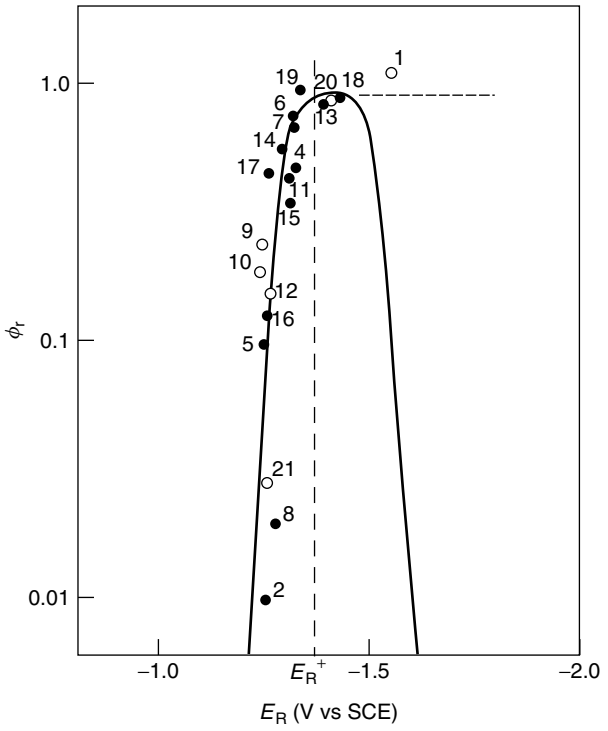


Figure 62. The effect of E_{red} on RQE for cyanine dyes. Numbers refer to the dye structure given in Ref. 254. (Reprinted with permission of the American Chemical Society.)

These hole bleaching experiments showed that there is a threshold of about $E_{ox} = +1.0$ V below which the hole cannot be transferred into the AgX valence band. The surface negative experiments showed that an electron cannot be transferred to the AgX conduction band when E_{red} is more positive than about -1.0 V. Also of note was that some dyes could also transfer both an electron and a hole to a grain. Of course, because the excitation energy of the dye is smaller than the band gap of silver halide, it is energetically impossible to transfer both carriers simultaneously. These dyes must first transfer an electron or hole from the excited state, depending on their HF and LV levels, and then transfer the complementary carrier later.

Correlations such as these have allowed generating energy-level diagrams for spectral sensitization that are qualitatively useful for understanding the mechanisms of spectral sensitizing dyes. One drawback of such electrochemical/photographic studies is that the dyes often aggregate. This shifts their energy relative to the adsorbed monomer, the state that would be expected to correlate best with the solution redox potentials. A possible way around this has been demonstrated by measuring redox potentials of an adsorbed dye in an emulsion coating by a combination of reflectance spectroscopy and treating a dyed-emulsion coating with a redox couple solution (257). The work showed that for the particular infrared dye studied, the major energetic effect in going from the monomer to the J aggregate of the dye was a lowering of the dye LV level. The generality of such an *in situ* measurement has yet to be demonstrated.

A general model of spectral sensitization has been developed that is based on the concepts expressed in Marcus electron-transfer theory (5,258). The model recognizes that there are two classes of spectral sensitizing dyes based on their mode of charge transfer. Some dyes function by first transferring an electron from the excited dye and then transferring a hole later. Others transfer a hole from the excited dye and then transfer an electron later. As would be expected, the class in which a particular dye falls depends on its energy levels relative to the conduction and valence bands of the silver halide. Using this model, very good correlation is found between the RQE of the dye and its solution redox potentials, as illustrated in Fig. 62 (254).

This model showed that a dye whose LV level is coincident with the minimum of the silver bromide conduction band had an $E_{red} = -0.92$ V. This value differs from the -1.3 V in Fig. 62 that corresponds to a spectral sensitizing dye with RQE approximately unity because one must also consider the reorganization energy in Marcus theory. This energy was found to be 0.42 eV (5,258), so that a dye with $E_{red} = -1.34 (= -0.92 - 0.42)$ V would have approximately unit efficiency of injecting an electron into the conduction band.

Spectral sensitization by hole-injecting dyes can be efficient if the dye ground-state energy level is at or

below the valence-band energy level, leading to a dye anion radical. The odd electron in this species lies higher in energy than it would if it were in the LV level of the excited dye (237). This is qualitatively explained by postulating electron–electron repulsion effects. Calculated energy differences between these two states are on the order of 0.2 eV (5), which makes transferring the odd electron more energetically favorable. In addition, there are no competing radiative or nonradiative de-excitation processes as in the case of an excited dye. A more circuitous route for transferring the odd electron to the grain involves transfer to a surface state, which forms a silver atom. Then, decay of the silver atom allows the electron to enter the conduction band.

Dye Desensitization

Dyes can decrease or even eliminate the intrinsic or blue sensitivity of the silver halide, depending on their energy levels and concentration (5,237). Although such effects are most easily quantified by comparing the intrinsic response with and without adsorbed dye, whatever desensitizing properties are at work also affect the response of the emulsion when exposed in the dye absorption region. As a result, there is a limit to the amount of dye that can be adsorbed on the grain, and if this limit is exceeded, the speed actually decreases. Thus, dye desensitization processes limit the maximum spectral speed that can be obtained from spectral sensitization.

Dye desensitization has it origins in several possible processes. Dyes whose LV level lies below the conduction band ("electron-trapping" dyes) can trap an electron from the conduction band. The trapping itself is often reversible, but secondary processes lead to irreversible loss of the electron. One secondary process involves recombination with a hole in the valence band, so that the dye acts as an additional recombination center. Another secondary process is loss of the trapped electron from the dye to O_2 in the coating which subsequently reacts with a hole and provides yet another recombination pathway (259). Evidence for such a process comes from exposures of films in vacuum that show markedly reduced dye desensitization (260).

As mentioned earlier, dyes that would normally be chosen for practical applications are those whose E_{red} is more negative than -1.3 V, or as close to this criterion as practically possible. As a result, electron-trapping desensitization is very unlikely for such dyes, but there is an alternative desensitization pathway for them. Because the dye LV level is positioned near or above the conduction band edge to maximize the yield of transferred electrons, by definition such dyes will introduce a new hole-trapping level associated with their HF level; these dyes can trap holes from the valence band, as shown in Fig. 63.

Such dyes can desensitize by two pathways. The first is complementary to that for electron-trapping dyes. The oxidized dye can act as a recombination center for electrons in the conduction band (237). The second pathway is based on the ability of the dye-trapped hole to migrate across the surface of a grain by hopping from one dye molecule to another (261). When the dye-trapped hole encounters a latent image or latent subimage, there is an electron transfer to the oxidized dye. This could be considered an alternative pathway for recombination, but it is more traditionally called regression.

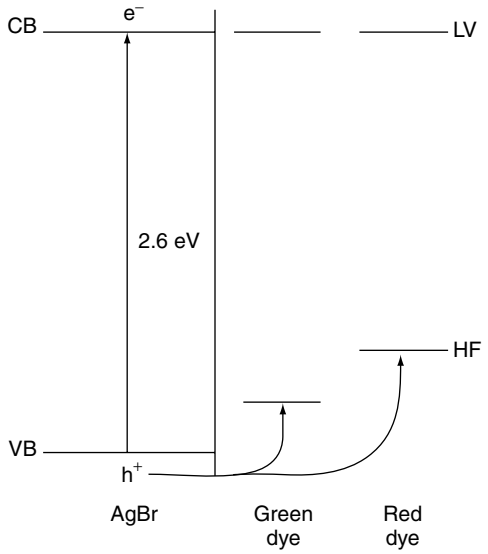


Figure 63. Ideal energy-level diagram for spectral sensitization by green- and red-absorbing cyanine dyes for maximum electron transfer efficiency. The dye HF levels provide additional hole-trapping sites for intrinsically generated holes that lead to desensitization.

Mechanistic studies of dye desensitization by good spectral sensitizing dyes have shown such dyes produce much lower desensitization when the location of the latent image is separated from that of the adsorbed dye (261–263). One way to accomplish this is to use emulsions that produce an exclusively internal image, for example, by sulfur-plus-gold sensitizing an emulsion and then precipitating more silver halide to produce a core–shell structure. Although not practically useful because the usual development is limited to a surface image, special developers can be formulated to reveal the internal image and allow such systems to provide mechanistic insight into dye desensitization processes. In fact, quantitative studies have shown that dye desensitization is completely eliminated in such systems even up to monolayer coverage of dyes that are very good hole traps ($E_{ox} < 0.60$ V) (261). Such results are strong evidence for the regression mechanism of dye desensitization.

Supersensitization

It should be clear by now that one of the important factors in achieving efficient spectral sensitization is that the dye LV level should be at or above the CB minimum, so that the quantum yield of CB electrons approaches one. However, such requirements become problematical, particularly for red and infrared dyes, because of dye desensitization mechanisms. Even though the dye leads to a high yield of CB electrons, most of them are eventually lost through desensitization processes. Another problem with this energy level criterion is that it forces the HF level to lie very high within the silver halide band gap. Such

dyes can be unstable because now they can be oxidized by other species in the coating, for example, O_2 (264).

From the standpoint of efficient electron transfer, such considerations suggest that it is necessary to use less than ideal dyes, for spectral sensitization. The electron transfer efficiencies of such dyes can be improved by a process known as supersensitization (265–267). This process uses an additional molecule, which may be another dye, whose HF level lies higher in energy than the spectral sensitizing dye HF level. The supersensitizer can transfer an electron to the half-filled HF level of the excited dye, which creates a reduced dye molecule. The advantage of this process is that now the electron residing in the former dye LV level is situated some 0.2 eV higher than an electron in the same level of the excited dye, as discussed before. The electron is more energetically disposed to transfer to the CB than it would be if just the usual excitation process was at work.

IMAGE PROCESSING

In image processing, the latent image is used as a catalyst to initiate development of the grain, that is, the silver ions in the crystal are reduced to silver atoms (4,5,268–270). In "black-and-white" development, the developed image consists of silver, whose extent determines the "blackness" of the image. Black-and-white development is an all or nothing process. Either the grain has received sufficient exposure to generate a latent image and all the silver ions of that grain are converted to silver atoms, or the grain does not develop at all. Here is where the silver halide detector derives its tremendous amplification capability. Considering that the minimum developable size of the latent image is three atoms and that a 1 μm cubic grain contains 2×10^{10} ion pairs, an amplification of the order of $10^{10}$is achieved!

In color photography, a dye image, rather than a silver one, is produced (271–273). To accomplish this, a by-product of the development reaction, the oxidized developer, is used by allowing it to react with an image dye precursor to form the image dye. The dye forms a "cloud" around the grain. In this type of development, a bleaching step must be part of the postdevelopment process, so that the developed silver can be removed from the final image. Otherwise, desaturation of the colors results.

For the most part, grains act independently of each other during exposure of a film layer. That is to say, absorption and subsequent latent-image formation in one grain does not affect the electronic and ionic processes in surrounding grains. During development, however, there can be considerable intergrain interactions, particularly in color development (271). These interactions can be fine-tuned to enhance the final image in sharpness and/or noise (see Detector Design Considerations section). In color photography, development is also restricted to only a small fraction of the grain so as to control the size of the dye clouds and reduce graininess in the final image. A comparison of grains developed in a black-and-white developer and a color developer is given in Fig. 64.

Microscopic examinations of developed grains have shown that they are usually masses of filamentary silver (269,274) often 15–25 nm in diameter and several

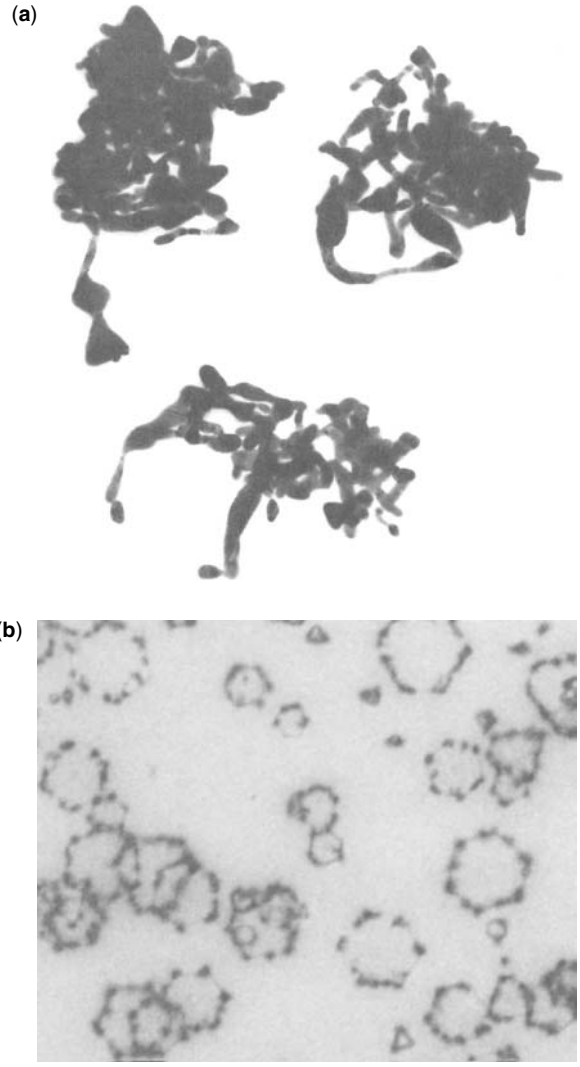


Figure 64. Micrographs of developed grains that show the difference between black-and-white development (**a**) and color development (**b**).

times that in length. The filament size has little dependence on the initial grain size. Developed silver is initially spherical, but, the developing site elongates into a filament at some point during development. There is no general agreement on the reason for this switchover, but some researchers have suggested that the reaction becomes limited by either diffusion of silver atoms (275) or silver ions (276,277), so the spherical geometry can no longer be maintained.

Electrochemistry of Development

Silver ions are reduced to silver atoms by reducing agents that are part of the developer formulation. As such, the process is best described by an electrochemical approach. The development reaction can be described by the following reaction:

$$D_{\text{red}} + n\text{Ag}^+ \longrightarrow n\text{Ag} + D_{\text{ox}} + mH^+ \qquad (80)$$

where D_{red} is the reduced form of the developer, D_{ox} is the oxidized form of the developer, n is the number of electrons transferred (1 or 2) and m is the number of protons produced (1 or 2). This reaction is composed of two half-cell reactions — one involves the silver ion, and the other involves the developing agent. These can be written as

$$Ag^+ + e^- \rightleftharpoons Ag, \quad E_{Ag} \tag{81}$$

$$D_{ox} + mH^+ \rightleftharpoons D_{red}, \quad E_{dev} \tag{82}$$

where E_{Ag} and E_{dev} are the electrochemical potentials of the half-cell reactions. The overall potential E_{cell} for the development reaction is given by

$$E_{cell} = E_{Ag} - E_{dev} \tag{83}$$

A thermodynamic relationship between E_{cell} and the free energy change ΔG is given by $\Delta G = -nFE_{cell}$, where F is Faraday's constant. Thus, when E_{cell} is positive, the free energy change is negative and reaction (80) occurs as written, that is, reduction occurs. But when E_{cell} is negative, the free energy change is positive, and reaction (80) occurs in reverse, that is, oxidation occurs. When E_{cell} is not 0, the reaction proceeds until $E_{cell} = 0$. Of course, if the by-products D_{ox} and H^+ can be removed, then E_{cell} never reaches 0 until one of the reactants is consumed.

The silver potential E_{Ag} is, of course, given by the Nernst equation which for a one-electron reaction at room temperature is expressed as

$$E_{Ag} = E^o_{Ag} - 0.059 \log \frac{[Ag]}{[Ag^+]} \tag{84}$$

and because [Ag] is assigned a value of unity and the solubility product K_{sp} controls $[Ag^+]$, we can recast the Nernst equation into

$$E_{Ag} = E^o_{Ag} + 0.059 \log K_{sp} - 0.059 \log[Br^-] \tag{85}$$

This equation indicates that the silver potential for developing AgBr is controlled by the Br^- concentration. Thus, most developer formulations contain Br^- to adjust the silver potential for optimum development.

The developing agents are usually organic compounds (268) that generally lose both an e^- and H^+ during development and can often lose a total of two of each, depending on the structure. It is useful to consider just the ionic equilibria first described by the following reversible reactions:

$$H_2R \longrightarrow HR^- + H^+ \longrightarrow R^{2-} + H^+ \tag{86}$$

$$R^{2-} + H^+ \longrightarrow HR^- + H^+ \longrightarrow H_2R \tag{87}$$

All three forms of the developing agent (H_2R, HR^-, and R^{2-}) can be reductants, although with different electron transfer abilities. For the two deprotonation reactions, the equilibrium constants are given by

$$K_1 = \frac{[H^+][HR^-]}{[H_2R]}; K_2 = \frac{[H^+][R^{2-}]}{[HR^-]} \tag{88}$$

Now, if we also consider the electron transfer reaction, the Nernst equation would look like

$$E_{dev} = E^o_{dev} + 0.0295 \log \frac{[H^+]^2[D_{ox}]}{[D_{red}]} \tag{89}$$

where $[D_{ox}]$ is the D_{ox} concentration and $[D_{red}]$ is the D_{red} concentration, which for this example would be given by

$$[D_{red}] = [HR] + [HR^-] + [R^{2-}] \tag{90}$$

Now substituting expressions from the ionic equilibria, (reactions (86) and (87)) we get for E_{dev}

$$E_{dev} = E^o_{dev} + 0.0259 \log \frac{[D_{ox}]}{[D_{red}]} + 0.0259 \log([H^+]^2 + K_1[H^+] + K_1K_2) \tag{91}$$

This equation indicates that E_{dev} will be affected by $[D_{red}]$, $[D_{ox}]$, and the pH.

Developing Agents

Most developing agents used today are based on appropriately substituted benzene derivatives and can be classified into the following three categories: (1) di- and polyhydroxy compounds (Fig. 65a), (2) amino hydroxy compounds (Fig. 65b), and (3) di- and polyamino compounds (Fig. 65c). The first two categories are used in developers for black-and-white images, whereas the third category is used in color developers. Some developers are based on ring structures that contain noncarbon nuclei such as in Fig. 65d.

A primary example of the polyhydroxy category is hydroquinone, a component in the majority of black-and-white developers. It exists in three forms, depending on the pH, as shown in Fig. 66. The pK values are useful because they indicate the pH at which the particular form will be half protonated. Many black-and-white developers work at around pH 10, so, the two forms of hydroquinone on the left and in the center of Fig. 66 will predominate. However, the ability to transfer electrons increases as the degree of deprotonation increases. The electron transfer reactions, illustrated in Fig. 67 for the fully deprotonated form, show that two electrons can be transferred to form the quinone molecule, D_{ox}. Similar reactions are at work for the other developing agent categories.

Developer Solutions

The driving force for development is the potential difference between the silver and developer half-cells, E_{cell}. If E_{cell} changes during development, then the rate of development will change and perhaps become so small that development ceases. E_{cell} changes because $[H^+]$ and $[D_{ox}]$ increase during development, whereas $[D_{red}]$ decreases, causing E_{dev} to increase. During development,

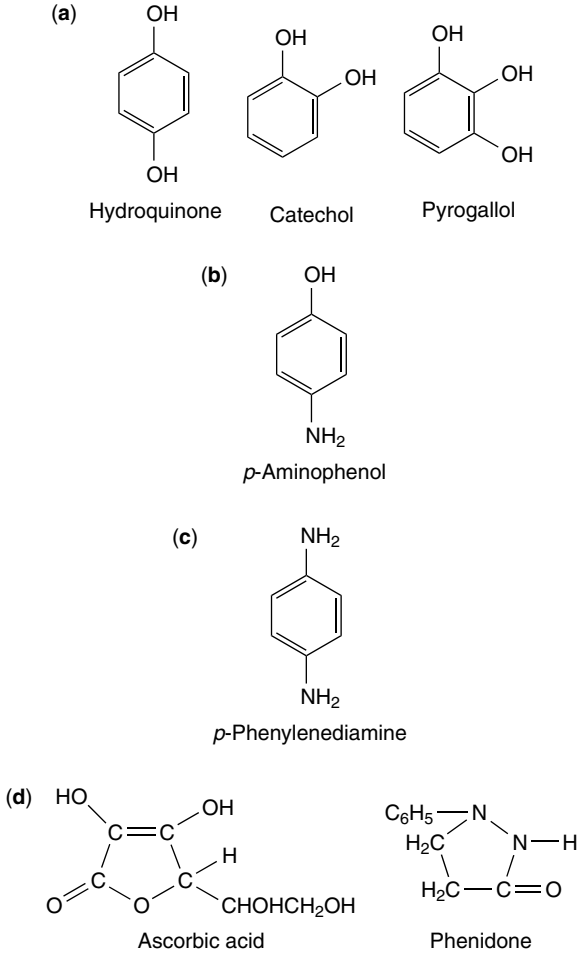


Figure 65. Structures of developing agents.

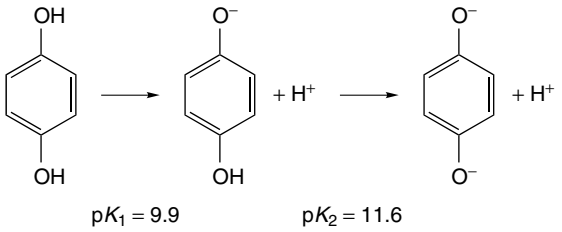


Figure 66. The deprotonation reactions of hydroquinone. The two pK_a values are indicated.

as silver ions are reduced to silver atoms, Br^- is also released into solution to maintain electrical neutrality within the developing grain. As a result, E_{Ag} decreases. These features of development are shown schematically in Fig. 68. The driving force obviously decreases over time. The developer contains at least two other components besides the developing agent to minimize some of these effects.

The developer contains a buffer that will react with the protons, released from the developing agent as a result of electron transfer, to prevent the local pH from decreasing. An example is Na_2CO_3, which buffers best at pH 10–11. This reagent reacts with water to give HCO_3^-, which

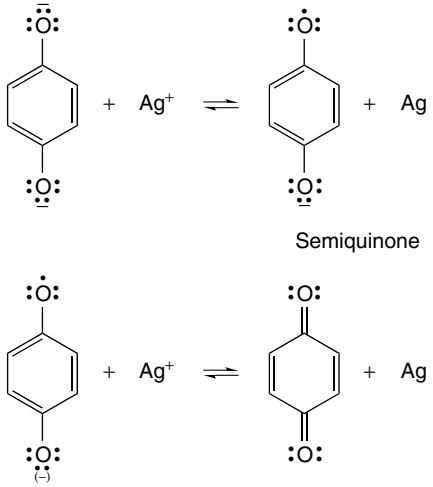


Figure 67. Electron-transfer reactions of fully deprotonated hydroquinone. The semiquinone is an intermediate stage of oxidation. The quinone form is depicted at the lower right.

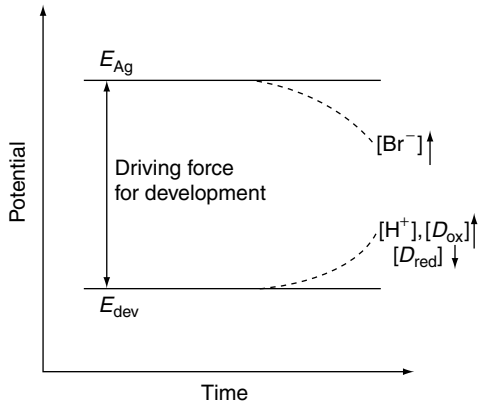


Figure 68. Schematic showing the effect of Br^- and H^+ release and D_{ox} formation on the silver and developer half-cell potentials (dashed curves).

undergoes a further reaction to give OH^- and H_2CO_3. As OH^- is consumed by the H^+ released from the oxidized developer, more HCO_3^- dissociates to maintain $[OH^-]$ constant. Other buffers are metaborate/tetraborate for pH 8–10 and phosphate/monohydrogen phosphate for pH 11–12.

Most developers contain sulfite that reacts with oxidized developing agents to produce inert forms, thereby preventing the buildup of D_{ox} and an increase in E_{dev}. Because dissolved O_2 can also react with the developing agent, which can catalyze further formation of D_{ox}, sulfite also acts as a preservative to allow developer solutions to remain stable for prolonged storage periods.

All developers also contain halide ion to adjust the silver potential, usually Br^- but sometimes Cl^- if AgCl emulsions are processed. The halide ion retards both development of image and fog, but more so the latter. Thus, it acts as an antifoggant. The antifoggant effect is primarily kinetic and allows substantially complete development of the image before the onset of fog, provided

that the proper development time is chosen. However, there is nothing in the developer to counteract the release of bromide during development. As a result, the silver potential decreases over time, but not to the extent that stops development. However, subsequent film developed in the same developer will probably not develop at the same speed because the halide concentration is higher than that for the first set of films.

In commercial developing and printing, often called "photofinishing," film or paper is continually being run through processing machines. Developing agent, sulfite, and buffer are continually consumed, and halide is continually generated. Under such conditions, it is impossible to maintain a constant E_{cell}, and the development rate slows down. To avoid this problem, a procedure called "replenishment" is employed in which a "replenisher" solution that contains fresh developing agent, sulfite, and buffer is continually added to the developer solution at a rate sufficient to compensate for the loss of developer components during development. To maintain a constant volume of developer, an equivalent amount of used developer flows out. The replenisher contains no halide, so the halide in the working developer solution is diluted. All of this allows E_{cell} to remain constant and provides constant developer activity during many processing cycles.

Types of Development

Development can be classified by the way the silver ions that take part in the development reaction are provided to the reacting site. In "chemical" development, the silver ions in the crystal lattice diffuse through the lattice to the developing site and take part in the reduction process, as shown in Fig. 69a. In "physical" development, the silver ions are made available by diffusion through the solution phase, and within this class are further distinguished as (1) solution physical development when the silver ions are provided by the crystal but diffuse through solution to arrive at the developing site; (2) prefixation physical development when the silver ions are provided in the developer formulation and the silver ions in the crystal are not used (Fig. 69b); and (3) postfixation physical development when the silver halide grains are first removed from the film layer, but not the latent image, and then the remaining film is subjected to a developer that contains the requisite silver ions.

Solution physical development occurs because components of the developer solution such as sulfite can solubilize lattice silver ions and provide transport through the solution to developing sites. Because Br^- is released during development and because it can increase the solubility of AgBr (see Fig. 10), there is always some solution physical development, even in the absence of any developer-induced effects.

Initiation of Development

After inserting of the film into the developer solution, there is a time during which nothing seems to be happening. The extent of this period will, of course, be a function of the detection method. This apparent period of inactivity is

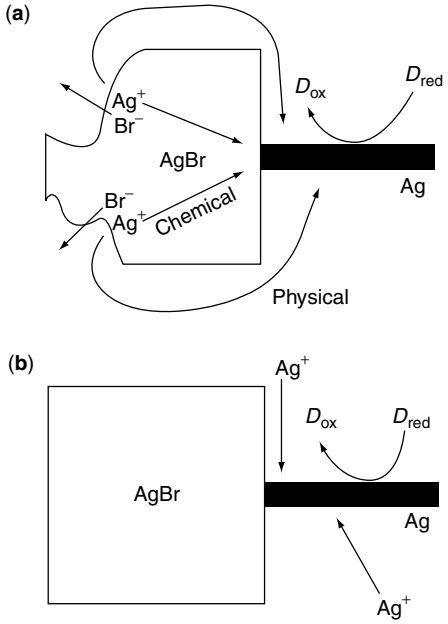


Figure 69. Illustration of the Ag^+ transport stage in development. In (**a**) the silver is transported either through the grain (chemical development) or by complexation of the silver ion and subsequent diffusion through the solution phase (solution physical development).Bromide must leave the grain to maintain electrical balance and leads to dissolution of the grain. (**b**) represents another form of physical development in which the silver ions are supplied via the developer. In this case, the silver ions of the grain are not used, and the grain does not dissolve.

usually due to the small size of the latent image, although partial burying of the latent image in the grain can be an additional factor. This latter effect might happen if some recrystallization occurs during the process of chemical sensitization and leaves the sulfur-plus-gold sensitizer site slightly buried beneath the surface. In the early stages of development, the silver nuclei are small, and the equation for E_{Ag} derived earlier, which was based on macroscopic quantities of silver, does not apply.

The Gibbs–Thomson concept can be used to arrive at an E_{Ag} that would be useful in understanding initiation by using the following equation (269):

$$E_{\text{Ag}} = E_\infty - \frac{2\sigma_s V_m}{rF} \tag{92}$$

where E_∞ is the potential of bulk silver, r is the silver nuclei radius, σ_s is the free surface energy, V_m is the molar volume, and F is Faraday's constant. Here we see that as r becomes larger, the second term tends toward zero and the calculated E_{Ag} is just that from Eq. (85). But, for small nuclei, the second term becomes significant, and the calculated E_{Ag} and therefore, the driving force for development is smaller than that expected. Thus, the initial rate of development will be smaller than that observed at later times.

Rate of Development

The rate of development can be measured by plotting density versus development time for a fixed exposure.

Alternatively, the mass of developed silver can be determined and then plotted versus development time. The degree of swelling of the film layer, which is mainly determined by the extent of hardening of the layer, is one factor that affects the rate of development. Other factors are the rate of diffusion of the developer components into the layer; the adsorption of the reduced form of the developing agent on the developing silver; the electron-transfer rate from the developing agent to the developing silver; and the desorption, diffusion, or reaction of the oxidized form of the developing agent. Additionally, hydration and diffusion of the halide ion can be important as is the dissolution rate of the silver halide grain and the diffusion rate of soluble silver-ion complexes.

Revisiting Eq. (85), we see that the E_{Ag} is controlled by the K_{sp} for a fixed molarity of the halide ion. So, as one moves from AgCl to AgBr and then to AgI, the E_{Ag} decreases, which, in turn, decreases the driving force for development, provided that no changes are made to the developer half-cell reaction. Thus, it is well known that AgCl emulsions develop more rapidly than AgBr emulsions. AgI emulsions develop too slowly to be photographically useful.

Superadditivity

Two developing agents are actually present in all black-and-white developers. This is done because it has been found that several developing agent combinations produce synergistic effects (278). The development rate is greater in a certain concentration range than a simple additive effect would predict. The mechanism of this superadditivity effect appears to be a regeneration of one of the developing agents by the other. Thus, in a superadditivity combination A/B, as developing agent A is converted into A_{ox} via development, it is converted back into A by an electrochemical reaction with B in the solution phase.

Postdevelopment Processes

After development, there are developed and undeveloped grains in the coating. Without additional processing, the image will be unstable because the undeveloped grains are still photosensitive and will gradually blacken through the photolytic formation of silver. These undeveloped grains must be removed from the film layer to make the image stable; this process is called "fixing" (279). Silver halide is removed from the film by forming soluble silver complexes. The complexation must be strong enough to prevent reprecipitation during the final washing phase. The steps involved in fixing are (1) diffusion of the fixing agent into the film, (2) dissolution of silver halide to form soluble silver-ion complexes, and (3) diffusion of soluble silver-ion complexes out of the film into the bulk solution.

The most common fixing agent is the thiosulfate anion. The 1:1 complex is insoluble but the 1:2 and 1:3 complexes are water soluble. The sequence of reactions is as follows

$$Ag^+ \text{ (surface)} + S_2O_3^{2-} \rightleftharpoons AgS_2O_3^- \text{ (surface)} \quad (93)$$

$$AgS_2O_3^- \text{ (surface)} + S_2O_3^{2-} \rightleftharpoons Ag(S_2O_3)_2^{3-} \text{ (surface)} \quad (94)$$

$$Ag(S_2O_3)_2^{3-} \text{ (surface)} \rightleftharpoons Ag(S_2O_3)_2^{3-} \text{ (solution)} \quad (95)$$

Surprisingly, the specific cation used with the thiosulfate anion affects the fixing rate because the diffusion rate of the cations is in the order: $NH_4^+ > K^+ > Na^+ > Li^+$. To avoid charge buildup, the cation must diffuse into the film at the same rate as the anion. It turns out that NH_4^+ can accelerate the fixing process when added to a sodium thiosulfate solution, although there is a optimum ratio, beyond which the acceleration effect can be lost.

The final stage of image processing is the wash step. Washing removes salts and other species accumulated during other stages in image processing. A major component to be removed is thiosulfate. If it remains in the dried film or print, it will attack the silver image over time. The principle of washing is to provide a driving force for thiosulfate to diffuse out of the film or paper by providing a solution of very low thiosulfate concentration. This principle can be met by continually providing changes of water by using a flowing water setup.

The degree of washing can be predicted by using the Bunsen–Ostwald dilution law (279),

$$X_n = \left(\frac{v}{\overline{V} + v}\right) X_o \quad (96)$$

where X_o is the initial concentration of thiosulfate, X_n is the thiosulfate concentration after n washings, v is the volume of washing fluid in and on the materials between stages, and $\overline{V}$ is the volume used in each wash bath. Commercial acceptability requires $X_n/X_o = 100$, but archival acceptability requires $X_n/X_o = 1{,}000$.

DETECTOR DESIGN CONSIDERATIONS

What has been discussed so far focuses entirely on how the silver halide grain is optimized for the desired (usually maximum) sensitivity. But the performance of any detector is based on its achievable signal-to-noise ratio (SNR). Noise in the silver halide detector is observed as graininess in an image (4,280,281). The image of an object whose radiance is uniform may or may not display the expected uniformity, depending on the characteristics of the detector. When the image does not reproduce the object's uniform radiance faithfully, it has a granular appearance. There are fluctuations in the image density that should not be present.

The appearance of graininess has to do with the randomness by which the silver halide grains are placed within the detector during the coating operation. Following development, the "silver grains" are also randomly placed and lead to areas of image density fluctuations. The human visual system is very sensitive to this randomness. A similar degree of density fluctuation but in an ordered array would be less noticeable. Although the human visual system cannot resolve the individual silver grains, their random placement does translate into the appearance of graininess in the final image.

A further feature of this randomness in silver halide detectors is that it increases with grain size. Consider an

emulsion coating that contains small grains and another that contains large grains. If both coatings are made so that they have comparable maximum density, then the large grain coating will have higher graininess because the larger grains can achieve the required maximum density with fewer grains. As a result, the intergrain distance will be larger with correspondingly larger fluctuations in density.

The correlation of graininess with grain size leads to a fundamental problem in designing photographic systems whose SNR is optimum. As discussed before, higher speed is achieved most often by increasing the grain size. But, now we see that the noise also increases with the sensitivity increase, so that the improvement in SNR will be less than expected. The exact relationship between SNR and grain size is related to a host of emulsion, coating, and design factors, so no simple rule of thumb can be given. Nevertheless, it is a fundamental issue in optimizing photographic systems for image capture.

Yet another feature of the silver halide detector is how well it reproduces the sharpness of the object being imaged (4,280,281). If we imagine that the object reflects a very narrow beam of light, we are concerned whether this beam is blurred in the final image. Imaging scientists use the concept of a "point-spread function" to characterize the amount of blurring seen in the image. Grain size is an important factor in determining the point-spread function in a silver halide detector. Because the sizes of the grains are on the order of the wavelength of light used in image capture, there is considerable scattering of light. The point of light is spread out laterally so that there is blurring. Although a bit of oversimplification, the tendency is for the spread to be greater for larger grains.

Taking these three system design factors — sensitivity, graininess, and sharpness — we can relate them by reference to the "photographic space" shown in Fig. 70 (282). The volume under the triangle represents the space available to the system designer. If higher sensitivity is needed, then the grain size can be increased. This will move the apex of the triangle further up the vertical axis, but will also pull the points inward where the triangle intercepts the horizontal axes — the volume under the triangle remains constant, and one design attribute is optimized at the expense of others. Equal but opposite effects occur when grain size is decreased to reduce graininess and increase sharpness.

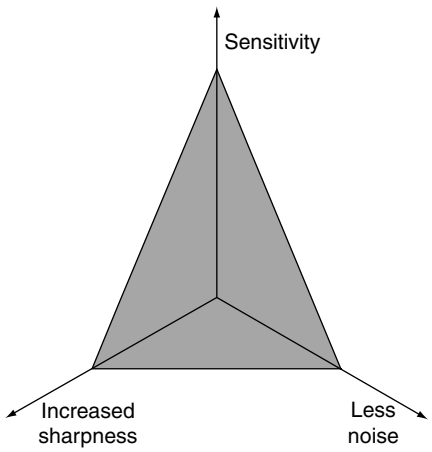


Figure 70. Schematic that illustrates detector design space. The area under the triangle is the "photographic space."

One way to overcome these design constraints is to improve the efficiency of latent-image formation. Using this approach, the sensitivity can be increased without increasing the grain size. Thus, the volume under the triangle increases, and the system designer has more degrees of freedom. If higher sensitivity is not needed for a particular application, then the efficiency-increasing technology can be applied to smaller grains to increase their sensitivity to the design requirements, and the smaller grain size ensures improved sharpness and less graininess. For this reason, photographic film manufacturers are constantly looking for ways to improve the efficiency of latent-image formation.

IMAGING SYSTEMS

This section briefly describes imaging systems based on silver halide detector technology; much more information can be found in other articles. The detector in systems designed to produce black-and-white images, would be prepared much as described earlier (4). Because the detector must have spectral sensitivity that spans the major part of the visible spectrum, both green and red dyes must be adsorbed on the grain surface. The emulsion would have a polydisperse grain-size distribution, so that the characteristic curve has a low-contrast (about 0.6), long-latitude response to minimize the risk of over- or underexposure. The final print image is made by exposing through the negative onto a negative-working silver halide emulsion on a reflective support. The characteristic curve of the emulsion intended for print viewing must have high contrast (about 1.8) to produce a positive image that has the correct variation of image gray value (tone scale) with scene luminance (4,14).

Systems designed to produce color prints also use a low-contrast emulsion to capture the image and a high-contrast emulsion to print the image. However, such systems are designed to produce dye images rather than silver images (271–273). As mentioned in the Image Processing section, this is done by allowing the oxidized developer to react with dye precursors to produce the dye image and then bleaching the developed silver back to silver halide for removal during fixing. Movie films are designed similarly to still camera films, except that the final-stage "print" is made using a transparent support, so that the image can be projected on a screen in a darkened room.

For color reproduction, the image must be separated into its blue, green, and red components using separate layers in the film that are sensitive to these spectral regions. This color separation is accomplished by using spectral sensitization to produce separate layers that are sensitive to blue through a combination of the silver halide

grain and an adsorbed blue spectral sensitizing dye (B), blue plus green by using a green spectral sensitizing dye (G), and blue plus red by using a red spectral sensitizing dye (R), as shown in Fig. 71. Representative spectral sensitivity curves for the different layers are shown in Fig. 72. By coating the B layer on the top and coating underneath it a yellow filter layer (which absorbs blue light), the G and R layers receive only minus-blue light. Then, by coating the G layer on top of the R layer, the latter receives only red light. Thus, the image has been separated into its B, G, and R components.

Systems that produce slides also use similar color technology, but now the image capture medium and the image display medium are the same film. Positive images are obtained by first developing the negative image in a black-and-white developer, then chemically treating the remaining undeveloped grains to make them developable, and finally developing those grains with a color developer. Then all the developed silver is bleached back to silver halide and fixed out. To produce the right tone scale in the final image, the film latitude must be considerably shorter than that in systems designed to produce prints and requires much more care in selecting optimum exposure conditions.

Systems designed for medical diagnosis using X-rays produce negative black-and-white images. Because film is a very poor absorber of X rays, these systems use a screen–film combination to minimize patient exposure (283). The screens are composed of heavy-element particles that are good X-ray absorbers and also emit radiation in the near UV or visible region. In these systems, the film need be sensitive only to the particular wavelength region where the screen emits. Films are designed with different contrasts to optimize the image for the particular diagnosis performed.

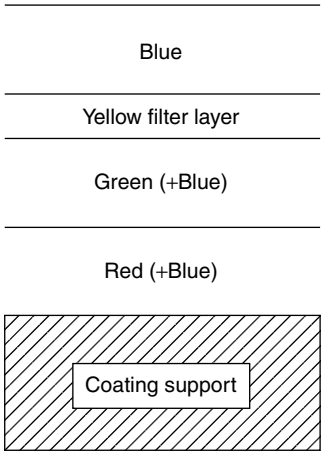


Figure 71. Arrangement of layers in a simple color film.

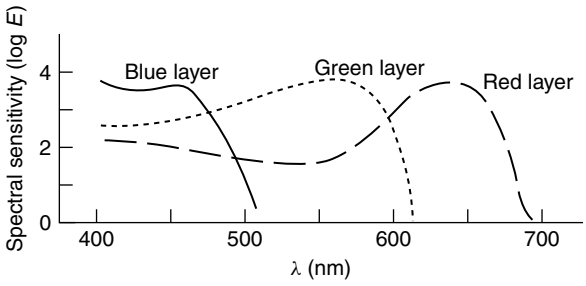


Figure 72. Spectral sensitivity of layers in color film. Data were obtained on separately coated layers.

ABBREVIATIONS AND ACRONYMS

E	exposure
t	exposure time
T	transmittance
R	reflectance
I	irradiance
D	image density
ISO	international standards organization
γ	contrast
AgCl	silver chloride
AgBr	silver bromide
AgI	silver iodide
Ag^+	silver ion
Br^-	bromide ion
I^-	iodide ion
K_{sp}	solubility product
pAg	negative logarithm of the silver ion concentration
pBr	negative logarithm of the bromide ion concentration
S	supersaturation
ΔG	free energy
Q	one-dimensional representation of a lattice distortion
η_n	nucleation efficiency
ω	recombination index
N	minimum number of silver/gold atoms in the developable latent image
QS	quantum sensitivity
F	fraction of grains developable
P	mean absorbed photons/grain
LIRF	low-irradiance reciprocity failure
HIRF	high-irradiance reciprocity failure
S	sulfur sensitized
S + Au	sulfur plus gold sensitized
LV	lowest vacant molecular orbital
HF	highest filled molecular orbital
RQE	relative quantum efficiency
E_{red}	electrochemical reduction potential
E_{ox}	electrochemical oxidation potential
D_{red}	reduced form of the developing agent
D_{ox}	oxidized form of the developing agent
E_{Ag}	electrochemical silver potential
E_{dev}	electrochemical developer potential
E_{cell}	difference between E_{Ag} and E_{dev} ($= E_{Ag} - E_{dev}$)
SNR	signal-to-noise ratio

BIBLIOGRAPHY

1. E. Ostroff, ed., *Pioneers of Photography. Their Achievements in Science and Technology*, SPSE, The Society for Imaging Science and Technology, Springfield, VA, 1987.

2. W. H. F. Talbot, *The Process of Calotype Photogenic Drawing* (communicated to the Royal Society, June 10, 1841) J. L. Cox & Sons, London, 1841.
3. G. F. Dutton, *Photographic Emulsion Chemistry*, Focal Press, London, 1966.
4. B. H. Carroll, G. C. Higgins, and T. H. James, *Introduction to Photographic Theory*, J Wiley, NY, 1980.
5. T. Tani, *Photographic Sensitivity*, Oxford University Press, NY, 1995.
6. T. H. James, T. H. James, ed., *The Theory of the Photographic Process*, 4th ed., Macmillan, NY, 1977.
7. J. F. Hamilton, *Adv. Phys.* **37**, 359 (1988).
8. D. M. Sturmer and A. P. Marchetti, in J. Sturge, V. Walworth, and A. Shepp, eds., *Imaging Materials and Processes*, Neblette's 8th ed., Van Nostrand Reinhold, NY, 1989, Chap. 3.
9. D. J. Locker, in *Kirk Othmer Encyclopedia of Chemical Technology*, vol. 18, 4th ed., J Wiley, NY, 1996, pp. 905–963.
10. R. S. Eachus, A. P. Marchetti, and A. A. Muenter, in H. L. Strauss, G. T. Babcock, and S. R. Leone, eds., *Annual Review of Physical Chemistry*, vol. 50, Annual Reviews, Palo Alto, CA, 1999, pp. 117–144.
11. H. A. Hoyen and X. Wen, in C. N. Proudfoot, ed., *Handbook of Photographic Science and Engineering*, 2nd ed., The Society for Imaging Science and Technology, Springfield, VA, 1997, pp. 201–224.
12. J. H. Altman, in T. H. James, ed., *The Theory of the Photographic Process*, 4th ed., Macmillan, NY, 1977, Chap. 17.
13. P. Kowalski, in T. H. James, ed., *The Theory of the Photographic Process*, 4th ed., Macmillan, NY, 1977, Chap. 18.
14. C. N. Nelson, in T. H. James, ed., *The Theory of the Photographic Process*, 4th ed., Macmillan, NY, 1977, Chap. 19, Sec. I.
15. American National Standard PH2.5-1972.
16. American National Standard PH2.27-1965.
17. C. Kittel, *Introduction to Solid Sate Physics*, 4th ed., J Wiley, NY, 1971.
18. E. Vogl and W. Waidelich, *Z. Agnew. Phys.* **25**, 98 (1968).
19. M. Bucher, *Phys. Rev. B* **35**, 2,923 (1987).
20. W. Nernst, *Z. Physik. Chem.* **4**, 129 (1889). Consult any analytical chemistry textbook.
21. I. H. Leubner, R. Jaganathan, and J. S. Wey, *Photogr. Sci. Eng.* **24**, 268 (1980).
22. E. J. Birr, *Stabilization of Photographic Silver Halide Emulsions*, Focal Press, London, 1974.
23. J. W. Mullin, *Crystallization*, 2nd ed., Butterworths, London, 1972, p. 222.
24. C. R. Berry and D. C. Skillman, *J. Photogr. Sci.* **16**, 137–147 (1968).
25. C. R. Berry and D. C. Skillman, *J. Phys. Chem.* **68**, 1,138 (1964).
26. C. R. Berry and D. C. Skillman, *J. Appl. Phys.* **33**, 1,900 (1962).
27. P. Claes and H. Borginon, *J. Photogr. Sci.* **21**, 155 (1973).
28. J. S. Wey and R. W. Strong, *Photogr. Sci. Eng.* **21**, 14–18 (1977).
29. T. Tanaka and M. Iwasaki, *J. Imaging Sci.* **29**, 86 (1985).
30. C. R. Berry and D. C. Skillman, *Photogr. Sci. Eng.* **6**, 159–165 (1962).
31. J. S. Wey and R. W. Strong, *Photogr. Sci. Eng.* **21**, 248 (1977).
32. I. H. Leubner, *J. Phys. Chem.* **91**, 6,069 (1987) and references cited therein.
33. C. R. Berry, S. J. Marino, and C. F. Oster Jr., *Photogr. Sci. Eng.* **5**, 332–336 (1961).
34. R. Jagannathan, *J. Imaging Sci.* **35**, 104–112 (1991).
35. M. Antonaides and J. S. Wey, *J. Imaging Sci. Technol.* **39**, 323–331 (1995).
36. C. R. Berry, in T. H. James, ed., *The Theory of the Photographic Process*, 4th ed., Macmillan, NY, 1977, Chap. 3.
37. A. H. Herz and J. Helling, *J. Colloid Sci.* **17**, 293 (1962).
38. W. L. Gardiner, D. Wrathall, and A. H. Herz, *Photogr. Sci. Eng.* **21**, 325–330 (1977).
39. S. Boyer, J. Cappelaere, and J. Pouradier, *Chim. Phys.* **56**, 495 (1959).
40. J. E. Maskasky, *J. Imaging Sci.* **30**, 247 (1986).
41. I. H. Leubner, *J. Imaging Sci.* **31**, 145 (1987).
42. M. J. Harding, *J. Photogr. Sci.* **27**, 1–12 (1979).
43. W. Markocki and A. Zaleski, *Photogr. Sci. Eng.* **17**, 289–294 (1973).
44. A. B. Holland and J. R. Sawers, *Photogr. Sci. Eng.* **17**, 295–298 (1973).
45. A. B. Holland and A. D. Feinerman, *J. Appl. Photogr. Eng.* **84**, 165 (1982).
46. D. L. Black and J. A. Timmons, *J. Imaging Sci. Technol.* **38**, 10–13 (1994).
47. J. E. Maskasky, *J. Imaging Sci.* **31**, 15–26 (1987).
48. J. E. Maskasky, *J. Imaging Sci.* **32**, 15–16 (1988).
49. B. K. Furman, G. H. Morrison, V. I. Saunders, and Y. T. Tan, *Photogr. Sci. Eng.* **25**, 121 (1981).
50. T. J. Maternaghan, C. J. Falder, R. Levi-Setti, and J. M. Chabala, *J. Imaging Sci.* **34**, 58–65 (1990).
51. J. F. Hamilton and L. E. Brady, *Surface Sci.* **23**, 389 (1970).
52. R. C. Baetzold, Y. T. Tan, and P. W. Tasker, *Surface Sci.* **195**, 579 (1988).
53. P. Tangyunyong, T. N. Rhodin, Y. T. Tan, and K. J. Lushington, *Surface Sci.* **255**, 259 (1991).
54. Y. T. Tan, K. J. Lushington, P. Tangyunyong, and T. N. Rhodin, *J. Imaging Sci. Technol.* **36**, 118 (1992).
55. P. W. M. Jacobs, J. Corish, and C. R. A. Catlow, *J. Phys. C Solid State Phys.* **13**, 1977 (1980).
56. W. G. Kleppmann and H. Bilz, *Commun. Phys.* **1**, 105 (1976).
57. H. Bilz and W. Weber, in A. Baldereschi, W. Czaja, E. Tosati, and M. Tosi, eds., *The Physics of Latent Image Formation in Silver Halides*, World Scientific, Singapore, 1984, p. 25.
58. R. J. Friauf, in A. Baldereschi, W. Czaja, E. Tosati, and M. Tosi, eds., *The Physics of Latent Image Formation in Silver Halides*, World Scientific, Singapore, 1984, p. 79.
59. J. Van Biesen, *J. Appl. Phys.* **41**, 1,910 (1970).
60. S. Takada, *J. Appl. Phys. Jpn.* **12**, 190 (1973).
61. S. Takada, *Photogr. Sci. Eng.* **18**, 500 (1974).
62. M. E. van Hull and W. Maenhout-van der Vorst, *Phys. Stat. Sol.* **39**(a), 253 (1977).
63. M. E. van Hull and W. Maenhout-van der Vorst, *Phys. Stat. Sol.* **40**(a), K57 (1977).
64. M. E. van Hull and W. Maenhout-van der Vorst, *Int. Cong. Photogr. Sci.*, Rochester, NY, 1978, Paper I 8.
65. M. E. van Hull and W. Maenhout-van der Vorst, *Phys. Stat. Sol.* **52**(a), 277 (1979).

66. F. Callens and W. Maenhout-van der Vorst, *Phys. Stat. Sol.* **50**(a), K175 (1978).
67. F. Callens and W. Maenhout-van der Vorst, *Phys. Stat. Sol.* **71**(a), K61 (1982).
68. F. Callens, W. Maenhout-van der Vorst, and L. Kettellapper, *Phys. Stat. Sol.* **70**(a), 189 (1982).
69. H. Pauly and H. P. Schwan, *Zeitsch. f. Naturforschung* **14b**, 125–131 (1950).
70. F. Callens, D. Vandenbroucke, L Soens, M. Van den Eeden, and F. Cardon, *J. Photogr. Sci.* **41**, 72–73 (1993).
71. J. Heick and F. Granzer, *J. Imaging Sci. Technol.* **38**, 464–474 (1995).
72. R. C. Baetzold, *Phys. Rev. B* **52**, 11,424–11,431 (1995).
73. K. Lehovec, *J. Chem. Phys.* **21**, 1,123 (1953).
74. K. L. Kliewer, *J. Phys. Chem. Solids* **27**, 705, 719–(1966).
75. R. B. Poeppel and J. M. Blakely, *Surface Sci.* **15**, 507 (1969).
76. Y. T. Tan and H. A. Hoyen Jr., *Surface Sci.* **36**, 242 (1973).
77. H. A. Hoyen, in A. Baldereschi, W. Czaja, E. Tosati, and M. Tosi, eds., *The Physics of Latent Image Formation in Silver Halides*, World Scientific, Singapore, 1984, p. 151.
78. H. A. Hoyen Jr. and Y. T. Tan, *J. Colloid Interface Sci.* **79**, 525–534 (1981).
79. F. Bassini, R. S. Knox, and W. B. Fowler, *Phys. Rev. A* **137**, 1,217 (1965).
80. W. B. Fowler, *Phys. Stat. Sol.(b)* **52**, 591 (1972).
81. F. Moser and R. Ahrenkiel, in T. H. James, ed., *The Theory of the Photographic Process*, 4th ed., Macmillan, NY, 1977, Chap. 1, Sect. IV.
82. R. C. Brandt and F. C. Brown, *Phys, Rev.* **181**, 1,241 (1969).
83. A. M. Stoneham, *Theory of Defects in Solids*, Clarendon, Oxford, 1975.
84. P. Langevin, *Ann. Chem. Phys.* **28**, 433 (1903).
85. M. Lax, *Phys. Rev.* **119**, 1,502 (1960).
86. R. M. Gibb, G. J. Rees, B. W. Thomas, B. L. H. Wilson, B. Hamilton, D. R. Wight, and N. F. Mott, *Philos. Mag.* **36**, 1,021 (1977).
87. Y. Toyozawa, *Semicond. Insulators* **5**, 175 (1983) and references therein.
88. H. Kanzaki and S. Sakuragi, *J. Phys. Soc. Jpn* **27**, 109 (1969).
89. F. Moser, R. K. Ahrenkeil, and S. L. Lyu, *Phys. Rev.* **161**, 897 (1967).
90. V. Platikanova and J. Malinowsi, *Phys. Stat. Sol.* **47**, 683 (1978).
91. R. E. Maerker, *J. Opt. Soc. Am.* **44**, 625 (1954).
92. M. Tamura, H. Hada, S. Fujiwara, and S. Ikenoue, *Photogr. Sci. Eng.* **15**, 200 (1971).
93. M. Kawasaki and H. Hada, *J. Soc. Photogr. Sci. Technol. Jpn.* **44**, 185 (1981).
94. H. Hada and M. Kawasaki, *J. Appl. Phys.* **54**, 1,644 (1983).
95. M. Kawasaki and H. Hada, *J. Imaging Sci.* **29**, 132 (1985).
96. M. Kawasaki and H. Hada, *J. Imaging Sci.* **31**, 267 (1987).
97. G. W. Luckey, *Discuss. Faraday Soc.* **28**, 113 (1959).
98. S. E. Sheppard, A. P. H. Trivelli, R. P. Loveland, *J. Franklin Inst.* **200**, 15 (1925).
99. R. W. Gurney and N. F. Mott, *Proc. R. Soc. London* **A 164**, 151 (1938).
100. N. F. Mott and R. W. Gurney, *Electronic Processes in Ionic Crystals*, Clarendon, Oxford, 1940.
101. J. W. Mitchell, *Rep. Prog. Phys.* **20**, 433 (1957).
102. J. W. Mitchell, *J. Phys. Chem.* **66**, 2,359 (1962).
103. J. W. Mitchell, *Photogr. Sci. Eng.* **22**, 249 (1978).
104. J. W. Mitchell, *Photogr. Sci. Eng.* **25**, 170 (1981).
105. J. F. Hamilton, in T. H. James, ed., *The Theory of the Photographic Process*, 4th ed., Macmillan, NY, 1977, Chap. 4.
106. J. F. Hamilton, in A. Baldereschi, W. Czaja, E. Tosati, and M. Tosi, eds., *The Physics of Latent Image Formation in Silver Halides*, World Scientific, Singapore, 1984, p. 203.
107. J. Malinowski, *Photogr. Sci. Eng.* **14**, 112 (1970).
108. J. Malinowski, *J. Photogr. Sci.* **18**, 363 (1974).
109. J. Malinowski, *Photogr. Sci. Eng.* **23**, 99 (1979).
110. E. Moisar, *Photogr. Sci. Eng.* **26**, 124–132 (1982).
111. E. Moisar, *Photogr. Sci. Eng.* **25**, 45–56 (1981).
112. E. Moisar and F. Granzer, *Photogr. Sci. Eng.* **26**, 1–14 (1982).
113. M. R. V. Sahyun, *Photogr. Sci. Eng.* **27**, 171–177 (1983).
114. M. R. V. Sahyun, *Photogr. Sci. Eng.* **28**, 157–161 (1984).
115. W. F. Berg, *Trans. Faraday Soc.* **39**, 115 (1943).
116. F. Seitz, *Rev. Mod. Phys.* **23**, 328 (1951).
117. B. E. Bayer and J. F. Hamilton, *J. Opt. Soc. Am.* **55**, 439–452 (1965).
118. P. C. Burton and W. F. Berg, *Photogr. J.* **86B**, 2 (1946).
119. P. C. Burton, *Photogr. J.* **86B**, 62 (1946).
120. P. C. Burton and W. F. Berg, *Photogr. J.* **88B**, 84 (1948).
121. P. C. Burton, *Photogr. J.* **88B**, 13 (1948).
122. P. C. Burton, *Photogr. J.* **88B**, 123 (1948).
123. W. F. Berg, *Rep. Prog. Phys.* **11**, 248 (1948).
124. R. K. Hailstone and J. F. Hamilton, *J. Imaging Sci.* **29**, 125–131 (1985).
125. J. F. Hamilton and B. E. Bayer, *J. Opt. Soc. Am.* **55**, 528–533 (1965).
126. J. F. Hamilton and B. E. Bayer, *J. Opt. Soc. Am.* **56**, 1,088–1,094 (1966).
127. J. F. Hamilton, *Photogr. Sci. Eng.* **26**, 263–269 (1982).
128. J. F. Hamilton, *Radiat. Effects* **72**, 103–106 (1983).
129. J. F. Hamilton and P. C. Logel, *Photogr. Sci. Eng.* **18**, 507–512 (1974).
130. R. K. Hailstone, N. B. Liebert, M. Levy, and J. F. Hamilton, *J. Imaging Sci.* **31**, 185–193 (1987).
131. R. K. Hailstone, N. B. Liebert, M. Levy, and J. F. Hamilton, *J. Imaging Sci.* **31**, 255–262 (1987).
132. P. Fayet, F. Granzer, G. Hegenbart, E. Moisar, B. Pischel, and L. Wöste, *Phys. Rev. Lett.* **55**, 3,002 (1985).
133. T. Leisner, C. Rosche, S. Wolf, F. Granzer, and L. Wöste, *Surf. Rev. Lett.* **3**, 1,105–1,108 (1996).
134. R. K. Hailstone and J. F. Hamilton, *J. Imaging Sci.* **31**, 229–238 (1987).
135. F. Trautweiler, *Photogr. Sci. Eng.* **12**, 138–142 (1968).
136. D. E. Powers, S. G. Hamsen, M. E. Geusic, D. L. Michalopoulos, and R. E. Smalley, *J. Chem. Phys.* **78**, 2,866–2,881 (1983).
137. M. Kawaski, Y. Tsujimura, and H. Hada, *Phys. Rev. Lett.* **57**, 2,796–2,799 (1986).
138. J. F. Hamilton and R. C. Baetzold, *Photogr. Sci. Eng.* **25**, 189–197 (1981).
139. R. C. Baetzold, *J. Phys. Chem.* **101**, 8,180–8,190 (1997).
140. T. H. James, *J. Photogr. Sci.* **20**, 182–186 (1972).

141. R. K. Hailstone and J. F. Hamilton, *J. Photogr. Sci.* **34**, 2–8 (1986).
142. J. F. Hamilton, *Photogr. Sci. Eng.* **14**, 122–130 (1970).
143. J. F. Hamilton, *Photogr. Sci. Eng.* **14**, 102–111 (1970).
144. V. I. Saunders, R. W. Tyler, and W. West, *Photogr. Sci. Eng.* **16**, 87 (1972).
145. R. S. Van Heyingen and F. C. Brown, *Phys. Rev.* **111**, 462 (1958).
146. H. E. Spencer and R. E. Atwell, *J. Opt. Soc. Am.* **54**, 498–505 (1964).
147. R. Deri and J. Spoonhower, *J. Appl. Phys.* **57**, 2,806 (1985).
148. J. Flad, H. Stoll, and H. Preuss, *Z. Phys. D—At. Mol. Clusters* **6**, 193–198 (1987).
149. J. Flad, H. Stoll, and H. Preuss, *Z. Phys. D—At. Mol. Clusters* **6**, 287–292 (1987).
150. J. Flad, H. Stoll, and H. Preuss, *Z. Phys. D—At. Mol. Clusters* **15**, 79–86 (1990).
151. P. G. Nutting, *Philos. Mag.* **26**(6), 423 (1913).
152. J. Gasper and J. J. DePalma, in T. H. James, ed., *The Theory of the Photographic Process*, 4th ed., Macmillan, NY, 1977, Chap. 20.
153. J. H. Webb, *J. Opt. Soc. Am.* **38**, 27 (1948).
154. G. C. Farnell, *J. Photogr. Sci.* **7**, 83 (1959).
155. G. C. Farnell, *J. Photogr. Sci.* **8**, 194 (1960).
156. R. K. Hailstone, N. B. Liebert, M. Levy, R. T. McCleary, S. R. Girolmo, D. L. Jeanmaire, and C. R. Boda, *J. Imaging Sci.* **32**, 113–124 (1988).
157. L. Silberstein, *Philos. Mag* **45**, 1,062 (1923).
158. G. C. Farnell and J. B. Chanter, *J. Photogr. Sci.* **9**, 73 (1961).
159. H. E. Spencer, *Photogr. Sci. Eng.* **15**, 468 (1971).
160. G. C. Attridge, *J. Photogr. Sci.* **30**, 197 (1982).
161. T. A. Babcock and T. H. James, *J. Photogr. Sci.* **24**, 19 (1976).
162. G. C. Farnell, *J. Photogr. Sci.* **17**, 116 (1969).
163. P. Broadhead and G. C. Farnell, *J. Photogr. Sci.* **30**, 176 (1982).
164. T. Tani, *J. Imaging Sci.* **29**, 93 (1985).
165. A. G. DiFrancesco, M. Tyne, C. Pryor, and R. Hailstone, *J. Imaging Sci. Technol.* **40**, 576–581 (1996).
166. T. Tani, *J. Soc. Photogr. Sci. Technol. Jpn.* **43**, 335 (1980).
167. J. W. Mitchell, *Photogr. Sci. Eng.* **25**, 170 (1981).
168. R. K. Hailstone, *J. Appl. Phys.* **86**, 1,363–1,369 (1999).
169. R. W. Bunsen and H. E. Roscoe, *Ann. Phys. Chem.* **2**(117), 529 (1862).
170. J. H. Webb, *J. Opt. Soc. Am.* **40**, 3 (1950).
171. T. A. Babcock, P. M. Ferguson, W. C. Lewis, and T. H. James, *Photogr. Sci. Eng.* **19**, 49–55 (1975).
172. T. A. Babcock, P. M. Ferguson, W. C. Lewis, and T. H. James, *Photogr. Sci. Eng.* **19**, 211–214 (1975).
173. G. A. Janusonis, *Photogr. Sci. Eng.* **22**, 297–301 (1978)
174. J. M. Harbison and H. E. Spencer, in T. H. James, ed., *The Theory of the Photographic Process*, 4th ed., Macmillan, NY, 1977, Chap. 5.
175. H. E. Spencer, *J. Photogr. Sci.* **24**, 34–39 (1976).
176. S. Sheppard, *Photogr. J.* **65**, 380 (1925).
177. S. Sheppard, *Photogr. J.* **66**, 399 (1926).
178. D. J. Cash, *J. Photogr. Sci.* **20**, 19 (1972).
179. D. J. Cash, *J. Photogr. Sci.* **20**, 77 (1972).
180. D. J. Cash, *J. Photogr. Sci.* **20**, 223 (1972).
181. M. Ridgway and P. J. Hillson, *J. Photogr. Sci.* **23**, 153 (1975).
182. P. H. Roth and W. H. Simpson, *Photogr. Sci. Eng.* **24**, 133 (1975).
183. D. A. Pitt, D. L. Rachu, and M. R. V. Sahyun, *Photogr. Sci. Eng.* **25**, 57 (1981).
184. J. E. Keevert and V. V. Gokhale, *J. Imaging Sci.* **31**, 243 (1987).
185. H. Kamzaki and Y. Tadakuma, *J. Phys. Chem. Solids* **55**, 631 (1994).
186. H. Kanzaki and Y. Tadakuma, *J. Phys. Chem. Solids* **58**, 221 (1997).
187. R. C. Baetzold, *J. Imaging Sci. Technol.* **43**, 375 (1999).
188. H. Frieser and W. Bahnmuller, *J. Photogr. Sci.* **16**, 38 (1968).
189. J. F. Hamilton, J. M. Harbison, and D. L. Jeanmaire, *J. Imaging Sci.* **32**, 17 (1988).
190. D. Zhang and R. K. Hailstone, *J. Imaging Sci. Technol.* **37**, 61 (1993).
191. K. Morimura and H. Mifune, *J. Soc. Photogr. Sci. Technol. Jpn.* **61**, 175 (1998).
192. T. Kaneda, *J. Imaging Sci.* **33**, 115–118 (1989).
193. T. Tani, *Photogr. Sci. Eng.* **27**, 75 (1983).
194. T. Tani, *J. Imaging Sci. Technol.* **39**, 386 (1995).
195. H. E. Spencer and R. E. Atwell, *J. Opt. Soc. Am.* **54**, 498–505 (1964).
196. H. E. Spencer and M. Levy, *J. Soc. Photogr. Sci. Technol. Jpn.* **46**, 514–524 (1983).
197. E. Moisar, *Photogr. Sci. Eng.* **25**, 45 (1981).
198. R. K. Hailstone, N. B. Liebert, and M. Levy, *J. Imaging Sci.* **34**, 169–176 (1990).
199. J. Pouradier, A. Maillet, and B. Cerisy, *J. Chim. Phys.* **63**, 469 (1966).
200. K. Tanaka, *Nippon Kagaku Kaishi* **12**, 2,264 (1973).
201. H. Hirsch, *J. Photogr. Sci.* **20**, 187 (1972).
202. H. E. Spencer, *J. Imaging Sci.* **32**, 28–34 (1988).
203. D. Spracklen, *J. Photogr. Sci.* **9**, 145–(1961).
204. P. Faelens, *Photogr. Korr.* **104**, 137 (1968).
205. D. Cash, *Photogr. Sci. Eng.* **27**, 156 (1983).
206. R. C. Baetzold, *J. Photogr.Sci.* **28**, 15–22 (1980).
207. J. M. Harbison and J. F. Hamilton, *Photogr. Sci. Eng.* **19**, 322 (1975).
208. G. W. Lowe, J. E. Jones, and H. E. Roberts, in J. W. Mitchell, ed., *Fundamentals of Photographic Sensitivity (Proc. Bristol Symp.)*, Butterworths, London, 1951, p. 112.
209. T. Tani, *Photogr. Sci. Eng.* **15**, 181 (1971).
210. T. Tani and M. Murofushi, *J. Imaging Sci. Technol.* **38**, 1 (1994).
211. T. Tani, *J. Imaging Sci. Technol.* **41**, 577 (1997).
212. T. Tani, *J. Imaging Sci. Technol.* **42**, 402 (1998).
213. S. Guo and R. Hailstone, *J. Imaging Sci. Technol.* **40**, 210 (1996).
214. M. Kawasakki and Y. Oku, *J. Imaging Sci. Technol.* **42**, 409 (1998).
215. A. P. Marchetti, A. A. Muenter, R. C. Baetzold, and R. T. McCleary, *J. Phys. Chem. B* **102**, 5,287–5,297 (1998).
216. T. Tani, *Imaging Sci. J.* **47**, 1 (1999).
217. E. Moisar, *Photogr. Korr.* **106**, 149 (1970).
218. S. S. Collier, *Photogr. Sci. Eng.* **23**, 113 (1979).
219. H. E. Spencer, L. E. Brady, and J. F. Hamilton, *J. Opt. Soc. Am.* **57**, 1,020 (1967).

220. H. E. Spencer, R. E. Atwell, and M. Levy, *J. Photogr. Sci.* **31**, 158 (1983).
221. H. E. Spencer, *Photogr. Sci. Eng.* **11**, 352 (1967).
222. T. Tani, *J. Imaging Sci.* **29**, 93 (1985).
223. T. Tani, *J. Imaging Sci.* **30**, 41 (1986).
224. R. K. Hailstone, N. B. Liebert, M. Levy, and J. F. Hamilton, *J. Imaging Sci.* **35**, 219–230 (1991).
225. S. S. Collier, *Photogr. Sci. Eng.* **26**, 98 (1982).
226. A. G. DiFrancesco, M. Tyne, and R. Hailstone, *IS & T 49th Annual Conf.*, Minneapolis, MN, 1996, p. 222.
227. R. S. Eachus and M. T. Olm, *Crystl. Lattice Deformation Amorphous Mater.* **18**, 297 (1989).
228. R. S. Eachus and M. T. Olm, *Annu. Rep. Prog. Chem. C* **83**, 3 (1989).
229. D. A. Corrigan, R. S. Eachus, R. E. Graves, and M. T. Olm, *J. Chem. Phys.* **70**, 5,676 (1979).
230. A. P. Marchetti and R. S. Eachus, in D. Volman, G. Hammond, and D. Neckers, eds., *Advances in Photochemistry*, vol. 17, J Wiley, NY, 1992, pp 145–216.
231. R. S. Eachus and R. E. Graves, *J. Chem. Phys.* **65**, 5,445–5,452 (1976).
232. W. Bahnmüller, *Photogr. Korr.* **104**, 169 (1968).
233. H. Zwickey, *J. Photogr. Sci.* **33**, 201 (1985).
234. B. H. Carroll, *Photogr. Sci. Eng.* **24**, 265–267 (1980).
235. R. S. Eachus and R. E. Graves, *J. Chem. Phys.* **65**, 1,530 (1976).
236. M. T. Olm, R. S. Eachus, and W. S. McDugle, *Bulg. Chem. Comm.* **26**, 350–367 (1993).
237. W. West and P. B. Gilman Jr., in T. H. James, ed., *The Theory of the Photographic Process*, 4th ed., Macmillan, NY, 1977, Chap. 10.
238. D. M. Sturmer and D. W. Heseltine, in T. H. James, ed., *The Theory of the Photographic Process*, 4th ed., Macmillan, NY, 1977, Chap. 8.
239. H. Kuhn, *J. Chem. Phys.* **17**, 1,198–1,212 (1949).
240. A. H. Herz, in T. H. James, ed., *The Theory of the Photographic Process*, 4th ed., Macmillan, NY, 1977, Chap. 9.
241. T. Tani and S. Kikuchi, *Bull. Soc. Sci. Photogr. Jpn.* **18**, 1 (1968).
242. T. Tani and S. Kikuchi, *J. Photogr. Sci.* **17**, 33 (1969).
243. H. Matsusaki, H. Hada, and M. Tamura, *J. Soc. Photogr. Sci. Technol. Jpn.* **31**, 204 (1968).
244. J. F. Padday, *Trans. Faraday Soc.* **60**, 1,325 (1964).
245. B. H. Carroll and W. West, in J. W. Mitchell, ed., *Fundamentals of Photographic Sensitivity (Proc. Bristol Symp.)*, Butterworths, London, 1951, p. 162.
246. E. Günther and E. Moisar, *J. Photogr. Sci.* **13**, 280 (1965).
247. H. Phillippaerts, W. Vanassche, F. Cleaes, and H. Borginon, *J. Photogr. Sci.* **20**, 215 (1972).
248. W. West and A. L. Geddes, *J. Phys. Chem.* **68**, 837 (1964).
249. T. Tani and S. Kikuchi, *Bull. Soc. Sci. Photogr. Jpn.* **17**, 1 (1967).
250. G. R. Bird, K. S. Norland, A. E. Rosenoff, and H. B. Michaud, *Photogr. Sci. Eng.* **12**, 196–206 (1968).
251. L. G. S. Brooker, F. L. White, D. W. Heseltine, G. H. Keyes, S. G. Dent Jr., and E. J. Van Lare, *J. Photogr. Sci.* **1**, 173 (1953).
252. R. L. Large, in R. Cox, ed., *Photographic Sensitivity*, Academic Press, NY, 1973, pp. 241–263.
253. J. Lenhard, *J. Imaging Sci.* **30**, 27–35 (1986).
254. T. Tani, T. Suzumoto and K. Ohzeki, *J. Phys. Chem.* **94**, 1,298–1,301 (1990).
255. R. A. Marcus, *Annu. Rev. Phys. Chem.* **15**, 155 (1964).
256. R. W. Berriman and P. B. Gilman Jr., *Photogr. Sci. Eng.* **17**, 235–244 (1973).
257. J. R. Lenhard and B. R. Hein, *J. Phys. Chem.* **100**, 17,287 (1996).
258. A. A. Muenter, P. B. Gilman Jr., J. R. Lenhard, and T. L. Penner, The Int. East-West Symp. Factors Influencing Photogr. Sensitivity, 1984. Mauii, Hawaii, Paper C-4.
259. T. H. James, *Photogr. Sci. Eng.* **18**, 100–109 (1974).
260. W. C. Lewis and T. H. James, *Photogr. Sci. Eng.* **13**, 54–64 (1969).
261. J. M. Simson and W. S. Gaugh, *Photogr. Sci. Eng.* **19**, 339–343 (1975).
262. F. J. Evans and P. B. Gilman Jr., *Photogr. Sci. Eng.* **19**, 333–339 (1975).
263. D. M. Sturmer and W. S. Gaugh, *Photogr. Sci. Eng.* **19**, 344–351 (1975).
264. J. R. Lenhard, B. R. Hein, and A. A. Muenter, *J. Phys. Chem.* **97**, 8,269–8,280 (1993).
265. P. B. Gilman Jr., *Photogr. Sci. Eng.* **11**, 222 (1967).
266. P. B. Gilman Jr., *Photogr. Sci. Eng.* **12**, 230 (1968).
267. P. B. Gilman Jr., *Photogr. Sci. Eng.* **18**, 418 (1974).
268. W. E. Lee and E. R. Brown, in T. H. James, ed., *The Theory of the Photographic Process*, 4th ed., Macmillan, NY, 1977, Chap. 11.
269. T. H. James, in T. H. James, ed., *The Theory of the Photographic Process*, 4th ed., Macmillan, NY, 1977, Chaps. 13, 14.
270. G. Haist, *Modern Photographic Processing*, vols. 1 and 2, J Wiley, NY, 1979.
271. J. Kapecki and J. Rodgers, in *Kirk Othmer Encyclopedia of Chemical Technology*, vol. 6, 4th ed., Wiley, NY, 1993, pp. 965–1002.
272. J. R. Thirtle, L. K. J. Tong, and L. J. Fleckenstein, in T. H. James, ed., *The Theory of the Photographic Process*, 4th ed., Macmillan, NY, 1977, Chap. 12.
273. P. Krause, in J. Sturge, V. Walworth, and A. Shepp, ed., *Imaging Materials and Processes*, Neblette's 8th ed., Van Nostrand Reinhold, New York, 1989, Chap. 4.
274. J. F. Hamilton, *Appl. Opt.* **11**, 13 (1972).
275. C. R. Berry, *Photogr. Sci. Eng.* **13**, 65 (1969).
276. H. D. Keith and J. W. Mitchell, *Philos. Mag.* **44**, 877 (1953).
277. D. C. Skillman, *Photogr. Sci. Eng.* **19**, 28 (1975).
278. W. E. Lee, in T. H. James, ed., *The Theory of the Photographic Process*, 4th ed., Macmillan, NY, 1977, Chap. 14, Sect. II.
279. G. I. P. Levensen, in T. H. James, ed., *The Theory of the Photographic Process*, 4th ed., Macmillan, NY, 1977, Chap. 15.
280. M. A. Kriss, in T. H. James, ed., *The Theory of the Photographic Process*, 4th ed., Macmillan, NY, 1977, Chap. 21.
281. J. Dainty and R. Shaw, *Image Science*, Academic Press, NY, 1974.
282. M. R. Pointer and R. A. Jeffreys, *J. Photog. Sci.* **39**, 100 (1991).
283. L. Erickson and H. R. Splettstosser, in T. H. James, ed., *The Theory of the Photographic Process*, 4th ed., Macmillan, NY, 1977, Chap. 23, Sect. III.

SINGLE PHOTON EMISSION COMPUTED TOMOGRAPHY (SPECT)

MARK T. MADSEN
University of Iowa
Iowa City, IA

INTRODUCTION

Single photon emission computed tomography (SPECT) is a diagnostic imaging modality that produces tomographic slices of internally distributed radiopharmaceuticals. It is routinely used in diagnosing coronary artery disease and in tumor detection. Projection views of the radiopharmaceutical distribution are collected by one or more scintillation cameras mounted on a gantry designed to rotate about a patient lying horizontally on a pallet. The projection information is mathematically reconstructed to obtain the tomographic slices. Most clinical SPECT studies are qualitative and have simplistic corrections for attenuation and scattered radiation. Quantitative SPECT requires corrections for attenuation, scatter and spatial resolution, although these have not been routinely implemented in the past because of their computational load. SPECT instrumentation has evolved to include coincidence imaging of positron-emitting radiopharmaceuticals, specifically ^{18}F Fluorodeoxyglucose.

RADIOTRACERS

Much of medical imaging depends on anatomic information. Examples include radiographs, X-ray computed tomography (CT), and magnetic resonance imaging (MRI). In SPECT imaging, functional information is obtained about tissues and organs from specific chemical compounds labeled with radionuclides that are used as tracers. These radiotracers, or radiopharmaceuticals, are nearly ideal tracers because they can be externally detected and they are injected in such small quantities that they do not perturb the physiological state of the patient. A radionuclide is an unstable atomic nucleus that spontaneously emits energy (1). As part of this process, it may emit some or all of the energy as high energy photons called gamma rays. Because gamma rays are energetic, a significant fraction of them are transmitted from their site of origin to the outside of the body where they can be detected and recorded. Only a small number of radionuclides are suitable as radiotracers. The most commonly used radionuclides in SPECT imaging are summarized in Table 1. The phrase "single photon" refers to the fact that in this type of imaging, gamma rays are detected as individual events. The term is used to distinguish SPECT from positron-emitting emission tomography (PET) which also uses radionuclides, but relies on coincidence imaging. The radionuclides used in PET emit positrons, which quickly annihilate with electrons to form two, collinear 511-keV photons. Both of the annihilation photons have to be detected simultaneously by opposed detectors to record a true event, as discussed in more detail in the SPECT/PET Hybrid section.

Diagnostic information is obtained from the way the tissues and organs of the body process the radiopharmaceutical. For example, some tumor imaging uses radiopharmaceuticals that have affinity for malignant tissue. In these scans, abnormal areas are characterized by an increased uptake of the tracer. In nearly all instances, the radiopharmaceutical is administered to the patient by intravenous injection and is carried throughout the body by the circulation where it localizes in tissues and organs. Because SPECT studies require 15–30 minutes to acquire, we are limited to radiopharmaceuticals whose distribution will remain relatively constant over that or longer intervals. Ideally, we also want the radiopharmaceutical to distribute only in abnormal tissues. Unfortunately, this is never the case, and the abnormal concentration of the radiotracer is often obscured by normal uptake of the radiopharmaceutical in surrounding tissues. This is why tomographic imaging is crucial. It substantially increases the contrast of the abnormal area, thereby greatly improving the likelihood of detection. The widespread distribution of the radiopharmaceutical in the body has other implications; the most important is radiation dose. The radiation burden limits the amount of radioactivity that can be administered to a patient, and for most SPECT studies, this limits the number of detected emissions and thereby, the quality of the SPECT images.

SPECT studies are performed for a wide variety of diseases and organ systems (2–6). Although myocardial perfusion imaging (Fig. 1) and tumor scanning (Fig. 2) are by far the most common SPECT applications, other studies include brain perfusion for evaluating stroke (Fig. 3) and dementias, renal function, and the evaluation of trauma.

Table 1. SPECT Radionuclides

Radionuclide	Decay Mode	Production Method	Half-Life	Principal photon Emissions (keV)
^{99m}Tc	IT	^{99}Mo generator	6 h	140
^{67}Ga	EC	^{68}Zn(p,2n)^{67}Ga	78 h	93, 185, 296
^{111}In	EC	^{111}Cd(p,n)^{111}In	67 h	172, 247
^{123}I	EC	^{124}Te(p,5n)^{123}I	13 h	159
^{131}I	β-	Fission by-product	8 days	364
^{133}Xe	β-	Fission by-product	5.3 days	30(X rays), 80
^{201}Tl	EC	^{201}Hg(d,2n)^{201}Tl	73 h	60–80(X rays), 167
^{18}F	β+	^{18}O(p,n)^{18}F	110 min	511

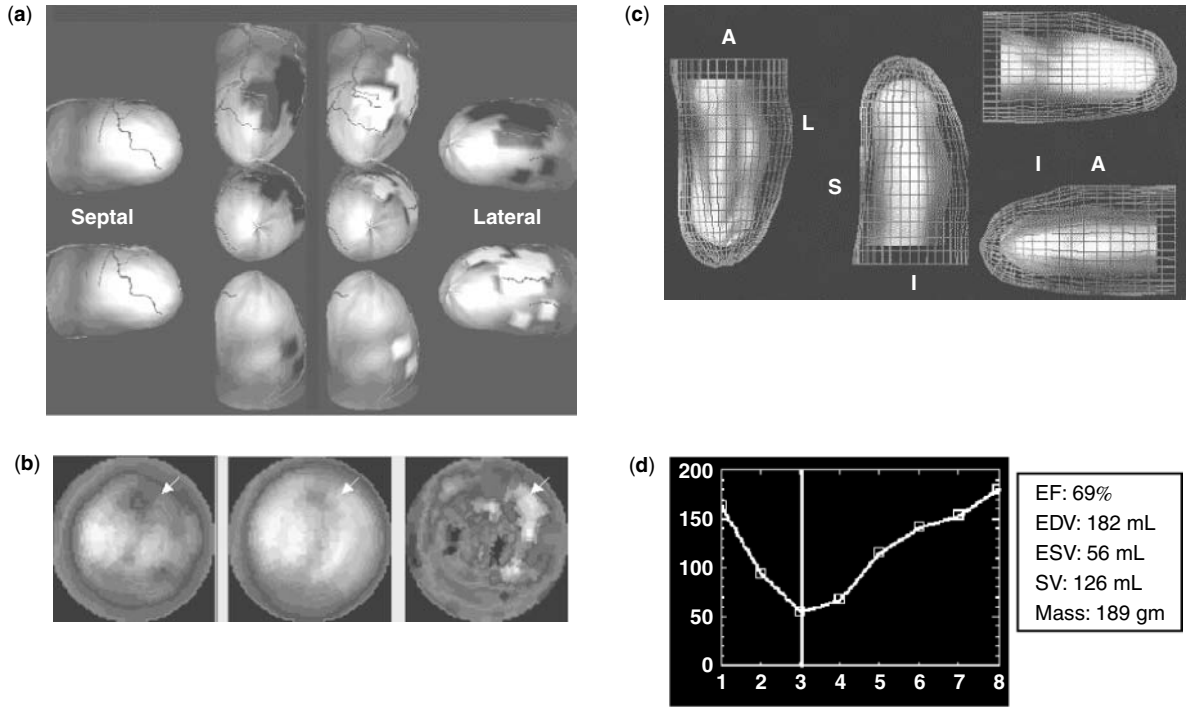


Figure 1. Myocardial perfusion SPECT. This common SPECT procedure is used to evaluate coronary artery disease. SPECT images show regional blood flow in the heart muscle under resting and stress conditions. Both bullet (**a**) and bull's eye displays (**b**) are used to compare the 3-D rest and stress images. Myocardial SPECT studies can also be gated to evaluate wall motion (**c**) and ejection fraction (**d**).

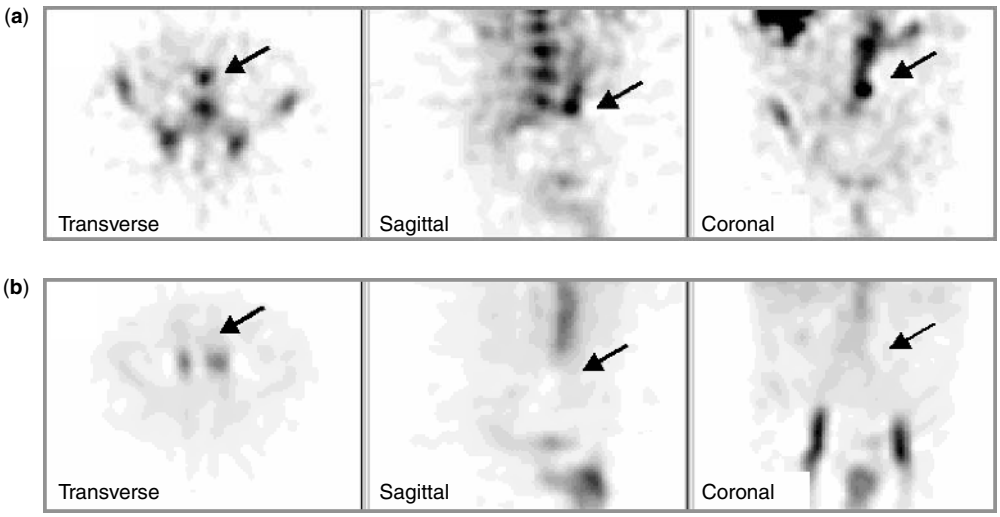


Figure 2. Tumor scanning. This shows images from scan for prostate cancer. Because of the difficulty in distinguishing abnormal uptake from circulating tracer, additional studies using a nonspecific radiotracer are acquired simultaneously. The upper set of images (**a**) shows tumor uptake (arrows) and no corresponding uptake in the corresponding blood pool image in the lower set (**b**).

Table 2 summarizes several of the most common SPECT studies along with radiation dose estimates.

Gamma Ray Interactions

To understand the detection and imaging of gamma rays, we must first review gamma ray interactions with different materials (7). The intensity of a gamma ray beam decreases as it traverses through a material because of interactions between the gamma rays and the electrons in the material. This is referred to as attenuation. Attenuation is an exponential process described by

$$I(x) = I_o \exp(-\mu x), \tag{1}$$

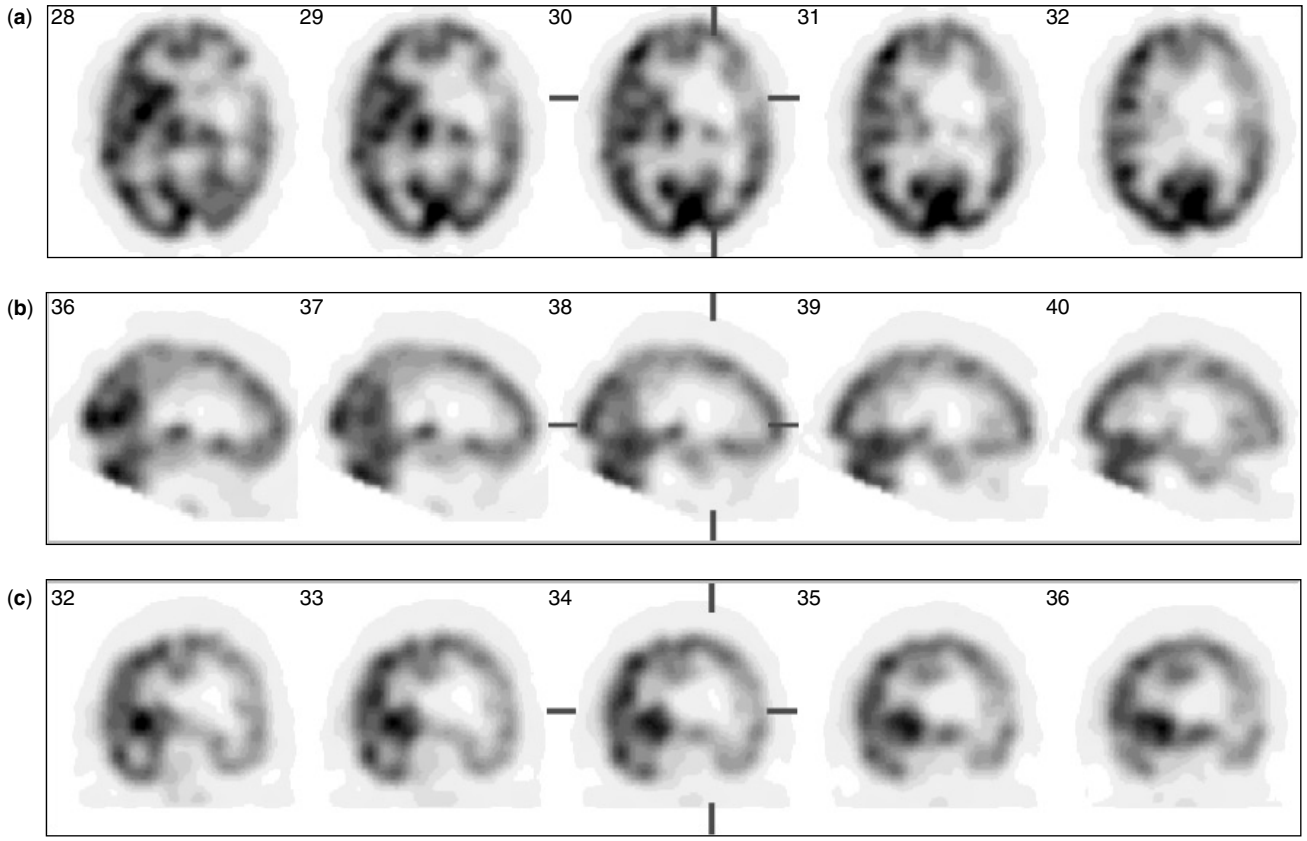


Figure 3. Brain perfusion study. These images show the asymmetric distribution of blood flow in the brain resulting from a stroke. Selected transverse (**a**), sagittal (**b**), and coronal views (**c**) are shown.

Table 2. SPECT Radiopharmaceuticals

Radiopharmaceutical		Application	Effective Dose (rem)
^{99m}Tc	Medronate(MDP), Oxidronate(HDP)	Bone scan	0.75
^{99m}Tc	Exametazine(HMPAO), Bicisate(ECD)	Brain perfusion	1.2
^{99m}Tc	Arcitumomab(CEA Scan)	Colon cancer	0.75
^{99m}Tc	Sestamibi, Tetrofosmin	Myocardial perfusion, breast cancer	1.2
^{123}I	Metaiodobenzylguanidine (MIBG)	Neuroendocrine tumors	0.7
^{67}Ga	Citrate	Infection, tumor localization	2.5
^{111}In	Capromab Pendetide(ProstaScint)	Prostate cancer	2.1
^{111}In	Pentetreotide(OctreoScan)	Neuroendocrine tumors	2.1
^{201}Tl	Thallous Chloride	Myocardial perfusion	2.5
^{18}F	Fluoro-2-deoxyglucose(FDG)	Tumor localization, Myocardial viability	1.1

where I_o is the initial intensity, $I(x)$ is the intensity after traveling a distance x through the material, and μ is the linear attenuation coefficient of the material. Over the range of gamma ray energies used in radionuclide imaging, the two primary interactions that contribute to the attenuation coefficient are photoelectric absorption and Compton scattering. Photoelectric absorption refers to the total absorption of the gamma ray by an inner shell atomic electron. It is not an important interaction in body tissues, but it is the primary interaction in high

Z materials such as sodium iodide (the detector material used in the scintillation camera) and lead. Photoelectric absorption is inversely proportional to the cube of the gamma ray energy, so that the efficiency of detection falls sharply as photon energy increases.

Compton scattering occurs when the incoming gamma ray interacts with a loosely bound outer shell electron. A portion of the gamma ray energy is imparted to the electron, and the remaining energy is left with the scattered photon. The amount of energy lost in scattering depends on the angle between the gamma ray and the scattered photon. The cross section of Compton scattering is inversely proportional to the gamma ray energy, and it is the dominant interaction in body tissues.

Scintillation Cameras

All imaging studies in nuclear medicine (SPECT and conventional planar) are acquired on scintillation cameras (also referred to as Anger or gamma cameras) invented by H. O. Anger in 1953 (8,9). The detector of the scintillation camera is a large, thin sodium iodide crystal (Fig. 4). Typical dimensions of the crystal are 40×50 cm and 9.5 mm thick. Sodium iodide, NaI(Tl), is a scintillator; it converts absorbed gamma ray energy into visible light. The magnitude of the light flash is proportional to the energy absorbed, so that information about the event energy as well as location is available. Photomultiplier tubes, which convert the scintillation into an electronic pulse, are arranged in a close-packed array that covers the entire sensitive area of the crystal. Approximately sixty 7.5-cm photomultiplier tubes are required for the scintillation camera dimensions given before. The location of the detected event is determined by the position-weighted average of the electronic pulses generated by the photomultiplier tubes in the vicinity of the event. This approach yields an intrinsic spatial resolution in the range of 3–4 mm. In addition to estimating the position of the event, the photomultiplier tube signals are also combined to estimate the energy absorbed in the interaction. The energy signal is used primarily to discriminate against Compton scattered radiation that occurs in the patient and to normalize the position signals so that the size of the image does not depend on the gamma ray energy. It also makes it possible to image distributions of radiotracers labeled with different radionuclides simultaneously. This is often referred to as dual isotope imaging; however, modern gamma cameras can acquire simultaneous images from four or more energy ranges.

Because the response of the crystal and photomultiplier tubes is not uniform, additional corrections are made for position-dependent shifts in the energy signal (referred to as Z or energy correction) and in determining the event location (referred to as L or spatial linearity correction). Thus, when a gamma ray is absorbed, the scintillation camera must determine the position and energy of the event, determine if the energy signal falls within a selected pulse height analyzer window, and apply spatial linearity correction. At this point, the location within the image matrix corresponding to the event has its count value increased by one. A scintillation camera image is generated from the accumulation of many (10^5–10^6) detected events.

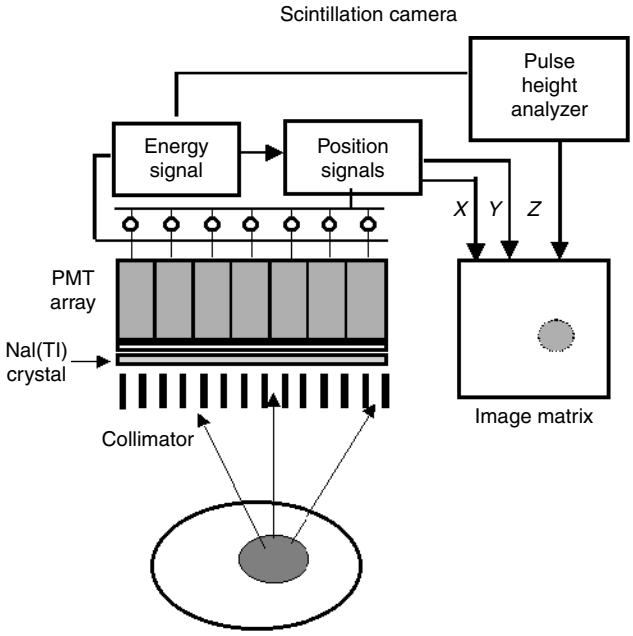


Figure 4. Scintillation camera. Virtually all SPECT imaging is performed with scintillation cameras. The scintillation camera determines the location and energy of each gamma ray interaction through the weighted averaging of photomultiplier signals. A collimator is required to form the image on the NaI(Tl) detector.

The time it takes to process an event is ultimately limited by the scintillation relaxation time [$t = 250$ ns for NaI(Tl)]. For most SPECT imaging, this does not present any problem. However, it becomes a severe constraint for coincidence imaging, discussed in detail later. Typical performance specifications of scintillation cameras are given in Table 3.

Gamma rays cannot be focused because of their high photon energy. Therefore, a collimator must be used to project the distribution of radioactivity within the patient onto the NaI(Tl) crystal (10). A collimator is a multihole lead device that selectively absorbs all gamma rays except those that traverse the holes (Fig. 5). This design severely restricts the number of gamma rays that can be detected. Less than 0.05% of the gamma rays that hit the front

Table 3. Scintillation Camera Specifications

Parameter	Specification
Crystal size	40×50 cm
Crystal thickness	9.5 mm
Efficiency at 140 keV	0.86 (photopeak); 0.99 (total)
Efficiency at 511 keV	0.05 (photopeak); 0.27 (total)
Energy resolution	10%
Intrinsic spatial resolution	3.5 mm
System count sensitivity[a]	250 counts/min/μCi
System spatial resolution at 10 cm[a]	8.0 mm
Maximum count rate(SPECT)	250,000 counts/s
Maximum count rate (coincidence)	>1,000,000 counts/s

[a]Specified for high-resolution collimator.

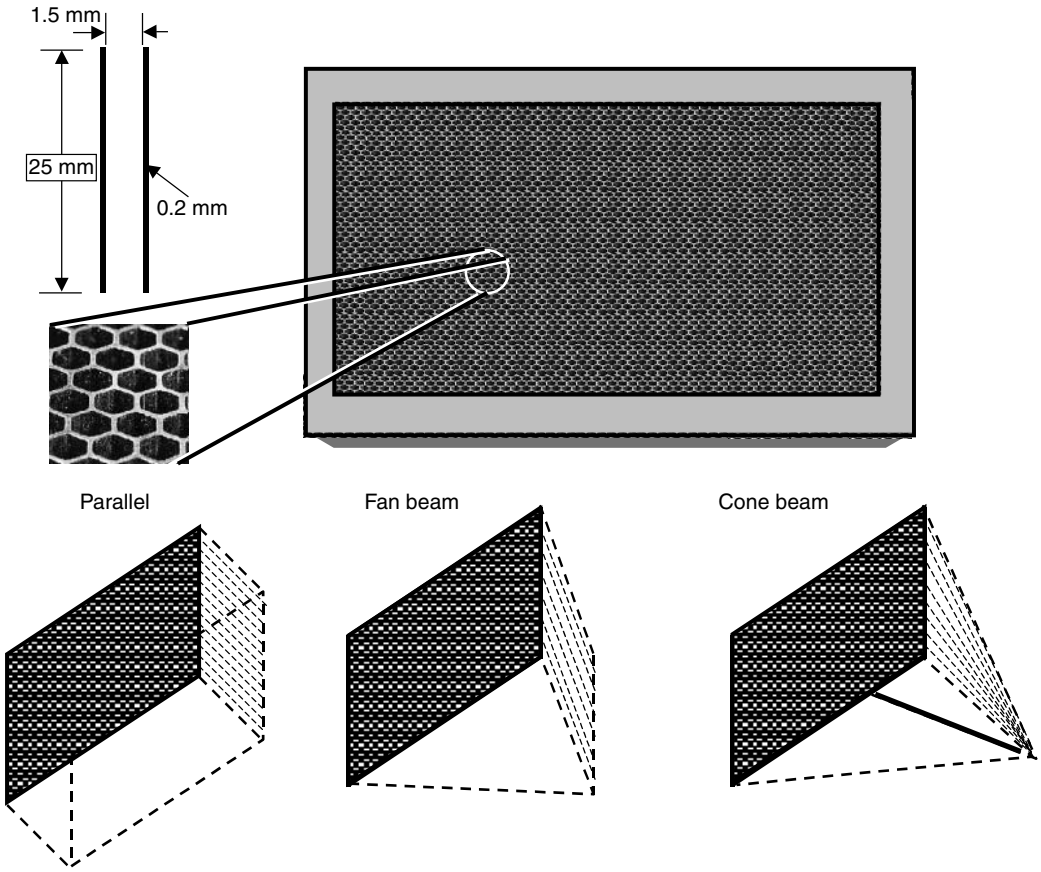


Figure 5. Collimators. Collimators are the image-forming apertures of the scintillation camera. They can be configured in parallel, fan-beam, and cone-beam geometries.

surface of the collimator are transmitted through to the crystal. Several parameters enter in the design of collimators. Most collimators have parallel holes that map the gamma ray distribution one-to-one onto the detector. Trade-offs are made in optimizing of the design for count sensitivity and spatial resolution. The sensitivity of the collimator is proportional to the square of the ratio of the hole size d and length l) ($\varepsilon \propto d^2/l^2$). The spatial resolution (R_{col}), characterized by the full-width-at-half-maximum (FWHM) of the line spread function, is proportional to d/l. The desire is to maximize ε while minimizing R_{col}. Because the optimal design often depends on the specific imaging situation, most clinics have a range of collimators available. Some typical examples are given in Table 4. For low-energy studies ($E\gamma < 150$ keV) either high-resolution or ultrahigh resolution collimators are typically used. Because the lead absorption of gamma rays is inversely proportional to the gamma ray energy, the design of collimators is influenced by the gamma ray energies that are imaged. As the photon energy increases, thicker septa are required, and to maintain count sensitivity, the size of the holes is increased which compromises spatial resolution.

Parallel hole geometry is not the most efficient arrangement. Substantial increases in count sensitivity are obtained by using fan- and cone-beam geometries (11) (Fig. 5). The disadvantage of these configurations is that the field of view becomes smaller as the source-to-collimator distance increases. This presents a problem for SPECT imaging in the body where portions of the radiopharmaceutical distribution are often truncated. Fan-beam collimators are routinely used for brain imaging, and hybrid cone-beam collimators are available for imaging the heart.

In addition to the dependence on hole size and length, the spatial resolution of a collimator depends on the source-to-collimator distance, as shown in Fig. 6. The overall system spatial resolution R_{sys} can be estimated from

$$R_{\text{sys}} = \sqrt{R_{\text{col}}^2 + R_{\text{int}}^2}. \tag{2}$$

For most imaging, the collimator resolution is substantially larger than the intrinsic spatial resolution ($R_{\text{int}} \sim 3.5$ mm) and is the dominant factor. Therefore, it is very important for the collimator to be as close to the patient as possible.

SPECT Systems

A SPECT system consists of one or more scintillation cameras mounted on a gantry that can revolve about a fixed axis in space, the axis of rotation (8,9,12,13) (Fig. 7).

Table 4. Collimator Specifications

Collimator	Energy (keV)	Hole Size (mm)	Hole Length (mm)	Septal Thickness (mm)	Relative Sensitivity	R_{col} at 10 cm (mm)
Parallel general purpose	140	1.4	24	0.20	1.0	8.3
Parallel high-sensitivity	140	2.0	24	0.20	2.0	11.9
Parallel high-resolution	140	1.1	24	0.15	0.7	6.3
Parallel ultrahigh resolution	140	1.1	36	0.15	0.3	4.5
Ultrahigh fan-beam	140	1.4	35	0.20	0.8	7.1
Cone-beam	140	1.9	41	0.25	1.2	7.1
Parallel medium-energy	300	3.5	50	1.30	0.8	11.6
Parallel high-energy	364	4.2	63	1.30	0.5	13.4
Parallel ultrahigh energy	511	4.2	80	2.40	0.3	12

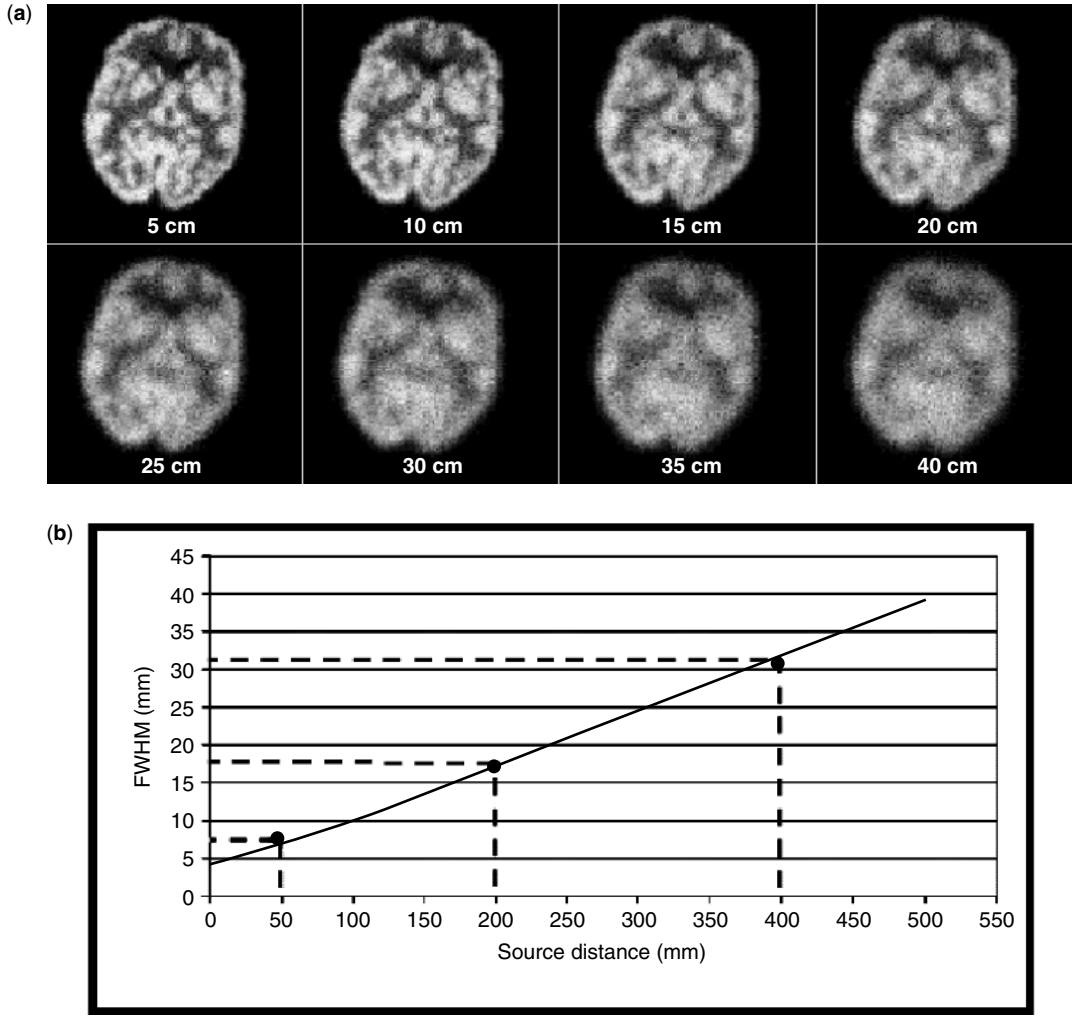


Figure 6. Collimator spatial resolution as a function of source distance. The spatial resolution of a collimator falls continuously as the source distance increases. This is shown in the quality of the images (**a**) and the plot of the calculated FWHM (**b**).

SPECT studies are usually acquired over a full 360° arc. This yields better quality images than 180° acquisitions because it tends to compensate somewhat for the effects of attenuation. One exception to this practice is myocardial perfusion studies, which are acquired using views from only 180° (see later). SPECT acquisitions are performed either in step-and-shoot mode or in a continuous rotational mode. In the step-and-shoot mode, the detector rotates to its angular position and begins collecting data after the detector stops for a preselected frame duration. In the

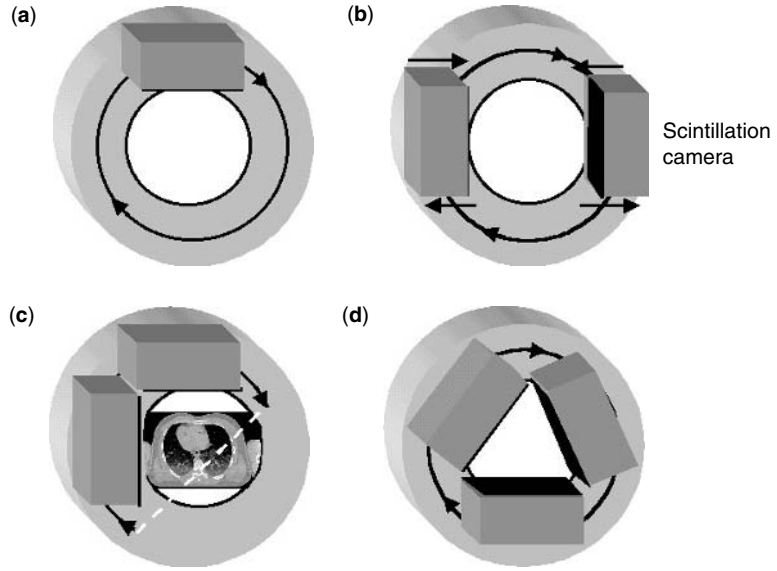


Figure 7. SPECT systems. SPECT systems consist of one or more scintillation cameras mounted on a gantry that allows image collection from 360° around a patient. The most common configuration has two scintillation cameras. To accommodate the 180° sampling of myocardial perfusion studies, many systems can locate the scintillation cameras at either 90° or 180°.

continuous rotational mode, the duration of the entire study is selected, and the detector rotational speed is adjusted to complete one full orbit. Data are collected continually and are binned into a preselected number of projections. Typically 60 to 120 projection views are acquired over 360°. Another feature of SPECT acquisition is body contouring of the scintillation cameras. Because spatial resolution depends on the source-to-collimator distance, it is crucial to maintain close proximity to the body as the detector rotates about the patient. Although a number of different approaches have been used to accomplish this, the most common method moves the detectors radially in and out as a function of rotational angle.

Myocardial perfusion studies are the most common SPECT procedures. Because the heart is located in the left anterior portion of the thorax, gamma rays originating in the heart are highly attenuated for views collected from the right lateral and right posterior portions of the arc. For this reason, SPECT studies of the heart are usually collected using the 180° arc that extends from the left posterior oblique to the right anterior oblique view (14) (Fig. 7c). This results in reconstructed images that have the best contrast, although distortions are often somewhat more pronounced than when 360° data are used (15). Because of the widespread use of myocardial perfusion imaging, many SPECT systems have been optimized for 180° acquisition by using two detectors arranged at ~90° (Fig. 7c). This reduces the acquisition time by a factor of 2 compared to single detectors and is approximately 30% more efficient than triple-detector SPECT systems. Positioning the detectors at 90° poses some challenges for maintaining close proximity. Most systems rely on the motion of both the detectors and the SPECT table to accomplish this.

The heart is continually moving during the SPECT acquisition, and this further compromises spatial resolution. Because the heart beats many times per minute, it is impossible to acquire a stop-action SPECT study directly. However, the heart's motion is periodic, so it is possible to obtain this information by gating the SPECT acquisition (16). In a gated SPECT acquisition, the cardiac cycle is subdivided, and a set of eight images that span the ECG R–R interval is acquired for each angular view. These images are placed into predetermined time bins based on the patient's heart rate, which is monitored by the ECG R wave interfaced to the SPECT system. As added benefits of gating, the motion of the heart walls can be observed, and ventricular volumes and ejection fractions can be determined (17) (Fig. 1).

Although most SPECT imaging samples a more or less static distribution of radionuclides, some SPECT systems can perform rapid sequential studies to monitor tracer clearance. An example of this is determining regional cerebral blood from the clearance of ^{133}Xe (18). Multiple 1-minute SPECT studies are acquired over a 10-minute interval. When one acquisition sample is completed, the next begins automatically. To minimize time, SPECT systems that perform these studies can alternately reverse the acquisition direction, although at least one SPECT system uses slip-ring technology, so that the detectors can rotate continuously in the same direction.

SPECT Image Reconstruction

The projection information collected by the SPECT system has to be mathematically reconstructed to obtain tomographic slices (8,19–21). The information sought is the distribution of radioactivity for one selected transaxial plane, denoted by $f(x, y)$. A projection through this distribution consists of a set of parallel line integrals,

$p(s, \theta)$, where s is the independent variable of the projection and θ is the angle along which the projection is collected. Ignoring attenuation effects for the moment,

$$p(s, \theta) = \int_{-\infty}^{\infty} f(x', y')\, dt, \tag{3}$$

where

$$x' = s\cos\theta - t\sin\theta;$$
$$y' = s\sin\theta + t\cos\theta. \tag{4}$$

If a complete set of projections can be collected over θ, then one can analytically reconstruct $f(x, y)$ by using several different equivalent methods. The most common approach used in SPECT is filtered backprojection:

$$f(x, y) = \int_{0}^{2\pi} p^*[x\cos(\theta) + y\sin(\theta)]\, d\theta, \tag{5}$$

where p^* is the measured projection altered by a reconstruction filter $[R(\omega)]$:

$$p^*(s, \theta) = \mathrm{FT}^{-1}\{\mathrm{FT}[p(s, \theta)] \times R(\omega)\}. \tag{6}$$

For an ideal projection set (completely sampled and no noise), $R(\omega) = |\omega|$ and is commonly referred to as a ramp filter. The amplification of high-frequency information by this filter requires adding a low-pass filter when real projections are reconstructed, as discussed later.

Operationally, SPECT imaging proceeds as follows. Projection views are collected with a scintillation camera at multiple angles about the patient. The field of view of the scintillation camera is large, so that information is acquired from a volume where each row of the projection view corresponds to a projection from a transaxial plane. To reconstruct a tomographic slice, the projections associated with that plane are gathered together, as shown in Fig. 8. This organization of the projections by angle is often referred to as a sinogram because each source in the plane completes a sinusoidal trajectory. The reconstruction filter is usually applied to the projections in the frequency domain, and the filtered projections are then backprojected to generate the tomographic slice (Fig. 9).

The noise level of the acquired projections is typically high. When the ramp reconstruction filter is applied, amplification of the noise-dominant higher frequencies overwhelms the reconstructed image. To prevent this, the reconstruction filter is combined with a low-pass filter (apodization). Many different low-pass filters have been used in this application. One common example is the Butterworth filter.

$$B(\omega) = \frac{1}{\left[1 + \left(\frac{\omega}{\omega_c}\right)^{2N}\right]}, \tag{7}$$

and the apodized reconstruction filter is

$$R(\omega) = \frac{|\omega|}{\left[1 + \left(\frac{\omega}{\omega_c}\right)^{2N}\right]}. \tag{8}$$

The adjustable parameters, the cutoff frequency (ω_c), and the order (N) allow configuring the reconstruction filter for different imaging situations. A low-cutoff frequency is desirable when the projections are noisy. When the count density's high, a low cutoff yields an overly smoothed result, as shown in Fig. 10.

An accurate description of SPECT imaging requires including attenuation. A more appropriate model of the measured projections (excluding scatter and resolution

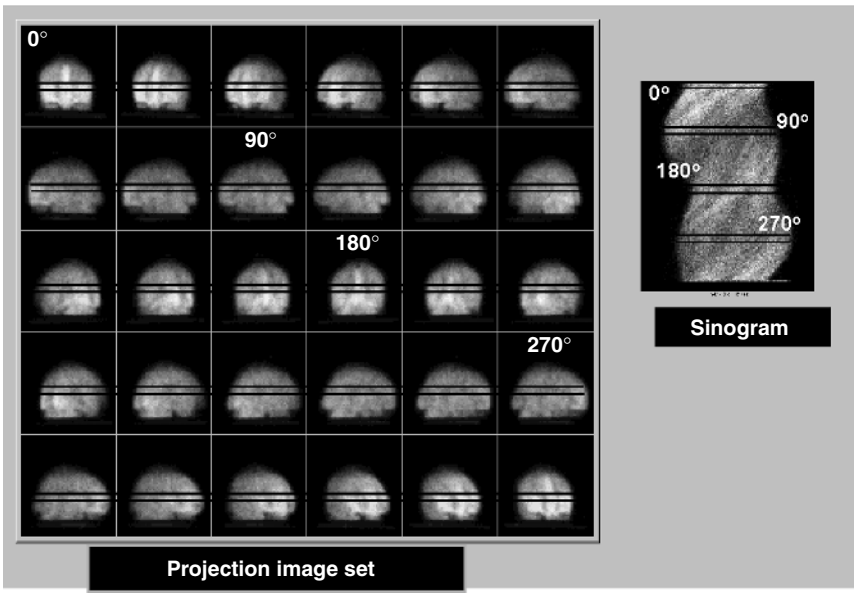


Figure 8. Projection data sets. The scintillation camera collects information from multiple projections in each view. A projection set consists of a stack of image rows from each of the angular views. The organization of projections by angle is commonly referred to as a sinogram. Madsen, M.T. Introduction to emission CT. Radiographics 15:975–991, 1995.

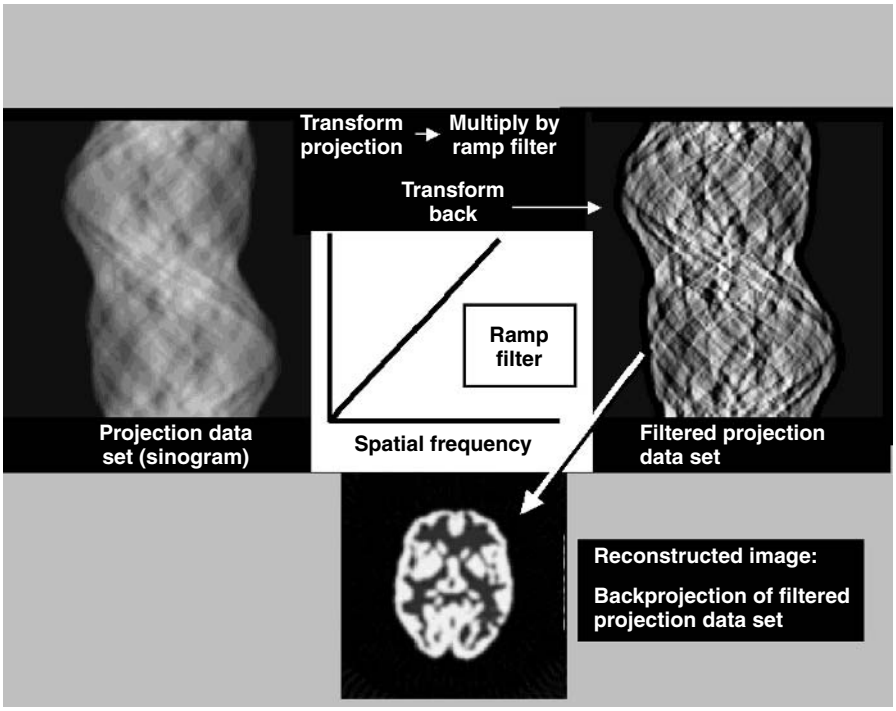


Figure 9. Filtered backprojection reconstruction. The projections are modified by the reconstruction filter and then backprojected to yield the tomographic image. For an ideal projection set, the ramp filter provides an exact reconstruction.

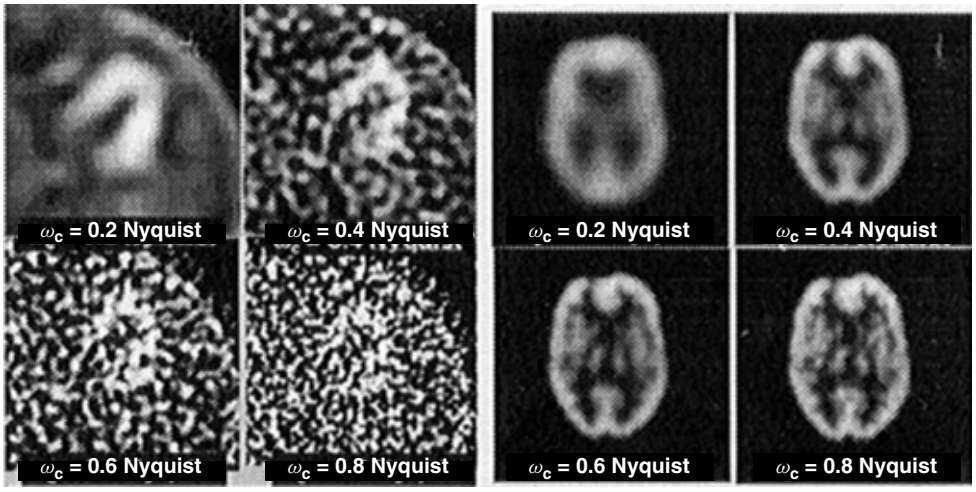


Figure 10. SPECT reconstruction filters and noise. Because of the statistical fluctuations in the projection views, it is necessary to suppress the high-frequency components of the ramp reconstruction filter by selecting appropriate filter parameters. Because noise suppression also reduces detail, the optimal filter choice depends on the organ system imaged and the count density. Madsen, M.T. Introduction to emission CT. Radiographics 15:975–991, 1995.

effects) is

$$p(s,\theta) = \int_{-\infty}^{\infty} f(x',y') \exp\left[-\int_{t'}^{\infty} \mu(x',y')\,dt\right] dt. \qquad (9)$$

This formulation is known as the attenuated Radon transform. Unfortunately, there is no analytic solution to this problem. Until recently, there were two ways of handling this problem for clinical studies. The first, and still very common, is simply to reconstruct the acquired projections using filtered backprojection and accept the artifacts that accompany the inconsistent data set. The second is to apply a simple attenuation correction. The most commonly used attenuation correction is the first-order Chang method in which a calculated correction map is applied to the reconstructed images (22). The correction factors are calculated by assuming that activity is uniformly distributed in a uniformly attenuating elliptical contour. The size of the ellipse is determined from the anterior and lateral projections. This attenuation correction is fairly adequate for parts of the body such as the abdomen and head but is not useful for the thorax where the assumptions are far from valid.

The other approach to image reconstruction uses iterative algorithms (23,24) (Fig. 11). Iterative algorithms are more time-consuming than filtered backprojection, but they have several important advantages. These include the elimination of radial streak artifacts, often seen in images reconstructed using filtered backprojection; accurate correction of physical degradations such as scatter, attenuation and spatial resolution; and better performance where a wide range of activities is present or where limited angle data are available. Iterative algorithms for image reconstruction were introduced in the 1970s resulting from the advent of X-ray computed tomography. These algorithms were extensions of general approaches to solving linear systems by using sparse matrices. Significant progress in iterative algorithms for emission computed tomography was made in 1982 when the maximum likelihood expectation maximization (ML-EM) algorithm of Shepp and Vardi was introduced (25). In the ML-EM approach, the Poisson nature of the gamma rays is included in the derivation. The likelihood that the measured projections are consistent with the estimated emission distribution is maximized to yield

$$\lambda_j^{new} = \frac{\lambda_j}{\sum_i c_{ij}} \sum_i \frac{c_{ij} y_i}{b_i}, \tag{10}$$

where

$$b_i = \sum_k c_{ik} \lambda_k \tag{11}$$

In this formulation, λ_j is the emission distribution (i.e., the SPECT image), y is the set of measured projections, and b is the set of calculated projections from the current estimate of λ. The c_{ij} are backprojection weighting factors that can also encompass appropriate factors for other physical effects such as attenuation, spatial resolution, and scatter. This yields an algorithm with several nice features. First, it is easy to see that because the updating of the estimate in each iteration depends on a ratio, it automatically restricts results to positive numbers. Second, the algorithm conserves the total image counts in each iteration. Unfortunately, the ML-EM algorithm converges slowly, and 20–50 iterations are often required for a satisfactory result.

One reason for the slow convergence of the ML-EM algorithm is that the SPECT estimate is updated only at the end of each iteration. One way of significantly reducing the number of iterations is the ordered subset (OS-EM) approach introduced by Hudson and Larkin (26). Using OS-EM, the projection set is split into multiple equal-sized projection sets. For example, a projection set of 64 angular samples might be split into eight subsets of eight samples each. The members of each set are

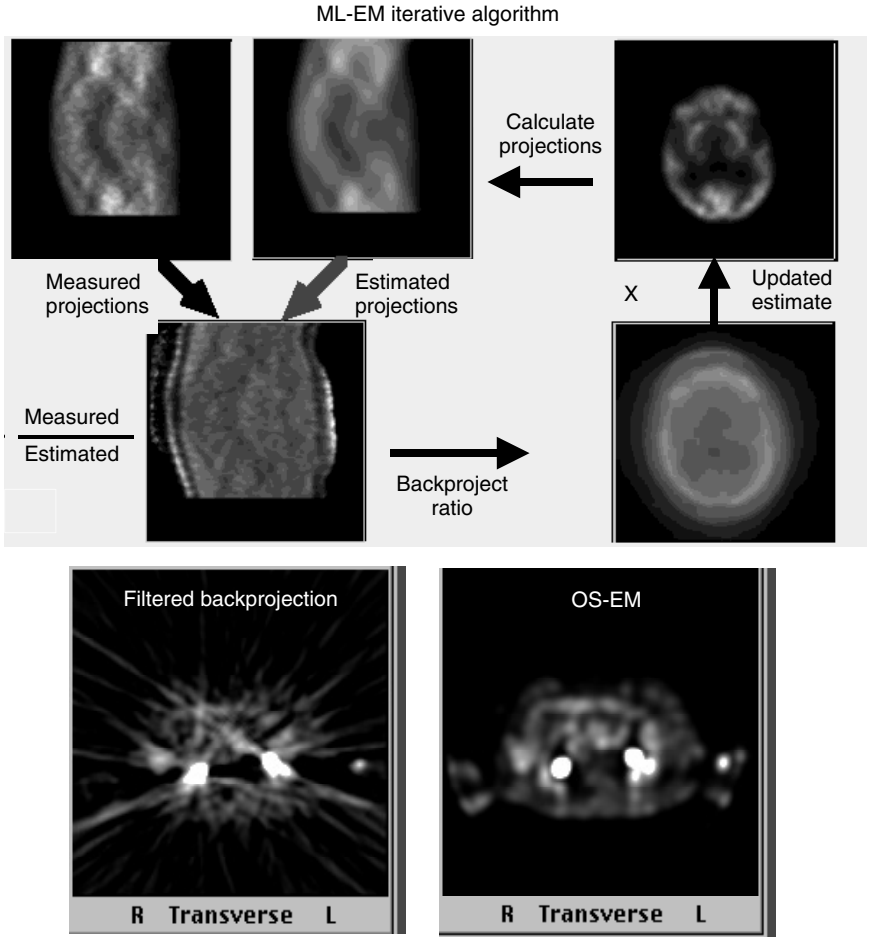


Figure 11. Iterative reconstruction. In iterative reconstruction, the initial uniform estimate of the tomographic slice is continually updated by backprojecting the ratio of the measured and calculated projections from the latest estimate. Although computationally intensive, iterative reconstructions allow accurate correction for attenuation and other physical degradations. They also reduce streak artifacts and perform better than filtered backprojection when the projection set is undersampled.

chosen to span the angular range. Set one would consist of projections 1, 9, 17, 25, 33, 41, 49, and 57, and set two would have projections 2, 10, 18, 26, 34, 42, 50, and 58, and so on. The ML-EM algorithm is applied sequentially to each subset, and one iteration is completed when all the subsets have been operated on. Thus, the estimated emission distribution is updated multiple times in each iteration. This approach decreases the number of required iterations by a factor approximately equal to the number of subsets. As a result, five or less iterations of the OS-EM algorithm are sufficient for most SPECT reconstructions. Although strict convergence has not been demonstrated for the OS-EM algorithm, it is now the most commonly used iterative algorithm for emission tomographic applications.

The reconstruction process produces a set of contiguous transaxial slices. These slices can be viewed individually, in a group, or even in a cine format. However, often the best way to visual the information is by using views that are parallel to the long axis of patients. These views can be generated directly from the transaxial slices. Sagittal slices are oriented at 0° (the side of the body) and proceed laterally from the right to the left side. Coronal slices are oriented at 90° (the front of the body) and proceed from posterior to anterior. These orientations are useful because many of the organs are aligned with the long axis of the body. An exception is the heart, which points down and to the left. Oblique views parallel and perpendicular to the long axis of the heart are generated for myocardial perfusion studies (see Figs. 1 and 14). Because myocardial SPECT is common, automatic routines exist to generate these views.

The transverse, sagittal, and coronal views are very useful, but they require that the observer view multiple slices. Another useful display is the maximum pixel intensity reprojection. Projections through the reconstructed slice volumes are calculated for typically 20–30 viewing angles over 360°. Instead of summing the information, the highest count pixel value is projected for each ray. Often this value will also be distance weighted. Then, the set of reprojected images is viewed in a cine format yielding a high-contrast, three-dimensional display. The maximum pixel reprojection displays are most useful for radiopharmaceuticals that accumulate in abnormal areas. Examples of these SPECT displays are shown in Fig. 12.

SPECT imaging is susceptible to many artifacts if not performed carefully (27,28). Many of the artifacts are a direct consequence of the fundamental assumptions of tomography. The primary assumption is that the external measurements of the distribution reflect true projections, i.e., line integrals. It has already been noted that attenuation and scatter violate this assumption. In addition, it is critical that an accurate center of rotation is used in the backprojection algorithm. The center of rotation is the point on a projection plane that maps the center of the image field, and it must be known to within one-half pixel. Errors larger than this distort each reconstructed point into a "doughnut" shape (Fig. 13a). It

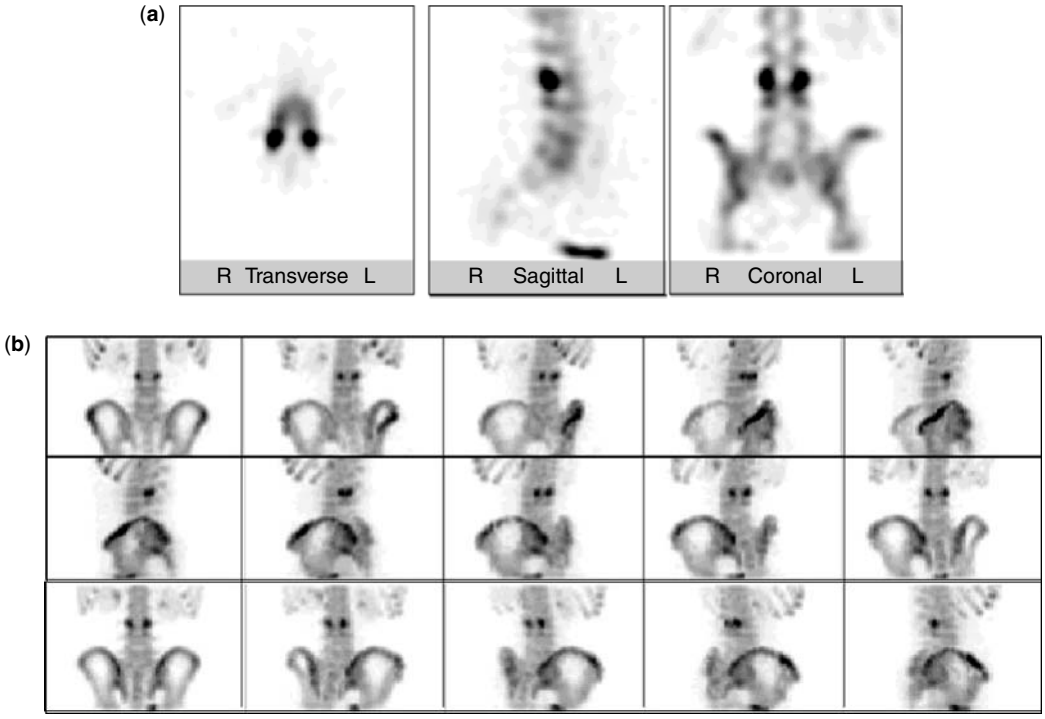


Figure 12. SPECT tomographic displays. Because SPECT uniformly samples a large volume, multiple transverse slices are available. This data can be resliced (**a**) to yield sagittal (side views parallel to the long axis of the body) and coronal (front views parallel to the long axis of the body), or any oblique view. Another useful display is (**b**) the maximum pixel reprojection set that is often viewed in cine mode.

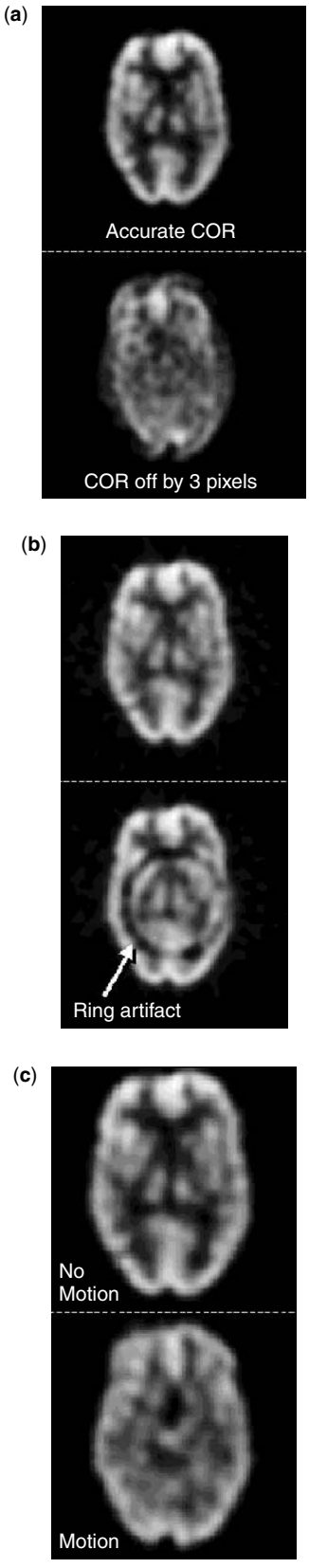


Figure 13. SPECT artifacts. SPECT images are susceptible to a variety of artifacts. (**a**) Inaccurate center of rotation values blurs each image point. (**b**) Nonuniformities in the scintillation camera cause ring artifacts. (**c**) Motion during SPECT acquisition can cause severe distortions.

is also a fundamental assumption of emission tomography that the detector field is uniform. Nonuniformities in the detector result in ring-shaped artifacts (Fig. 13b). This problem is especially acute for uniformities that fall near the center of rotation, where deviations as small as 1% can result in a very distinct artifact. A change of the source distribution from radiotracer kinetics or patient motion also induces artifacts (Fig. 13c).

Although much of SPECT imaging today uses the methods described before, the equations presented earlier oversimplify of the actual imaging problem. Both the spatial resolution of SPECT system and the scatter contributions correlate information from other planes. An accurate description of SPECT requires a 3-D formulation such as (29)

$$p(s,\theta) = c \int_{\Omega} h(t,\omega;r) \int_{-\infty}^{\infty} f(r) \exp\left[- \int_{r}^{\infty} \mu(u)\, dt' \right] dt\, d\omega. \tag{12}$$

Here $h(t,\omega;r)$ represents a three-dimensional system transfer function that includes both the effects of spatial resolution and scattered radiation, $f(r)$ is the emission distribution and $\mu(u)$ is the attenuation distribution. This more accurate model has not been routinely implemented for clinical situations because of high computational costs. However, investigations have shown measurable improvements in image quality, and it is likely that the 3-D formulation will be standard in the near future. This is discussed in greater detail in the Quantitative SPECT section.

SPECT SYSTEM PERFORMANCE

The system performance of SPECT is summarized in Table 5. The scintillation cameras and the associated collimation determine the count sensitivity of a SPECT system. SPECT spatial resolution is generally isotropic and has a FWHM of 8–10 mm for brain imaging and 12–18 mm for body imaging. Spatial resolution is affected by the collimation, the organ system imaged, and the radiopharmaceutical used. This becomes clear when the components of the spatial resolution are examined. SPECT

Table 5. SPECT System Performance (High-Resolution Collimator)

Parameter	Specification
Number of scintillation cameras	1, 2 or 3
Count sensitivity per camera	250 cpm/μCi per detector
Matrix size	64 × 64; 128 × 128
Pixel size	6 mm; 3 mm
Spatial resolution (brain studies)	8 mm
Spatial resolution (heart studies)	14 mm
SPECT uniformity	15%
Contrast of 25.4-mm sphere[a]	0.45

[a] Measured in a cylindrical SPECT phantom of 22-cm diameter at a detector orbit radius of 20 cm (58).

spatial resolution is quantified by

$$R_{\mathrm{SPECT}} = \sqrt{R_{\mathrm{col}}^2 + R_{\mathrm{filter}}^2 + R_{\mathrm{int}}^2}. \quad (13)$$

The intrinsic spatial resolution is a relatively minor factor in this calculation. The collimator resolution depends on the type of collimator selected and how close the scintillation camera can approach the patient during acquisition. The type of collimation depends on the expected gamma ray flux and the energy of the gamma rays emitted. If a ^{99m}Tc radiopharmaceutical is used that concentrates with high uptake in the organ of interest, then a very high-resolution collimator can be used to minimize the R_{col} component. However, if a high-energy gamma emitter such as ^{131}I is used, the appropriate collimator will perform significantly more poorly. Keeping the collimator close to the patient during acquisition is extremely critical for maintaining good spatial resolution. The best results are obtained for head imaging where a radius of less than 15 cm is possible for most studies. In the trunk of the body, it is difficult to maintain close proximity, and there is a corresponding loss of spatial resolution. In addition, count density is a major consideration. Low count density requires more smoothing within the reconstruction filter, and this imposes additional losses in spatial resolution.

QUANTITATIVE SPECT

As stated before, until recently, most SPECT imaging relied on either filtered backprojection with no corrections or corrections that use simple physical models that often are poor descriptors of the actual imaging situation (Fig. 14a,b). Routines are available in clinical SPECT systems for enhancing contrast and suppressing noise by using Metz or Wiener filters in conjunction with

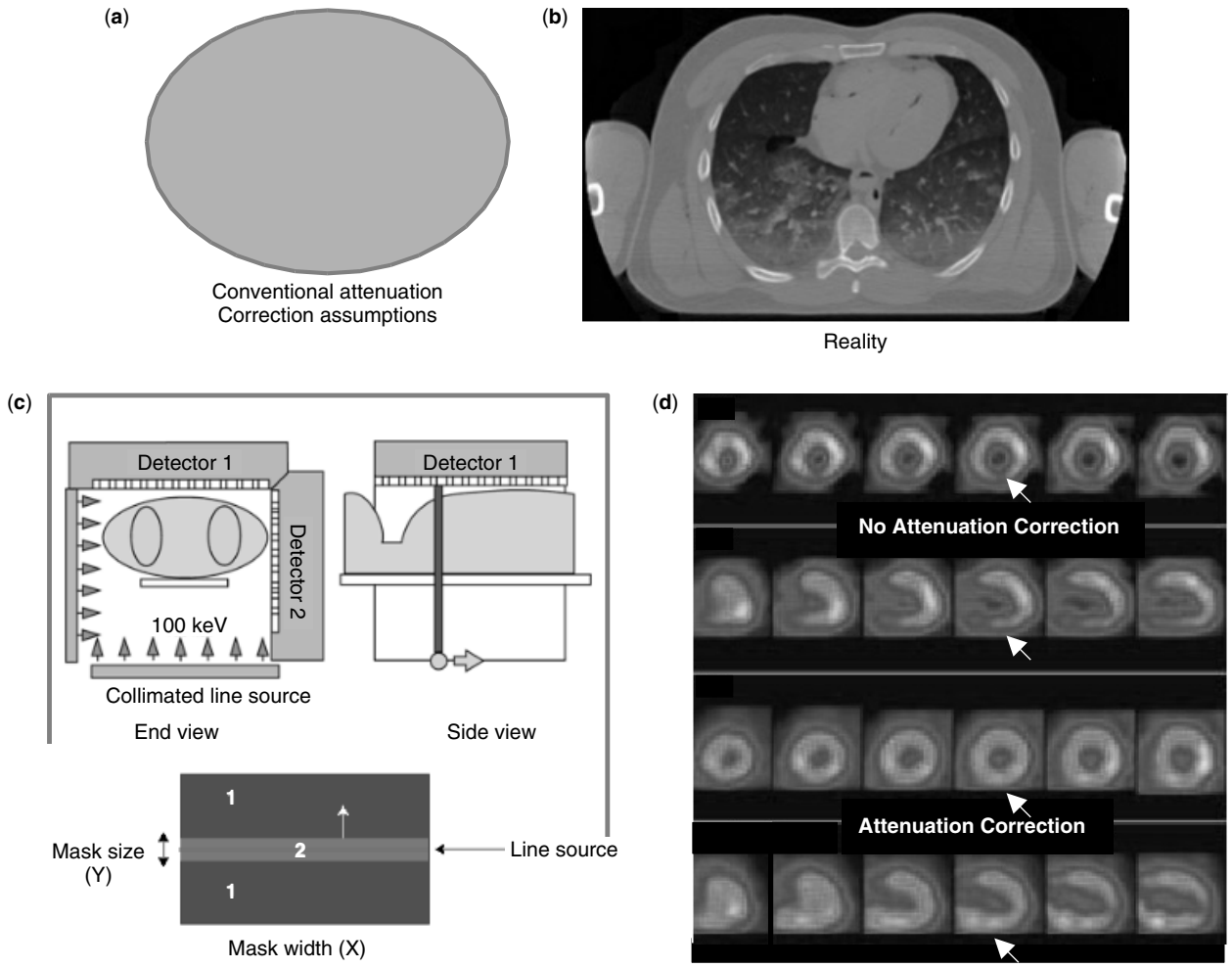


Figure 14. SPECT attenuation correction. Accurate attenuation correction is important for myocardial perfusion imaging because of the heterogeneous distribution of tissues that are different from the assumptions used in the simplified correction schemes (**a**) and (**b**). Accurate attenuation correction requires an independent transmission study using an external gamma ray source such as that shown in (**c**). Attenuation correction removes artifacts that mimic coronary artery disease (**d**). Photos courtesy of GE Medical Systems.

filtered backprojection (30,31). These filters are most often applied to the projection set before reconstruction and have a resolution-restoration component. Then, the resulting projections can be reconstructed by using a ramp filter because the noise is already suppressed. This prefiltering improves image quality significantly, but it still does not accurately correct for attenuation, distance-dependent spatial resolution, or scatter. An iterative approach is required to accomplish that (14,32,33). In the past, iterative algorithms have not been used because of the computational load required to implement the corrections accurately. However, the combination of faster computers and improved reconstruction algorithms in recent years has made these corrections feasible.

Gamma ray attenuation by the body destroys the desired linear relationship between the measured projections and the true line integrals of the internally distributed radioactivity. Reconstructing of the measured projections without compensating for attenuation results in artifacts (34). This is especially a big problem in the thorax where the artifacts from diaphragmatic and breast attenuation mimic the perfusion defects associated with coronary artery disease (Fig. 14d). To correct accurately for attenuation, the attenuation distribution needs to be known for each slice. Many different approaches have been investigated to obtain attenuation maps. These range from using information in the emission data to acquiring transmission studies (35,36). Transmission acquisition is the standard approach used today. In this approach, an external source (or sources) is mounted on the gantry opposite a detector, and transmission measurements are acquired at the same angles as the emission data (Fig. 14c). All of the manufacturers of SPECT systems have options for obtaining transmission studies by using external sources to measure the attenuation distribution of the cardiac patients directly using the scintillation camera as a crude CT. Most of the transmission devices allow simultaneous acquisition of emission and transmission information. Therefore, the transmission sources must have energy emissions different from the radiotracers used in the clinical study. Radionuclides that have been used as sources for transmission studies include Am-241, Gd-153, Ba-133 and Cs-137, and at least one vendor uses an X-ray tube. Different source configurations have been used to collect the transmission studies (37), but all of the commercial systems use one or more line sources. In some systems, the line source is translated across the camera field of view (38). One vendor uses an array of line sources that spans the field of view (39). The information collected from these transmission measurements is corrected for cross talk by using the emission gamma rays, and the transmission views are reconstructed to yield an attenuation map. If the photon energy of the transmission source is significantly different from that of the emission radionuclide, the map has to be scaled to the appropriate attenuation coefficients. This is a relatively easy mapping step and can be done with sufficient accuracy. Then, the scaled attenuation map can be used in the iterative algorithm. Because of time, count rate and sensitivity constraints, the quality of the attenuation maps is poor. However, there is generally sufficient information to provide useful attenuation compensation. It is desirable to minimize noise in the attenuation map, but because the correction is integral, the propagation of noise is considerably less than in multiplicative corrections such as in PET studies.

Compton scattered radiation accounts for about 30–40% of the acquired counts in SPECT imaging. This occurs despite the energy discrimination available in all SPECT systems. This is illustrated in Fig. 15a which shows a plot of the energy of a Compton scattered photon as a function of the scattering angle for different energy gamma rays. Future SPECT systems may have substantially better energy resolution than the 9–10% that is available from NaI(Tl) detectors, but for now, it is necessary to correct for this undesirable information. Scattered

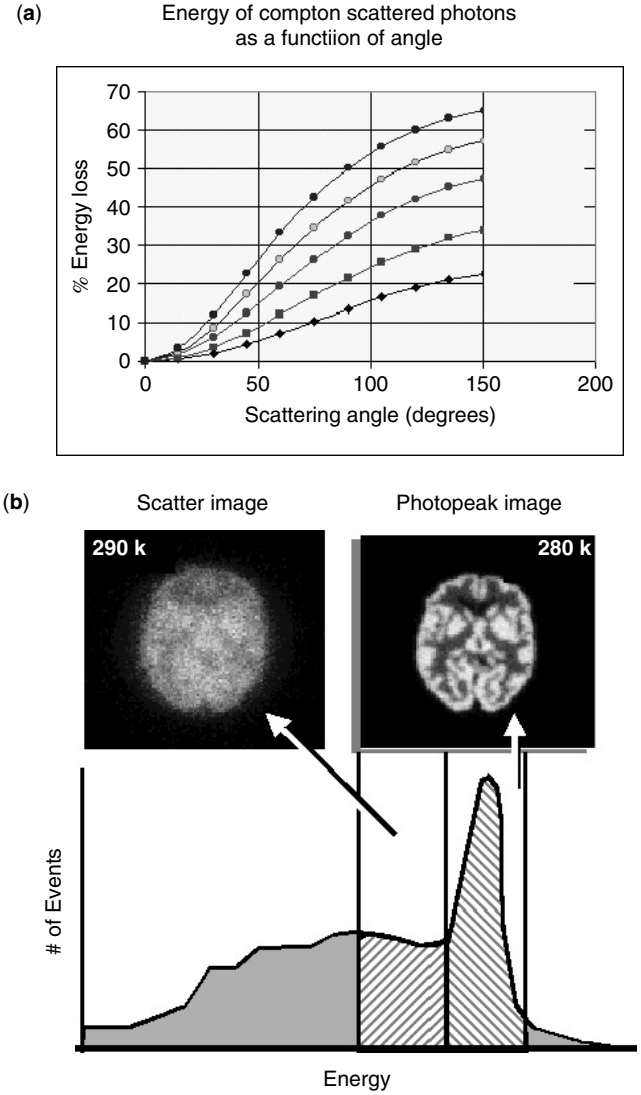


Figure 15. Scattered radiation. Compton scattered radiation degrades contrast and compromises attenuation correction. (**a**) shows a plot of the energy loss of scattered photons as a function of angle and energy. (**b**) Scattered radiation can be compensated for by acquiring additional data simultaneously from an energy window below the photopeak.

radiation decreases contrast and can impact other corrections. For example, when attenuation correction is applied without also correcting for scattered radiation, the heart walls near the liver may be overenhanced.

Scatter has been corrected in several different ways (33,40–44). The easiest to implement is the subtraction method where information is simultaneously acquired in a second energy window centered below the photopeak in the Compton scatter region of the energy spectrum (Fig. 15b). After establishing an appropriate normalization factor, the counts from the scatter window are subtracted from the photopeak window. Then, the corrected projections are used in the reconstruction algorithm. The disadvantage of this approach is that it increases noise and it is difficult to establish an accurate normalization factor. An alternate approach is to model the scatter as part of the forward projection routine using a more or less complicated photon-transport model (30,31,42,45,46). This requires information about tissue densities, which is available from transmission measurements. This approach has the potential to provide the best correction; however, it is computationally intensive. It is likely that this problem will be overcome by improvements in computer performance and by more efficient algorithms. For example, it has already been shown that one can calculate the scatter fractions on a coarser grid than the emission data because it has inherently low resolution (47). In addition, because of its low spatial frequency, it converges rapidly, and it is not necessary to update the calculated scatter data at every iteration.

Correction must also be made for the distance-dependent spatial resolution discussed earlier (Fig. 6). Although a number of approaches have been investigated for applying this correction by using filtered backprojection, the best results have been achieved from iterative algorithms. Like scatter correction, accurate modeling of spatial resolution requires a three-dimensional approach. This is computationally intensive because specific convolution kernels are required for each projection distance. The Gaussian diffusion method is a simpler and faster alternative (48). Using Gaussian diffusion, a convolution kernel is chosen that is sequentially applied at each row of the forward projection matrix. The repeated convolutions reproduce the distance dependence of the collimation fairly accurately.

SPECT/PET HYBRID SYSTEMS

As previously stated, the motivating force behind SPECT imaging is the availability of radiopharmaceuticals that provide crucial diagnostic information. In recent years, it has become apparent that the premier cancer imaging agent is ^{18}F Fluorodeoxyglucose (^{18}F FDG). Fluorodeoxyglucose is a glucose analog that reveals metabolic activity, and it has a sensitivity and specificity for detecting a large number of cancers, including lung, colon, breast, melanoma, and lymphoma. However, ^{18}F is a positron emitter. This makes it ideally suitable for positron-emission tomography, but unfortunately much less suited for SPECT. The main reason for this is the high energy of the annihilation radiation resulting from the positron emission.

High-energy photons are a problem in SPECT for two reasons. First, the relatively thin NaI(Tl) crystals have low efficiency for detection. At 511 keV, the photopeak efficiency is less than 10% for a 9.6-mm crystal. The second problem is that it is difficult to collimate these high-energy photons (49–51). Because thicker septa are required, the count sensitivity is very low. As a result, the spatial resolution is 30–50% worse than collimators used with ^{99m}Tc. This poor spatial resolution reduces the sensitivity of the test. There is one SPECT application where ^{18}F FDG performs adequately, and that is in heart imaging. Fluorodeoxyglucose provides information about the metabolic activity of the heart muscle and is a good indicator of myocardial viability. However, the imaging of ^{18}F FDG in tumors is substantially worse in SPECT than in PET tomographs.

As stated previously, the dual detector SPECT system is the most common configuration. One obvious solution to the problem of collimated SPECT using positron emitting radiotracers is to resort to coincidence imaging (Table 6 and Fig. 16). When a positron is emitted from a nucleus during radioactive decay, it dissipates its energy over a short distance and captures an electron. The electron positron pair very quickly annihilates each other and produces two collinear 511-keV photons. This feature of annihilation radiation can be exploited in coincidence detection where simultaneous detection by opposed detectors is required. Two opposed scintillation cameras whose collimators are removed can have additional electronics added to enable coincidence detection and essentially turn a SPECT system into a PET tomograph (52,53). Although this may sound easy, there are many problems to overcome.

Detection of annihilation photons at the two scintillation cameras represents independent events. The overall efficiency for detection is equal to the product of the individual efficiencies. With a singles efficiency of about 10% (i.e., the efficiency for either detector to register one event), the coincidence efficiency drops to 1%. Although this is very low compared to the detection efficiency at 140 keV (86%), the overall coincidence efficiency is actually very high compared to system efficiency using collimators. But there are still problems. The detection efficiency for detecting only one photon (singles efficiency) is an order of magnitude higher than the coincidence efficiency. This leads to problems with random coincidences. Random coincidences are registered coincidence events that do not result from a

Table 6. SPECT/PET Hybrid System Performance

Parameter	Specification
Number of scintillation cameras	2
NaI(Tl) thickness(mm)	15.9–19
Matrix size	128 × 128
Pixel size	3 mm
Maximum singles rate(counts/s)	1,000,000–2,000,0000
Maximum coincidence rate(counts/s)	10,000
Spatial resolution (mm)	5

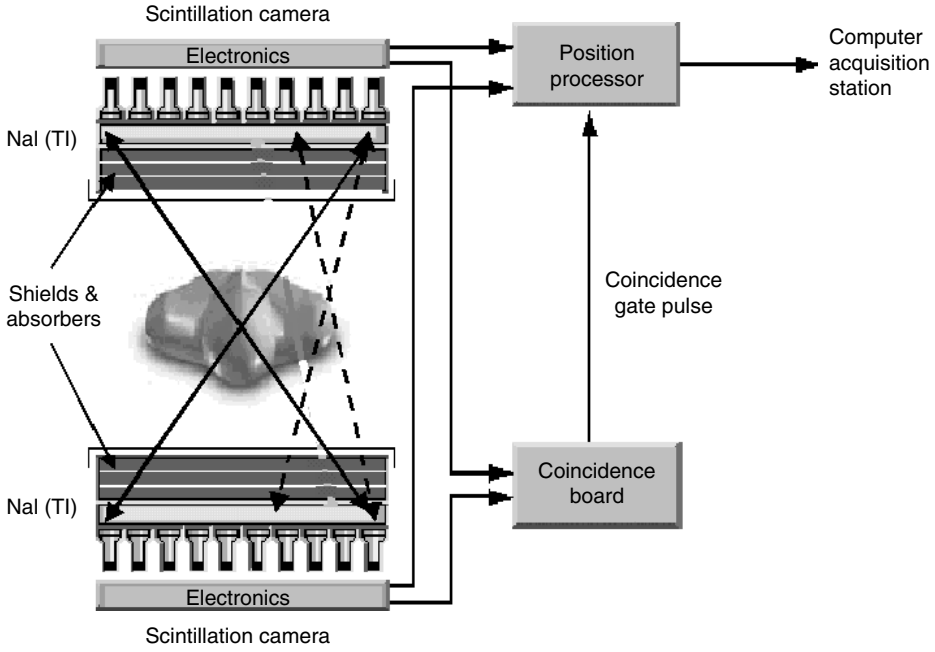


Figure 16. Coincidence imaging. Opposed scintillation cameras can acquire PET studies by adding of coincidence electronics to record the locations of simultaneously detected events. Photos courtesy of GE Medical Systems.

single annihilation. These randoms present a background that needs to be subtracted. Because of the low coincidence efficiency, the random rates in coincidence scintillation cameras are quite high and further compromise image quality. One way to improve this problem is to increase efficiency by using thicker crystals. All of the vendors have done this. The crystals have been increased from 9.6 mm to at least 15.9 mm and to as much as 19 mm. As NaI(Tl) crystals increase in thickness, there is a loss of intrinsic spatial resolution that limits the thickness of the crystals that can be used.

In most SPECT imaging studies, there are essentially no count rate losses from finite temporal resolution. The amount of radioactivity that can be safely administered and the low sensitivity of the collimation put the observed count rate well within the no-loss range. In fact, most SPECT studies would benefit from higher gamma ray flux. This is not so in coincidence imaging. Once the collimators are removed, the wide open NaI(Tl) crystals are exposed to very large count rates. These count rates are so high that count rate losses are unavoidable, and they become the limiting factor in performance. Much effort has been devoted to improving this situation. In the early 1990s, the maximum observed counting rate for a scintillation camera was in the 200,000–400,000 count/second range. As the SPECT systems were redesigned for coincidence imaging, this rate has been extended to more than 1,000,000 counts/second by shortening the integration time on the pulses and implementing active baseline restoration. The limiting count rate factor in the scintillation camera is the persistence of the scintillation. The $1/e$ scintillation time for NaI(Tl) is 250 nanoseconds. At the lower energies at which the scintillation camera typically operates, it is advantageous to capture the entire scintillation signal to optimize energy resolution and intrinsic spatial resolution. In coincidence imaging, shortening the pulse is mandatory. Fortunately, the increased signal obtained from a 511-keV interaction (compared to the typical 140-keV) allows shortening the pulse integration without extreme degradation. Note that, even with these efforts, the coincidence cameras are still limited by count rate and the amount of activity that can be in the camera field of view at the time of imaging is restricted to less than 3 mCi. Thus, one cannot make up for the loss in sensitivity by giving more activity to the patient.

Other measures have been taken to help reduce the count rate burden. One of these is a graded absorber. The scintillation camera has to process every event that the crystal absorbs. The only events we care about are the 511-keV photons, but many low-energy photons that result from scatter within the subject also interact with the detector. The photoelectric cross section is inversely proportional to the cube of the gamma ray energy. This means that a thin lead shield placed over the detector will freely pass most 511-keV photons but will strongly absorb low-energy photons. If one uses only lead, there is a problem with the lead characteristic X rays that are emitted as part of the absorption process. These can be absorbed by a tin filter. In turn, the tin characteristic X rays are absorbed by a copper filter and the copper characteristic X rays by an aluminum filter.

Even though the detectors are thin, the uncollimated detectors present a large solid angle to the annihilation photons. To achieve maximum sensitivity, it is desirable to accept all coincidences, even those at large angles. Several problems are associated with this. First, it is apparent that the camera sensitivity is highly dependent on position. Sources at the central axis of the detectors have a large solid angle, whereas those at the edge have a very small solid angle. In addition, including of the large angle coincidence events drastically increases the scatter component to more than 50%. Because of this problem, many manufacturer's use lead slits aligned perpendicular

to the axial direction to restrict the angular extent of the coincidences. This reduces the scatter component to less than 30% and also reduces the solid angle variability.

The intrinsic spatial resolution of hybrid systems is comparable to that of the dedicated PET systems whose FWHM is 4–5 mm. However, the count sensitivity is at least an order of magnitude lower. This, along with the maximum count rate constraint, guarantees that the coincidence camera data will be very count poor and therefore require substantial low-pass filtering when reconstructed. As a result, the quality of the reconstructed images is perceptibly worse than the dedicated PET images (Fig. 17). In head-to-head comparisons, it has been found that the hybrid systems perform well on tumors greater than 2 cm in diameter located in the lung (54–56). Tumors smaller than 1.5 cm and those located in high background areas are detected by a much lower sensitivity. These results are important because they provide a guide for the useful application of the coincidence camera.

Other improvements have also been made on scintillation camera performance for coincidence imaging. As discussed before, the conventional method for determining the location of an interaction on the detector is through a weighted average of PMT signals. At high count rates, this produces unacceptable positioning errors. For example, if two gamma rays are absorbed simultaneously in the opposing corners of the detector, the Anger logic will place a single event in the center of the camera. New positioning algorithms have been developed that use maximum likelihood calculations and can correctly handle that scenario. The projection information collected by coincidence cameras requires correction for random coincidences, scatter, and attenuation for accurate tomographic images. Typically, randoms are either monitored in a separate time window or are calculated from the singles count rate and are subtracted. Scatter correction is sometimes ignored or is accomplished by subtracting data collected in a separate scatter energy window, as discussed for SPECT imaging. Attenuation correction requires information about the transmission of gamma rays through the body at the coincidence lines of response. Some systems ignore this correction and just reconstruct the random- and scattered-corrected projections. This creates rather severe artifacts but still shows the accumulation of ^{18}F FDG in tumor sites (Fig. 17). When attenuation is corrected, a separate transmission study is performed using an external source, and attenuation maps are formed in a manner similar to that discussed in myocardial SPECT. Cs-137 has been used for this purpose for coincidence cameras. Attenuation correction factors for coincidence imaging are very large for annihilation radiation going through thick portions of the body where they can approach a value of 100.

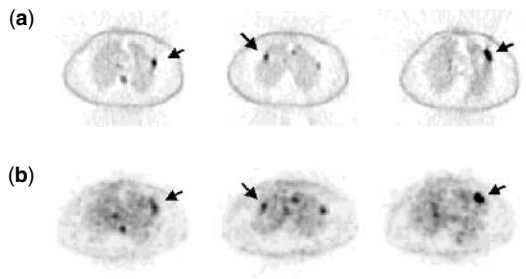


Figure 17. Comparison of PET images from a SPECT/PET hybrid system and a dedicated PET system. (**a**) The top row of images shows several slices obtained by a dedicated PET system from a patient who has lung cancer. (**b**) The bottom row shows the corresponding images obtained from a SPECT/PET hybrid system. The intensified body edge seen in both the dedicated PET and SPECT/PET hybrid system is an attenuation artifact.

SUMMARY

SPECT imaging is expected to play a continuing and important role in medical imaging. Future improvements in SPECT instrumentation are likely to include new detectors and collimation schemes. The coincidence scintillation cameras will also continue their evolution by adding more cameras and multidetector levels optimized for SPECT and coincidence imaging. Improvements in reconstruction algorithms will include prior information about the tomographic image such as smoothness constraints and anatomic distributions. The primary motivating factor in SPECT imaging will continue to be the creation and implementation of new radiopharmaceuticals. SPECT will continue in wide use for myocardial perfusion imaging, but SPECT use in tumor imaging will probably experience the largest growth. Applications will include treatment planning for internal radiation therapy as well as diagnostic studies (57).

ABBREVIATIONS AND ACRONYMS

FWHM	full-width-at-half-maximum
MRI	magnetic resonance imaging
ML-EM	maimum likelihood expectation maximization
OS-EM	ordered subsets expectation maximization
PMT	photomultiplier tube
PET	positron emission tomography
SPECT	single photon emission computed tomography
CT	X-ray computed tomography

BIBLIOGRAPHY

1. J. A. Patton, *Radiographics* **18**, 995–1,007 (1998).
2. T. R. Miller, *Radiographics* **16**, 661–668 (1996).
3. R. Hustinx and A. Alavi, *Neuroimaging Clin. N. Am.* **9**, 751–766 (1999).
4. B. L. Holman and S. S. Tumeh, *JAMA* **263**, 561–564 (1990).
5. R. E. Coleman, R. A. Blinder, and R. J. Jaszczak, *Invest. Radiol.* **21**, 1–11 (1986).
6. J. F. Eary, *Lancet* **354**, 853–857 (1999).
7. D. J. Simpkin, *Radiographics* **19**, 155–167; quiz 153–154 (1999).
8. R. J. Jaszczak and R. E. Coleman, *Invest. Radiol.* **20**, 897–910 (1985).
9. F. H. Fahey, *Radiographics* **16**, 409–420 (1996).
10. S. C. Moore, K. Kouris, and I. Cullum, *Eur. J. Nucl. Med.* **19**, 138–150 (1992).
11. B. M. Tsui and G. T. Gullberg, *Phys. Med. Biol.* **35**, 81–93 (1990).

12. J. M. Links, *Eur. J. Nucl. Med.* **25**, 1,453–1,466 (1998).
13. W. L. Rogers and R. J. Ackermann, *Am. J. Physiol. Imaging* **7**, 105–120 (1992).
14. B. M. Tsui et al., *J. Nucl. Cardiol.* **5**, 507–522 (1998).
15. K. J. LaCroix, B. M. Tsui, and B. H. Hasegawa, *J. Nucl. Med.* **39**, 562–574 (1998).
16. M. P. White, A. Mann, and M. A. Saari, *J. Nucl. Cardiol.* **5**, 523–526 (1998).
17. D. S. Berman and G. Germano, *J. Nucl. Cardiol.* **4**, S169–171 (1997).
18. P. Bruyant, J. Sau, J. J. Mallet, and A. Bonmartin, *Comput. Biol. Med.* **28**, 27–45 (1998).
19. H. Barrett, in A. Todd-Pokropek and M. Viergever, eds., *Medical Images: Formation, Handling and Evaluation*, Springer-Verlag, NY, 1992, pp. 3–42.
20. M. T. Madsen, *Radiographics* **15**, 975–91 (1995).
21. T. F. Budinger et al., *J. Comput. Assist. Tomogr.* **1**, 131–145 (1977).
22. L. Chang, *IEEE Trans. Nucl. Sci.* **NS-25**, 638–643 (1978).
23. J. W. Wallis and T. R. Miller, *J. Nucl. Med.* **34**, 1,793–1,800 (1993).
24. B. F. Hutton, H. M. Hudson, and F. J. Beekman, *Eur. J. Nucl. Med.* **24**, 797–808 (1997).
25. L. Shepp and Y. Vardi, *IEEE Trans. Medical Imaging* **1**, 113–122 (1982).
26. H. Hudson and R. Larkin, *IEEE Trans. Medical Imaging* **13**, 601–609 (1994).
27. L. S. Graham, *Radiographics* **15**, 1,471–1,481 (1995).
28. H. Hines et al., *Eur. J. Nucl. Med.* **26**, 527–532 (1999).
29. T. Budinger et al., *Mathematics and Physics of Emerging Biomedical Imaging*, National Academy Press, Washington, D.C., 1996.
30. J. M. Links et al., *J. Nucl. Med.* **31**, 1,230–1,236 (1990).
31. M. A. King, M. Coleman, B. C. Penney, and S. J. Glick, *Med. Phys.* **18**, 184–189 (1991).
32. M. A. King et al., *J. Nucl. Cardiol.* **3**, 55–64 (1996).
33. B. M. Tsui, X. Zhao, E. C. Frey, and W. H. McCartney, *Semin. Nucl. Med.* **24**, 38–65 (1994).
34. S. L. Bacharach and I. Buvat, *J. Nucl. Cardiol.* **2**, 246–255 (1995).
35. M. T. Madsen et al., *J. Nucl. Cardiol.* **4**, 477–486 (1997).
36. A. Welch, R. Clack, F. Natterer, and G. T. Gullberg, *IEEE Trans. Medical Imaging* **16**, 532–541 (1997).
37. M. A. King, B. M. Tsui, and T. S. Pan, *J. Nucl. Cardiol.* **2**, 513–524 (1995).
38. R. Jaszczak et al., *J. Nucl. Med.* **34**, 1,577–1,586 (1993).
39. A. Celler et al., *J. Nucl. Med.* **39**, 2,183–2,189 (1998).
40. I. Buvat et al., *J. Nucl. Med.* **36**, 1,476–1,488 (1995).
41. F. J. Beekman, C. Kamphuis, and E. C. Frey, *Phys. Med. Biol.* **42**, 1,619–1,632 (1997).
42. F. J. Beekman, H. W. de Jong, and E. T. Slijpen, *Phys. Med. Biol.* **44**, N183–192 (1999).
43. D. R. Haynor, M. S. Kaplan, R. S. Miyaoka, and T. K. Lewellen, *Medical Phys.* **22**, 2,015–2,024 (1995).
44. M. S. Rosenthal et al., *J. Nucl. Med.* **36**, 1,489–1,513 (1995).
45. A. Welch et al., *Medical Phys.* **22**, 1,627–1,635 (1995).
46. A. Welch and G. T. Gullberg, *IEEE Trans. Medical Imaging* **16**, 717–726 (1997).
47. D. J. Kadrmas, E. C. Frey, S. S. Karimi, and B. M. Tsui, *Phys. Med. Biol.* **43**, 857–873 (1998).
48. V. Kohli, M. A. King, S. J. Glick, and T. S. Pan, *Phys. Med. Biol.* **43**, 1,025–1,037 (1998).
49. W. E. Drane et al., *Radiology* **191**, 461–465 (1994).
50. J. S. Fleming and A. S. Alaamer, *J. Nucl. Med.* **37**, 1,832–1,836 (1996).
51. P. K. Leichner et al., *J. Nucl. Med.* **36**, 1,472–1,475 (1995).
52. T. K. Lewellen, R. S. Miyaoka, and W. L. Swan, *Nucl. Med. Commun.* **20**, 5–12 (1999).
53. P. H. Jarritt and P. D. Acton, *Nucl. Med. Commun.* **17**, 758–766 (1996).
54. P. D. Shreve et al., *Radiology* **207**, 431–437 (1998).
55. P. D. Shreve, R. S. Steventon, and M. D. Gross, *Clin. Nucl. Med.* **23**, 799–802 (1998).
56. R. E. Coleman, C. M. Laymon, and T. G. Turkington, *Radiology* **210**, 823–828 (1999).
57. P. B. Zanzonico, R. E. Bigler, G. Sgouros, and A. Strauss, *Semin. Nucl. Med.* **19**, 47–61 (1989).
58. L. S. Graham et al., *Medical Phys.* **22**, 401–409 (1995).

STEREO AND 3-D DISPLAY TECHNOLOGIES

DAVID F. MCALLISTER
North Carolina State University
Raleigh, NC

INTRODUCTION

Recently, there have been rapid advancements in 3-D techniques and technologies. Hardware has improved and become considerably cheaper, making real-time and interactive 3-D available to the hobbyist, as well as to the researcher. There have been major studies in areas such as molecular modeling, photogrammetry, flight simulation, CAD, visualization of multidimensional data, medical imaging, teleoperations such as remote vehicle piloting and remote surgery, and stereolithography. In computer graphics, the improvements in speed, resolution, and economy make interactive stereo an important capability. Old techniques have been improved, and new ones have been developed. True 3-D is rapidly becoming an important part of computer graphics, visualization, virtual-reality systems, and computer gaming. Numerous 3-D systems are granted patents each year, but very few systems move beyond the prototype stage and become commercially viable. Here, we treat the salient 3-D systems.

First, we discuss the major depth cues that we use to determine depth relationships among objects in a scene.

DEPTH CUES

The human visual system uses many depth cues to eliminate the ambiguity of the relative positions of objects in a 3-D scene. These cues are divided into two categories: physiological and psychological.

Physiological Depth Cues

Accommodation. *Accommodation* is the change in focal length of the lens of the eye as it focuses on specific regions of a 3-D scene. The lens changes thickness due to a change in tension of the ciliary muscle. This depth cue is normally used by the visual system in tandem with convergence.

Convergence. *Convergence*, or simply *vergence*, is the inward rotation of the eyes to converge on objects as they move closer to the observer.

Binocular Disparity. *Binocular disparity* is the difference in the images projected on the retinas of the left and right eyes in viewing a 3-D scene. It is the salient depth cue used by the visual system to produce the sensation of depth, or *stereopsis*. Any 3-D display device must be able to produce a left- and right-eye view and present them to the appropriate eye separately. There are many ways to do this as we will see.

Motion Parallax. *Motion parallax* provides different views of a scene in response to movement of the scene or the viewer. Consider a cloud of discrete points in space in which all points are the same color and approximately the same size. Because no other depth cues (other than binocular disparity) can be used to determine the relative depths of the points, we move our heads from side to side to get several different views of the scene (called *look around*). We determine relative depths by noticing how much two points move relative to each other: as we move our heads from left to right or up and down, the points closer to us appear to move more than points further away.

Psychological Depth Cues

Linear Perspective. *Linear perspective* is the change in image size of an object on the retina in inverse proportion to the object's change in distance. Parallel lines moving away from the viewer, like the rails of a train track, converge to a vanishing point. As an object moves further away, its image becomes smaller, an effect called *perspective foreshortening*. This is a component of the depth cue of retinal image size.

Shading and Shadowing. The amount of light from a light source that illuminates a surface is inversely proportional to the square of the distance from the light source to the surface. Hence, the surfaces of an object that are further from the light source are darker (shading), which gives cues of both depth and shape. Shadows cast by one object on another (shadowing) also give cues to relative position and size.

Aerial Perspective. Distant objects tend to be less distinct and appear cloudy or hazy. Blue has a shorter wavelength and penetrates the atmosphere more easily than other colors. Hence, distant outdoor objects sometimes appear bluish.

Interposition. If one object occludes, hides, or overlaps (interposes) another, we assume that the object doing the hiding is closer. This is one of the most powerful depth cues.

Retinal Image Size. We use our knowledge of the world, linear perspective, and the relative sizes of objects to determine relative depth. If we view a picture in which an elephant is the same size as a human, we assume that the elephant is further away because we know that elephants are larger than humans.

Textural Gradient. We can perceive detail more easily in objects that are closer to us. As objects become more distant, the textures become blurred. Texture in brick, stone, or sand, for example, is coarse in the foreground and grows finer as distance increases.

Color. The fluids in the eye refract different wavelengths at different angles. Hence, objects of the same shape and size and at the same distance from the viewer often appear to be at different depths because of differences in color. In addition, bright-colored objects will appear closer than dark-colored objects.

The human visual system uses all of these depth cues to determine relative depths in a scene. In general, depth cues are additive; the more cues, the better the viewer can determine depth. However, in certain situations, some cues are more powerful than others, and this can produce conflicting depth information. Our interpretation of the scene and our perception of the depth relationships result from our knowledge of the world and can override binocular disparity.

A TECHNOLOGY TAXONOMY

The history of 3-D displays is well summarized in several works. Okoshi (1) and McAllister (2) each present histories of the development of 3-D technologies. Those interested in a history beginning with Euclid will find (3) of interest.

Most 3-D displays fit into one or more of three broad categories: stereo pair, holographic, and multiplanar or volumetric. Stereo pair-based technologies distribute left and right views of a scene independently to the left and right eyes of the viewer. Often, special viewing devices are required to direct the appropriate view to the correct eye and block the incorrect view to the opposite eye. If no special viewing devices are required, then the technology is called *autostereoscopic*. The human visual system processes the images and if the pair of images is a stereo pair, described later, most viewers will perceive depth. Only one view of a scene is possible per image pair which means that the viewer cannot change position and see a different view of the scene. We call such images "virtual." Some displays include head tracking devices to simulate head motion or "look around." Some technologies allow displaying multiple views of the same scene providing motion parallax as the viewer moves the head from side to side. We discuss these technologies here.

In general, holographic and multiplanar images produce "real" or "solid" images, in which binocular disparity, accommodation, and convergence are consistent with the

apparent depth in the image. They require no special viewing devices and hence, are autostereoscopic. Holographic techniques are discussed elsewhere. Multiplanar methods are discussed later.

Stereo Pairs

The production of stereoscopic photographs (stereo pairs or *stereographs*) began in the early 1850s. Stereo pairs simulate the binocular disparity depth cue by projecting distinct (normally flat) images to each eye. There are many techniques for viewing stereo pairs. One of the first was the stereoscope, which used stereo card images that can still be found in antique shops. A familiar display technology, which is a newer version of the stereoscope, is the ViewMaster and its associated circular reels (Fig. 1).

Because some of the displays described are based on the stereo pair concept, some stereo terminology is appropriate.

Terminology. Stereo pairs are based on presenting two different images, one for the left eye (L) and the other for the right eye (R). Stereo images produced photographically normally use two cameras that are aligned horizontally and have identical optics, focus, and zoom. To quantify what the observer sees on the two images, we relate each image to a single view of the scene.

Consider a point P in a scene being viewed by a binocular viewer through a window (such as the film plane of a camera). A point P in the scene is projected on the window surface, normally a plane perpendicular to the observer's line of sight, such as the camera film plane, the face of a CRT, or a projection screen. This projection surface is called the *stereo window* or *stereo plane*. We assume that the y axis lies in a plane that is perpendicular to the line through the observer's eyes. The distance between the eyes is called the *interocular distance*. Assigning a Cartesian coordinate system to the plane, the point P will appear on the left eye view at coordinates (x_L, y_L) and in the right eye view at coordinates (x_R, y_R). These two points are called *homologous*. The horizontal parallax of the point P is the distance $x_R - x_L$ between the left- and right-eye views; the vertical parallax is $y_R - y_L$ (Fig. 2).

Positive parallax occurs if the point appears behind the stereo window because the left-eye view is to the left of the right-eye view. Zero parallax occurs if the point is at the same depth as the stereo window; zero parallax defines the stereo window, and negative parallax occurs if the point lies in front of the stereo window (Fig. 3).

Given the previous geometric assumptions, vertical parallax or *vertical disparity* should always be zero. Misaligned cameras can produce nonzero vertical parallax. Observers differ about the amount they can tolerate before getting side effects such as headache, eye strain, nausea,

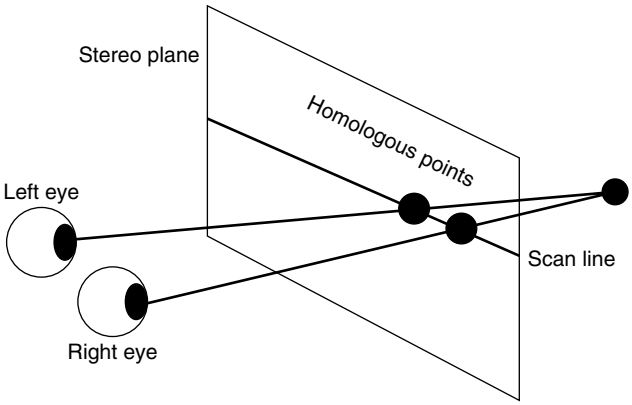


Figure 2. Horizontal parallax.

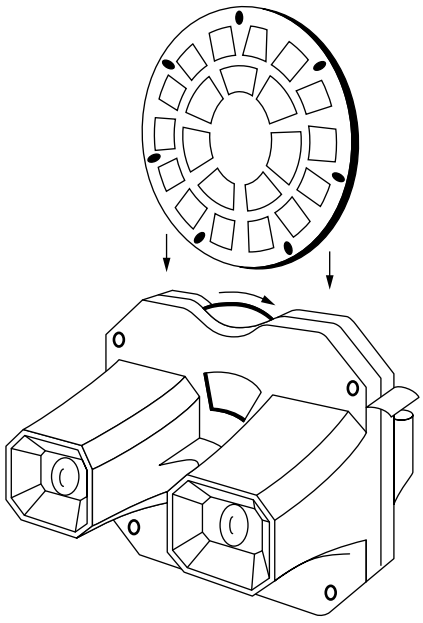


Figure 1. ViewMaster.

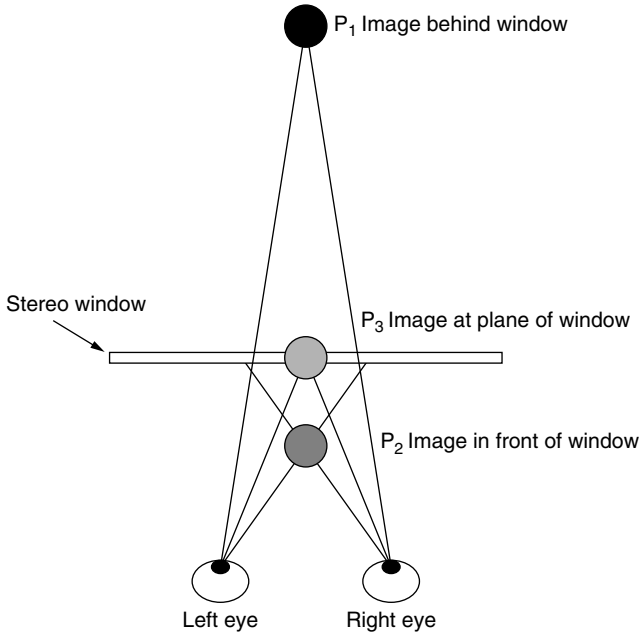


Figure 3. Positive/negative parallax.

or other uncomfortable physical symptoms. Henceforth, the term *parallax* will mean horizontal parallax.

If the horizontal parallax is too large and exceeds the maximum parallax, to view the points, our eyes must go *wall-eyed*, a condition where the eyes each move to the outside to view the image. After lengthy exposure, this can produce disturbing physical side effects. Images in which the parallax is reversed are said to have *pseudostereo*. Such images can be very difficult to fuse; the human visual system will have difficulty recognizing the binocular disparity. Other depth cues compete and overwhelm the visual system.

Parallax and convergence are the primary vehicles for determining perceived depth in a stereo pair; the observer focuses both eyes on the plane of the stereo window. Hence, accommodation is fixed. In such cases accommodation and convergence are said to be "disconnected," and the image is "virtual" rather than "solid" (see the section on volumetric images later). This inconsistency between accommodation and convergence can make stereo images difficult for some viewers to fuse. If you cannot perceive depth in a stereo pair, you may be a person who is "stereo-blind" and cannot fuse stereo images (interpret as a 3-D image rather than two separate 2-D images). There are many degrees of stereo-blindness, and the ability or inability to see stereo may depend on the presentation technique, whether the scene is animated, color consistency between the L/R pair, and many other considerations.

Computation of Stereo Pairs. Several methods have been proposed for computing stereo pairs in a graphics environment. Certain perception issues eliminate some techniques from consideration.

A common technique for computing stereo pairs involves rotating a 3-D scene about an axis parallel to the sides of the viewing screen, followed by a perspective projection. This process can cause vertical displacement because of the foreshortening that occurs in a perspective projection. Hence, the technique is not recommended.

Although parallel projection will not produce vertical displacement, the absence of linear perspective can create a "reverse" perspective as the result of a perceptual phenomenon known as Emmert's law: objects that do not obey linear perspective can appear to get larger as the distance from the observer increases. The preferred method for computing stereo pairs is to use two off-axis centers of perspective projection (corresponding to the positions of the left and right eyes). This method simulates the optics of a stereo camera where both lenses are parallel. For further details, see (2).

OVERVIEW OF DISPLAY TECHNOLOGIES

Separating Left- and Right-Eye Views

When viewing stereo pairs, a mechanism is required so that the left eye sees only the left-eye view and the right eye sees only the right-eye view. Many mechanisms have been proposed to accomplish this. The ViewMaster uses two images each directed to the appropriate eye by lenses. The images are shown in parallel, and there is no way one eye can see any part of the other eye view.

It is common in display technologies to use a single screen to reflect or display both images either simultaneously (time parallel) or in sequence (time multiplexed or field sequential). The technologies used to direct the appropriate image to each eye while avoiding mixing the left- and right-eye images require sophisticated electro-optics or shuttering. Some of the more common methods are described here.

Cross Talk

*S*tereo *cross talk* occurs when a portion of one eye view is visible in the other eye. In this case, the image can appear blurred, or a second or double image appears in regions of the scene being viewed that creates a phenomenon called *ghosting*. Cross talk can create difficulty in fusing L/R views. When using the same display surface to project both eye views, cross talk can be a problem. When stereo displays are evaluated, the cross talk issue should be addressed.

Field-Sequential Techniques

A popular method for viewing stereo by a single display device is the field-sequential or time-multiplexed technique. The L/R views are alternated on the display device, and a blocking mechanism to prevent the left eye from seeing the right eye view and vice versa is required. The technology for field-sequential presentation has progressed rapidly. Historically, mechanical devices were used to occlude the appropriate eye view during display refresh. A comparison of many of these older devices can be found in (4). Newer technologies use electro-optical methods such as liquid-crystal plates. These techniques fall into two groups: those that use active versus passive viewing glasses.

In a passive system, a polarizing shutter is attached to the display device, as in a CRT, or the screen produces polarized light automatically as in an LCD panel. The system polarizes the left- and right-eye images in orthogonal directions (linear or circular), and the user wears passive polarized glasses where the polarization axes are also orthogonal. The polarizing lenses of the glasses combine with the polarized light from the display device to act as blocking shutters to each eye. When the left eye view is displayed, the light is polarized along an axis parallel to the axis of the left-eye lens and the left eye sees the image on the display. Because the axis is orthogonal to the polarizer of the right eye, the image is blocked to the right eye.

The passive system permits several people to view the display simultaneously and allows a user to switch viewing easily from one display device to another because no synchronization with the display device is required. It also permits a larger field of view (FOV). The drawback is that the display device must produce a polarized image. Projector mechanisms must have polarizing lenses, and a CRT or panel display must have a polarizing plate attached to or hanging in front of the screen or the projector. When projecting an image on a screen, the screen must be coated with a material (vapor-deposited aluminum) that does not depolarize the light (the commercially available "silver"

screen). Polarization has the added disadvantage that the efficiency or *transmission* is poor; the intensity of the light to reach the viewer compared to the light emitted from the display device is very low, often in the range of 30%. Hence, images appear dark.

LCDs can also be used as blocking lenses. An electronic pulse provided by batteries or a cable causes the lens to "open" or admit light from the display device. When no electronic pulse is present, the lens is opaque and blocks the eye from seeing the display device. The pulses are alternated for each eye while the display device alternates the image produced. The glasses must be synchronized to the refresh of the display device, normally using an infrared signal or a cable connection. For CRT-based systems, this communication is accomplished using the stereo-sync or Left/Right (L/R) signal. In 1997, the Video Equipment Standards Association (VESA) called for the addition of a standard jack that incorporates the L/R signal along with a + 5 volt power supply output. Using this new standard, stereo equipment can be plugged directly into a stereo-ready video card that has this jack.

Active glasses have an advantage that the display device does not have to polarize the light before it reaches the viewer. Hence, efficiency is higher and back-projection can be used effectively. The disadvantage is obviously the requirement for synchronization.

Though the initial cost of the passive system is higher, the cost to add another user is low. This makes the passive system a good choice for theaters and trade shows, for example, where one does not want to expose expensive eyewear to abuse.

If the images in both systems are delivered at a sufficiently fast frame rate (120 Hz) to avoid flicker, the visual system will fuse the images into a three-dimensional image. Most mid- to high-end monitors can do this. A minimum of 100 Hz is acceptable for active eyewear systems. One may be able to use 90 Hz for a passive system without perceiving flicker, even in a well-lit room.

Time-Parallel Techniques

Time-parallel methods present both eye views to the viewer simultaneously and use optical techniques to direct each view to the appropriate eye.

Often, 3-D movies used the *anaglyph* method that requires the user to wear glasses that have red and green lenses or filters. Both images were presented on a screen simultaneously; hence, it is a time-parallel method. Many observers suffered headaches and nausea when leaving the theater, which gave 3-D, and stereo in particular, a bad reputation. (A phenomenon called ghosting or cross talk was a significant problem. Colors were not adjusted correctly, and the filters did not completely eliminate the opposite-eye view, so that the left eye saw its image and sometimes part of the right-eye image as well. Other problems included poor registration of the left- and right-eye images that caused vertical parallax and projectors out of sync.) The ViewMaster is another example of a time-parallel method.

An early technique for viewing stereo images on a CRT was the half-silvered mirror originally made for viewing microfiche (4). The device had polarizing sheets, and the user wore polarized glasses that distributed the correct view to each eye.

Polarizing filters can also be attached to glass-mounted slides. Incorrect positioning of the projectors relative to the screen can cause *keystoning*, in which the image is trapezoidal caused by foreshortening that results in vertical parallax.

If more than one projector is used, as is often the case when projecting 35-mm stereo slides, for example, orthogonal polarizing filters are placed in front of each projector, and both left- and right-eye images are projected simultaneously onto a nondepolarizing screen. Hence, the technique is time parallel. The audience wears passive glasses in this case. Using more than one projector always brings with it the difficulties of adjusting the images. L/R views should be correctly *registered*; there must be minimal luminosity differences, minimal size differences, minimal keystoning, minimal vertical parallax, minimal ghosting, and so forth.

Most nonautostereoscopic display systems use one of these methods. Following, we indicate which method.

3D DISPLAYS VIEWING DEVICES REQUIRED

Hard Copy

Anaglyphs. The anaglyph method has been used for years to represent stereo pairs, and it was a salient technique in old 3-D movies and comic books. Colored filters cover each eye; red/green, red/blue, or red/cyan filters are the most common. One eye image is displayed in red and the other in green, blue, or cyan, so that the appropriate eye sees the correct image. Because both images appear simultaneously, it is a time-parallel method. The technique is easy to produce using simple image processing techniques, and the cost of viewing glasses is very low. Gray-scale images are most common. Pseudocolor or *polychromatic* anaglyphs are becoming more common. If correctly done, anaglyphs can be an effective method for presenting stereo images.

Vectographs. Polaroid's Vectograph process was introduced by Edwin Land in 1940. The earliest Vectograph images used extensively were black-and-white polarizing images formed by iodine ink applied imagewise to oppositely oriented polyvinyl alcohol (PVA) layers laminated to opposite sides of a transparent base material. The iodine forms short polymeric chains that readily align with the oriented polymeric molecules and stain the sheet. The chemistry is analogous to that of uniformly stained iodine polarizers, such as Polaroid H-sheet, used in polarizing filters for stereo projection and in 3-D glasses used for viewing stereoscopic images [see (2) for more details].

In 1953, Land demonstrated three-color Vectograph images formed by successive transfer of cyan, magenta, and yellow dichroic dyes from gelatin relief images to Vectograph sheet. Unlike StereoJet digital ink-jet printing described next, preparation of Vectograph color images required lengthy, critical photographic and dye transfer steps. Although the process produced excellent images, it was never commercialized.

StereoJet. The StereoJet process, developed at the Rowland Institute for Science in Cambridge, Massachusetts, provides stereoscopic hard copy in the form of integral, full-color polarizing images. StereoJet images are produced by ink-jet printing that forms polarizing images by using inks formulated from dichroic dyes. Paired left-eye and right-eye images are printed onto opposite surfaces of a clear multilayer substrate, as shown in Fig. 4.

The two outer layers, formed of an ink-permeable polymer such as carboxymethylcellulose, meter the ink as it penetrates the underlying image-receiving layers. The image-receiving layers are formed of polyvinyl alcohol (PVA) molecularly oriented at 45° to the edge of the sheet. As the dye molecules are adsorbed, they align with the oriented polymer molecules and assume the same orientation. The two PVA layers are oriented at 90° to one another, so that the images formed have orthogonal polarization.

StereoJet transparencies are displayed directly by rear illumination or projected by overhead projector onto a nondepolarizing screen, such as a commercially available lenticular "silver" screen. No attachments to the projector are needed because the images themselves provide the polarization. StereoJet prints for viewing by reflected light have aluminized backing laminated to the rear surfaces of StereoJet transparencies.

ChromaDepth. Chromostereoscopy is a phenomenon in optics commercialized by Richard Steenblik (2). The technique originally used double prism-based glasses that slightly deflect different colors in an image, laterally displacing the visual positions of differently colored regions of an image by different amounts. The prisms are oriented in opposite directions for each eye, so that different images are presented to each eye, thereby creating a stereo pair (Fig. 5). Production chromostereoscopic glasses, marketed under the name

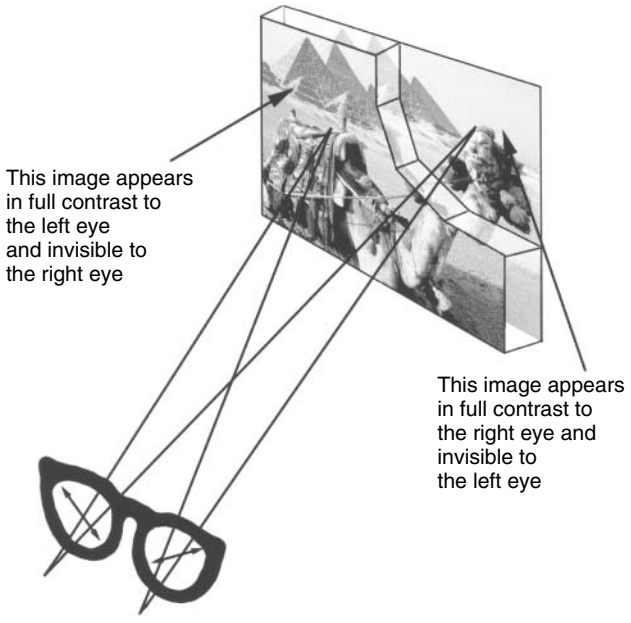


Figure 4. StereoJet imaging.

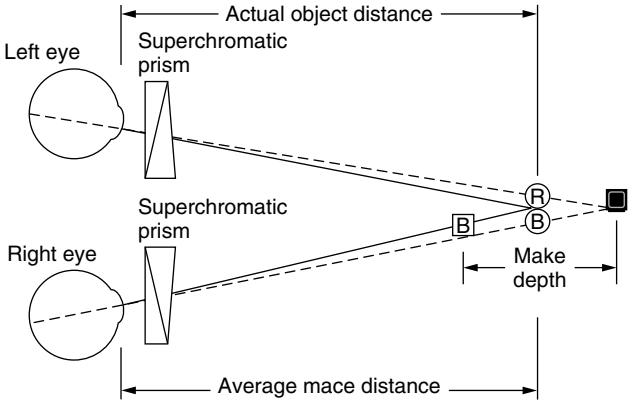


Figure 5. Superchromatic glasses.

ChromaDepth 3-D, use a unique micro-optic film that performs the same optical function as double-prism optics without the attendant weight and cost. Images designed for viewing with ChromaDepth 3-D glasses use color to encode depth information. A number of color palettes have been successfully employed; the simplest is the RGB on Black palette: on a black background, red will appear closest, green in the middleground, and blue in the background. Reversal of the optics results in the opposite depth palette: BGR on Black.

A peculiar feature of the ChromaDepth 3-D process is that the user does not have to create a stereo pair. A single ChromaDepth 3-D color image contains X, Y, and Z information by virtue of the image contrast and the image colors. The stereo pair seen by the user is created by the passive optics in the ChromaDepth 3-D glasses. The primary limitation of the ChromaDepth 3-D process is that the colors in an image cannot be arbitrary if they are to carry the image's Z dimension; so the method will not work on arbitrary images. The best effects are obtained from images that are specifically designed for the process and from natural images, such as underwater reef photographs, that have natural coloring fitting the required palette.

Another limitation is that some color "fringing" can occur when viewing CRT images. The light emitted from a CRT consists of different intensities of red, green, and blue; any other color created by a CRT is a composite of two or more of these primary colors. If a small region of a composite color, such as yellow, is displayed on a CRT, the optics of the ChromaDepth 3-D glasses may cause the composite color to separate into its primary components and blur the region. ChromaDepth 3-D high definition glasses reduce this problem by placing most of the optical power in one eye, leaving the other eye to see the image clearly.

The ChromaDepth 3-D technique can be used in any color medium. It has found wide application in laser shows and in print, video, television, computer graphic, photographic slide, and Internet images. Many areas of research have benefited from ChromaDepth 3-D, including interactive visualization of geographic and geophysical data.

Transparency Viewers. Cheap plastic and cardboard slide viewers are available from many companies like Reel 3-D Enterprises *(http://stereoscopy.com/reel3-D/index.html)* for viewing 35-mm stereo slides. The user places the left eye view in the left slot and the right eye view in the right slot and then holds them up to the light. This is a standard technique for checking the mounting of slides for correct registration.

Field-Sequential Devices

StereoGraphics Systems. Although there are many manufacturers of active and passive glasses systems, StereoGraphics is a well-known company that has produced high-quality CRT and RGB projector based stereo systems for years. The quality of their hardware is excellent, and we report on it here.

Active StereoGraphics shutters called CrystalEyes (Fig. 6) are doped, twisted-nematic devices. They "open" in about 3 ms and "close" in about 0.2 ms. The shutter transition occurs within the vertical blanking period of the display device and is all but invisible. The principal figure of merit for such shutters is the dynamic range, which is the ratio of the transmission of the shutter in its open state to its closed state. The CrystalEyes system has a ratio in excess of 1000 : 1. The transmission of the shutters is commonly 32%, but because of the 50% duty cycle, the effective transmission is half that. Their transmission should be neutral and impart little color shift to the image being viewed. The field of view (FOV) also varies. Ninety-seven degrees is typical. SGI can operate at a speed up to 200 fields per second. The cost for eyewear and emitter is $1000.

Passive systems have a lower dynamic range than active eyewear systems. The phosphor afterglow on the CRT causes ghosting, or image cross talk, in this type of system. Electrode segmentation can be used to minimize the time during which the modulator is passing an unwanted image. The modulator's segments change state moments before the CRT's scanning beam arrives at that portion of the screen. The consequence of this action is a modulator that changes state just as the information is changing. This increases the effective dynamic range of the system and produces a high-quality stereo image.



Figure 6. Active glasses CrystalEyes system.



Figure 7. Passive glasses ZScreen system.

This technique is used by StereoGraphics in their ZScreen system (Fig. 7). A Monitor ZScreen system costs $2200.

The above-and-below format is used on personal computers that do not have a stereo sync output. The left image is placed on the top half of the CRT screen, and the right image on the bottom half, thus reducing the resolution of the image. Chasm Graphics makes a software program called Sudden Depth that will format the images this way. Now, the stereo information exists but needs an appropriate way to send each L/R image to the proper eye. The StereoGraphics EPC-2 performs this task.

The EPC-2 connects to the computer's VGA connector and intercepts the vertical sync signal. When enabled, the unit adds an extra vertical sync pulse halfway between the existing pulses. The result causes the monitor to refresh at twice the original rate. In effect, this stretches the two images to fill the whole screen and show field-sequential stereo. The EPC-2 acts as an emitter for CrystalEyes or can be used as a device to create a left/right signal to drive a liquid crystal modulator or other stereo product. The EPC-2 is the same size as the other emitters and has approximately the same range. Its cost is $400.

The Pulfrich Technique. Retinal sensors require a minimum number of light photons to fire and send a signal to the visual system. By covering one eye with a neutral density filter (like a lens in a pair of sunglasses), the light from a source will be slightly delayed to the covered eye. Hence, if an object is in motion in a scene, the eye that has the filter cover sees the position of the object later than the uncovered eye. Therefore, the images perceived by the left and right eyes will be slightly different, and the visual system will interpret the result as a stereo pair.

If the motion of an object on a display device is right to left and the right eye is covered by the filter, then a point on the object will be seen by the left eye before the right eye. This will be interpreted by the visual system as positive parallax, and the object will appear to move behind the stereo window. Similarly, an object moving from left to right will appear in front of the display device. The reader can implement the technique easily using one lens of a pair of sunglasses while watching TV.

The Fakespace PUSH Display. Fakespace Lab's PUSH desktop display uses a box-shaped binocular viewing device that has attached handles and is mounted on a triad of cylindrical sensors (Fig. 8). The device allows the user to move the viewing device and simulate limited movement within a virtual environment. The field of view can be as large as 140° on CRT-based systems. The cost is US $25,000 for the 1024 × 768 CRT and US $9,995 for the 640 × 480 LCD version.

A variation that permits more viewer movement is the Boom (Fig. 9). The binocular viewing device is attached to a large arm configured like a 3-D digitizer that signals the position of the viewer using sensors at the joints of the arm. The viewer motion is extended to a circle 6 ft in diameter. Vertical movement is limited to 2.5 ft. The Boom sells for US $60,000. A hands-free version is available for US $85,000.

Workbench Displays. Smaller adjustable table-based systems such as the Fakespace ImmersaDesk R2 (Fig. 10) and ImmersaDesk M1 are available. The systems use the active glasses stereo technique. The fully portable R2 sells for approximately US $140,000, including tracking. The M2 sells for US $62,995.



Figure 8. The Fakespace Lab's PUSH desktop display.

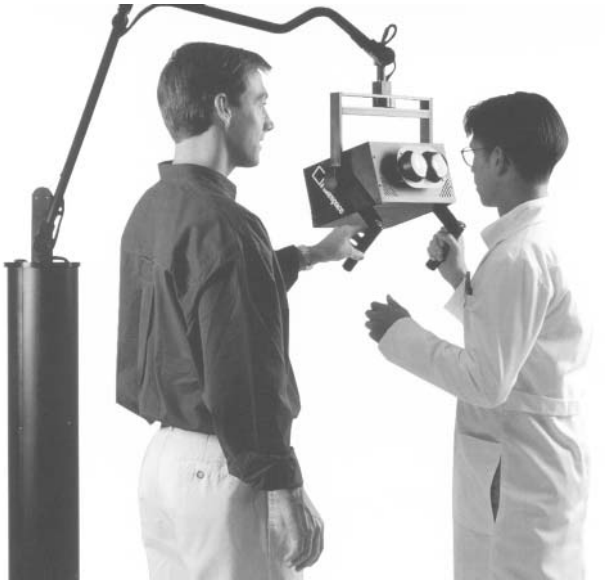


Figure 9. Fakespace Lab's Boom.



Figure 10. Fakespace ImmersaDesk R2.

VREX Micropolarizers. VREX has patented what they call the μPol (micropolarizer) technology, an optical device that can change the polarization of an LCD display line by line. It is a periodic array of microscopically small polarizers that spatially alternate between mutually perpendicular polarizing states. Each micropolarizer can be as small as 10 millionths of a meter. Hence, a μPol could have more than 6 million micropolarizers of alternating polarization states per square inch in a checkerboard configuration of more than 2500 lines per inch in a one-dimensional configuration. In practice, the μPol encodes the left-eye image on even lines and the right-eye image on odd lines. Passive polarized glasses are needed to view the image.

The format requires a single-frame stereoscopic image format that combines a left-eye perspective view with a right-eye perspective view to form a composite image, which contains both left- and right-eye information alternating line by line. VREX provides software to combine left- and right-eye views into a single image. All VREX hardware supports this image format. The advantages of μPol include the ability to run at lower refresh rates because both eyes are presented with a (lower resolution) image simultaneously and hence the presentation is time parallel.

LARGE FORMAT DISPLAYS

One of the objects of virtual reality is to give the user the feeling of immersion in a scene. This has been accomplished in various ways. Head-mounted displays are a common solution. In general, head-mounted displays have a limited field of view and low resolution. In addition, allowing the user to move in space requires position tracking which has been a difficult problem to solve. Position tracking results in image lag which is a result of the time required to sense that the viewer's position has changed, signal the change to the graphics system, render the scene change, and then transmit it to the head-mounted display. Any system that must track the viewer and change the scene accordingly must treat this problem. The lag can produce motion sickness in some people.

Projection systems have been developed that use large projection surfaces to simulate immersion. In some cases, the user is permitted to move about. In others, the user is stationary, and the scene changes.

IMAX

Most readers are familiar with the large screen IMAX system that employs a large flat screen to give the illusion of peripheral vision. When projecting stereo, IMAX uses the standard field-sequential polarized projection mechanism where the user wears passive glasses. Similar techniques are used in the Kodak flat screen 3-D movies at Disney.

Fakespace Systems Displays

Fakespace Systems markets immersive displays that are similar to immersive technologies produced by several other companies.

The walk-in, fully immersive CAVE is an extension of flat screen stereo. The CAVE system was developed at the Electronic Visualization Lab of the University of Illinois where the user is in a 10×10 ft room that has flat walls (Figs. 11 and 12). A separate stereo image is back-projected onto each wall; the floor and possibly the ceiling give the user the feeling of immersion. Image management is required so that the scenes on each wall fit together seamlessly to replicate the single surrounding environment. Because the system uses back-projection, it requires active shuttering glasses. The user can interact with the environment using 3-D input devices such as gloves and other navigational tools. The system sells for approximately US $325,000 to $500,000, depending on the projection systems used.

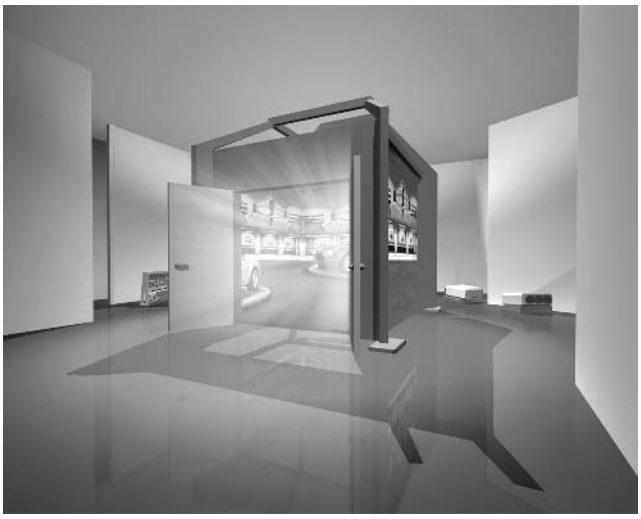


Figure 11. Fakespace CAVE, front view.



Figure 12. Fakespace CAVE, inside.

Fakespace also produces an immersive WorkWall whose screen size is up to 8×24 ft (Fig. 13). The system uses two or more projectors, and images blend to create a seamless image. As in the CAVE, the user can interact with the image using various 2-D and 3-D input devices. The cost is approximately US $290,000 for an 8×24 ft three-projector system.

The VisionDome

Elumens Corporation Vision Series displays (5–8) use a hemispherical projection screen that has a single projection lens. Previous dome-based systems relied on multiple projectors and seamed-together output from multiple computers, making them both complicated to configure and prohibitively expensive. The high cost, complexity, and nonportability of these systems made them suitable for highly specialized military and training applications, but they were impractical and out of reach for most corporate users. Available in sizes from 1.5 to 5 meters in diameter, which accommodate from one to forty

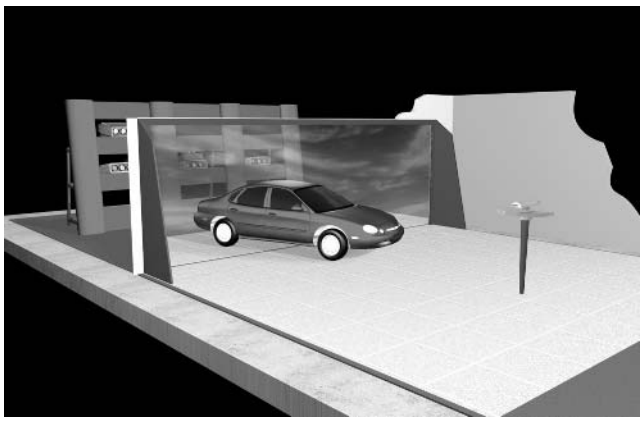


Figure 13. The Fakespace WorkWall.

people, the VisionDome systems range in price from US $15,000 to US $300,000.

The projector is equipped with a patented "fish-eye" lens that provides a 180° field of view. This single projection source completely fills the concave screen with light. Unlike other fish-eye lenses, whose projections produce focal "hot spots" and nonlinear distortions, the Vision Series lens uses linear angular projection to provide uniform pixel distribution and uniform pixel size across the entire viewing area. The lens also provides an infinite depth of field, so images remain in focus on screens from 0.5 meters away to theoretical infinity at all points on the projection surface.

The single-user VisionStation displays 1024×768 pixels at 1000 lumens; larger 3- to 5-meter VisionDomes display up to 1280×1024 pixels at 2000 lumens (Fig. 14).

Elumens provides an application programming interface called SPI (Spherical Projection of Images). Available for both OpenGL and DirectX applications, SPI is an image-based methodology for displaying 3-D data on a curved surface. It enables off-axis projection that permits arbitrary placement of the projector on the face plane of the screen. The image-based projection depends on the viewer position; if the viewer moves, the image must change accordingly, or straight lines become curved. The number of viewers within the viewing "sweet spot" increases as the screen diameter increases.

Field-sequential stereo imaging with synchronized shutter glasses is supported on Elumens products. The maximum refresh rate currently supported is 85 Hz (42.5 Hz stereo pair). Passive stereo that projects left- and right-eye images simultaneously but of opposite polarization is currently under development.



Figure 14. The VisionDome.

AUTOSTEREOSCOPIC DISPLAYS NO VIEWING DEVICES REQUIRED

Hard Copy

Free Viewing. With practice, most readers can view stereo pairs without the aid of blocking devices by using a technique called *free viewing*. There are two types of free viewing, distinguished by the way the left- and right-eye images are arranged. In *parallel*, or *uncrossed* viewing, the left-eye image is to the left of the right-eye image. In *transverse* or *cross* viewing, they are reversed and crossing the eyes is required to form an image in the center. Some people can do both types of viewing, some only one, some neither. In Fig. 15, the eye views have been arranged in left/right/left order. To parallel view, look at the left two images. To cross view, look at the right two images.

Figure 16 is a *random dot autostereogram* in which the scene is encoded in a single image, as opposed to a stereo pair (9). There are no depth cues other than binocular disparity. Using cross viewing, merge the two dots beneath the image to view the functional surface. Crossing your eyes even further will produce other images. [See (10) for a description of the method for generating these interesting images].

Holographic Stereograms. Most readers are familiar with holographic displays, which reconstruct solid images. Normally, a holographic image of a three-dimensional scene has the "look around" property. A popular combination of holography and stereo pair technology, called a *holographic stereogram*, involves recording a set of 2-D images, often perspective views of a scene, on a piece of holographic film. The film can be bent to form a cylinder, so that the user can walk around the cylinder to view the scene from any aspect. At any point, the left eye will see one view of the scene and the right eye another, or the user is viewing a stereo pair.

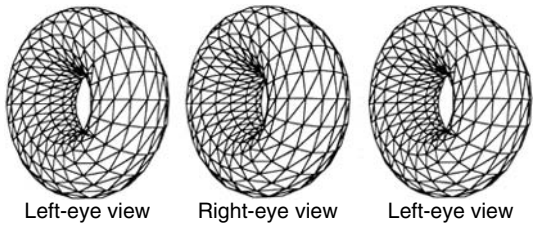


Figure 15. Free viewing examples.

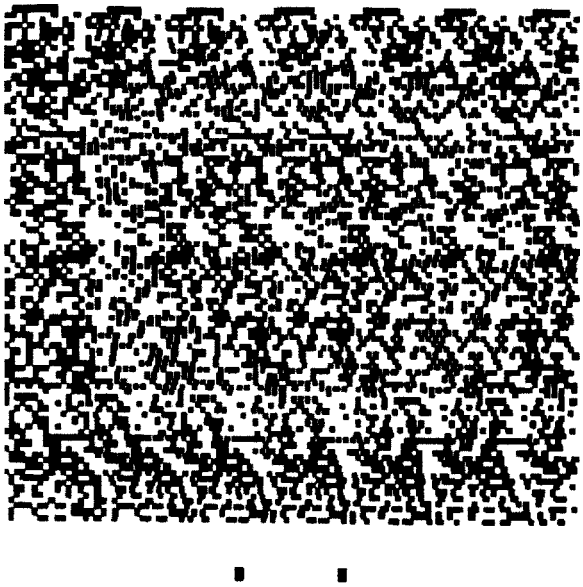


Figure 16. A random dot autostereogram $\cos[(x^2 + y^2)^{(1/2)}]$ for $-10 \leq x, y \leq 10$.

Conventional display holography has long been hampered by many constraints such as limitations with regard to color, view angle, subject matter, and final image size. Despite the proliferation of holographic stereogram techniques in the 1980s, the majority of the constraints remained. Zebra Imaging, Inc. expanded on the developments in one-step holographic stereogram printing techniques and has developed the technology to print digital full-color reflection holographic stereograms that have a very wide view angle (up to 110°), are unlimited in size, and have full parallax.

Zebra Imaging's holographic stereogram technique is based on creating an array of small (1- or 2-mm) square elemental holographic elements (hogels). Much like the pixels of two-dimensional digital images, hogel arrays can be used to form complete images of any size and resolution. Each hogel is a reflection holographic recording on panchromatic photopolymer film. The image recorded in each hogel is of a two-dimensional digital image on a spatial light modulator (SLM) illuminated by laser light in the three primary colors: red, green, and blue (Fig. 17).

Parallax Barrier Displays. A parallax barrier (2) consists of a series of fine vertical slits in an otherwise opaque medium. The barrier is positioned close to an image that has been recorded in vertical slits and backlit. If the vertical slits in the image have been sampled at the correct frequency relative to the slits in the parallax barrier and the viewer is the required distance from the barrier, then the barrier will occlude the appropriate image slits to the right and left eyes, respectively, and the viewer will perceive an autostereoscopic image (Fig. 18). The images can be made panoramic to some extent by recording multiple views of a scene. As the viewer changes position, different views of the scene will be directed by the barrier to the visual system. The number of views is limited by the optics and, hence, moving horizontally beyond a certain point will produce "image flipping" or cycling of the different views of the scene.

High resolution laser printing has made it possible to produce very high quality images: the barrier is printed on one side of a transparent medium and the image on the other. This technique was pioneered by Art^n in the early 1990s to produce hard-copy displays and is now being used by Sanyo for CRT displays.

Lenticular Sheets. A *lenticular sheet* (1,2) consists of a series of semicylindrical vertical lenses called "lenticles," typically made of plastic. The sheet is designed so the parallel light that enters the front of the sheet will be focused onto strips on the flat rear surface (Fig. 19). By recording an image in strips consistent with the optics of the lenticles, as in the parallax barrier display, an autostereoscopic panoramic image can be produced. Because the displays depend on refraction versus occlusion, the brightness of a lenticular sheet display is usually superior to the parallax barrier and requires no backlighting. Such displays have been mass produced for many years for such hard-copy media as postcards.

In these two techniques, the image is recorded in strips behind the parallax barrier or the lenticular sheet. Although the techniques are old, recent advances in printing and optics have increased their popularity for both hard-copy and autostereoscopic CRT devices.

In both the lenticular and parallax barrier cases, multiple views of a scene can be included to provide

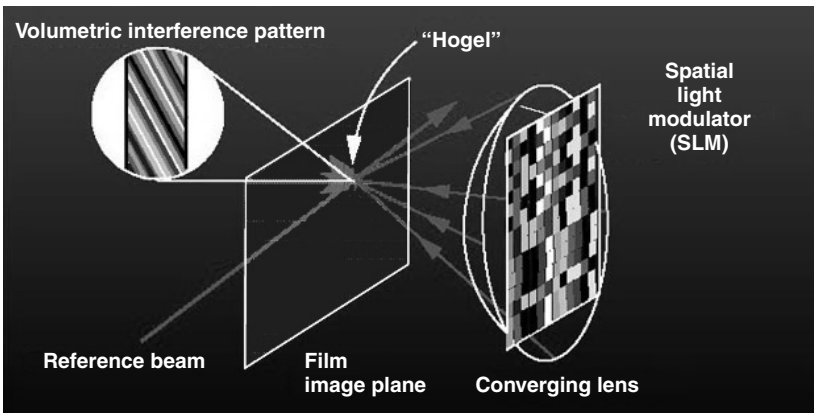


Figure 17. Zebra Imaging holographic stereogram recording.

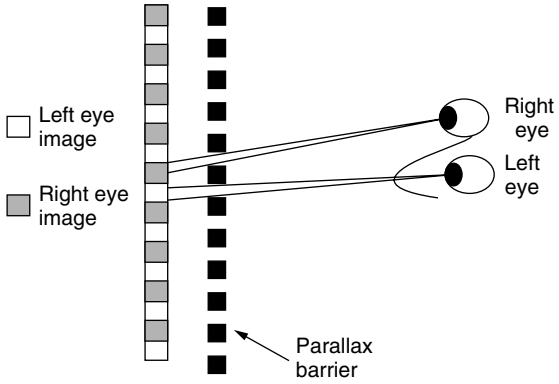


Figure 18. Parallax barrier display.

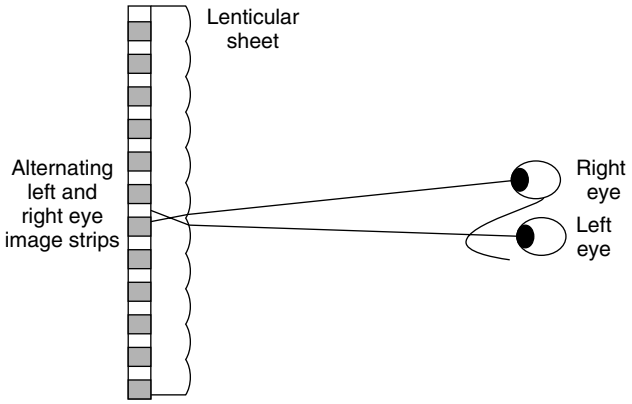


Figure 19. Lenticular sheet display.

motion parallax as viewers move their heads from side to side creating what is called a *panoramagram*. Recently, parallax barrier liquid-crystal imaging devices have been developed that can be driven by a microprocessor and used to view stereo pairs in real time without glasses. Some of these techniques are discussed later.

Alternating Pairs

The output from two vertically mounted video cameras are combined. An integrating circuit was designed to merge the two video streams by recording a fixed number of frames from one camera, followed by the same number of frames from the other camera. The technique imparts a vertical rocking motion to the image. If the scene has sufficient detail and the speed of the rocking motion and the angle of rotation are appropriate for the individual viewing the system, most viewers will fuse a 3-D image. The system was commercialized under the name VISIDEP. The technique can be improved using graphical and image processing methods. More details can be found in (2).

Moving Slit Parallax Barrier

A variation of the parallax barrier is a mechanical moving slit display popularized by Homer Tilton that he called the Parallactiscope (2). A single vertical slit is vibrated horizontally in front of a point-plotting output display such as a CRT or oscilloscope. The image on the display is synchronized with the vibration to produce an autostereoscopic image. Many variants have been proposed but to date the author knows of no commercially viable products using the technique.

The DTI System

The Dimension Technologies, Inc. (DTI) illuminator is used to produce what is known as a multiperspective autostereoscopic display. Such a display produces multiple images of a scene; each is visible from a well-defined region of space called a viewing zone. The images are all 2-D perspective views of the scene as it would appear from the center of the zones. The viewing zones are of such a size and position that an observer sitting in front of the display always has one eye in one zone and the other eye in another. Because the two eyes see different images in different perspectives, a 3-D image is perceived.

The DTI system is designed for use with an LCD or other transmissive display. The LCD is illuminated from behind, and the amount of light passing through individual elements is controlled to form a full-color image.

The DTI system uses an LCD backlight technology which they call parallax illumination (11). Figures 20 and 21 illustrate the basic concept. As shown in Fig. 20, a special illuminator is located behind the LCD.

The illuminator generates a set of very thin, very bright, uniformly spaced vertical lines. The lines are spaced with respect to pixel columns such that (because of parallax) the left eye sees all the lines through the odd columns of the LCD and the right eye sees them through even columns. There is a fixed relation between the distance of the LCD to the illumination plate, and the distance of the viewer from the display. This in part determines the extent of the "viewing zones." As shown in Fig. 21, viewing zones are diamond-shaped areas in front of the display where all of the light lines are seen behind the odd or even pixel columns of the LCD.

To display 3-D images, left- and right-eye images of a stereoscopic pair are placed in alternate columns of elements. The left image appears in the odd columns, and the right image is displayed in even columns. Both left and right images are displayed simultaneously, and hence

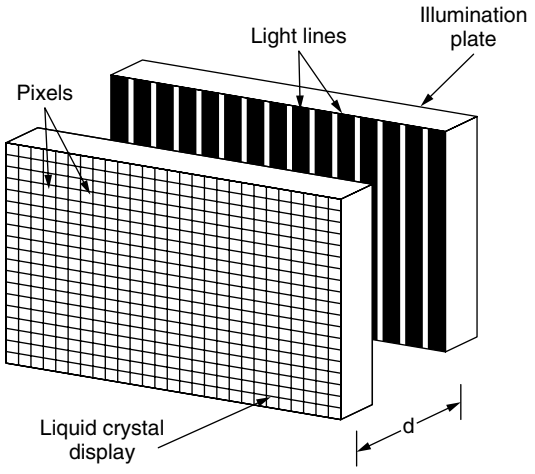


Figure 20. DTI illuminator.

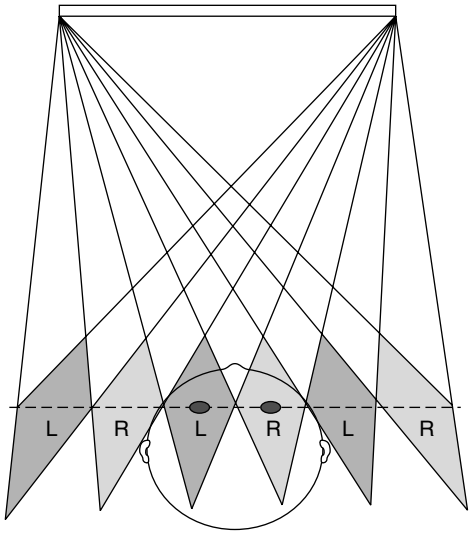


Figure 21. Viewing zones.

the display is time parallel. Because the left eye sees the light lines behind the odd columns, it sees only the left-eye image displayed in the odd columns. Similarly, the right eye sees only the right-eye image displayed in the even columns.

The 2-D/3-D Backlight System. There are many ways to create the precise light lines described before. One method that is used in DTI products is illustrated in Fig. 22 (12,13). The first component is a standard off-the-shelf backlight of the type used for conventional 2-D LCD monitors. This type of backlight uses one or two miniature fluorescent lamps as light sources in combination with a flat, rectangular light guide. Two straight lamps along the top and bottom of the guide are typically used for large displays. A single U-shaped lamp is typically used for smaller displays. An aluminized reflector is placed around the lamp(s) to reflect light into the light guide.

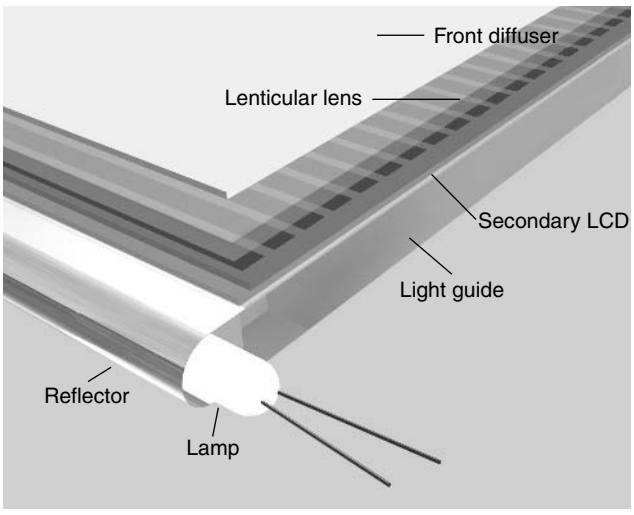


Figure 22. Backlight system.

The flat, rectangular light guide is typically made of acrylic or some other clear plastic. Light from the lamp enters the light guide from the sides and travels through it due to total internal reflection from the front and back surfaces of the guide.

The side of the light guide facing away from the LCD possesses a pattern of reflective structures designed to reflect light into the guide and out the front surface. Several possible choices for such structures exist, but current manufacturers usually use a simple pattern of white ink dots applied to the rear surface of the light guide in combination with a white reflective sheet placed behind the light guide.

The second component is a simple, secondary LCD which, in the "on" state, displays a pattern of dozens of thin, transparent lines that have thicker opaque black stripes between them. These lines are used for 3-D imaging as described in the previous section.

The third major component is a lenticular lens, again shown in Fig. 22 This lens consists of a flat substrate upon whose front surface of hundreds of vertical, parallel cylindrical lenslets are molded. Light coming through the dozens of thin transparent lines on the secondary LCD is reimaged into thousands of very thin, evenly spaced vertical lines by a lenticular lens array spaced apart from and in front of the secondary LCD. The lines can be imaged onto an optional front diffuser located in a plane at one focal length from the lenticular lenslets. The pitch (center-to-center distance) of the lines on the light guide and the lenticular lenses must be chosen so that the pitch of the light lines reimaged by the lenticular lenslets bears a certain relationship to the pitch of the LCD pixels.

Because the displays are likely to be used for conventional 2-D applications (such as word processing and spreadsheets) as well as 3-D graphics, the system must be capable of generating illumination so that each eye sees all of the pixels of the LCD and a conventional full-resolution 2-D image can be displayed by using conventional software.

Note that when the secondary LCD is off, in other words in the clear state where the lines are not generated, the even diffuse light from the backlight passes through it freely and remains even and diffuse after being focused by the lenticular lens. Therefore, when the secondary LCD is off, no light lines are imaged, and the observer sees even, diffuse illumination behind all of the pixels of the LCD. Therefore, each of the observer's eyes can see all of the pixels on the LCD, and full-resolution 2-D images can be viewed.

DTI sells two displays, a 15-inch at $1699 and optional video input at $300 extra, and an 18.1-inch at $6999, video included. Both have 2-D and 3-D modes and accept the standard stereo formats (field sequential, frame sequential, side by side, top/bottom).

Seaphone Display

Figure 23 shows a schematic diagram of the Seaphone display (14–16). A special transparent μPol-based color liquid-crystal imaging plate (LCD : SVGA 800 × 600) that has a lenticular sheet and a special backlight unit is used to produce a perspective image for each eye. The lenticular

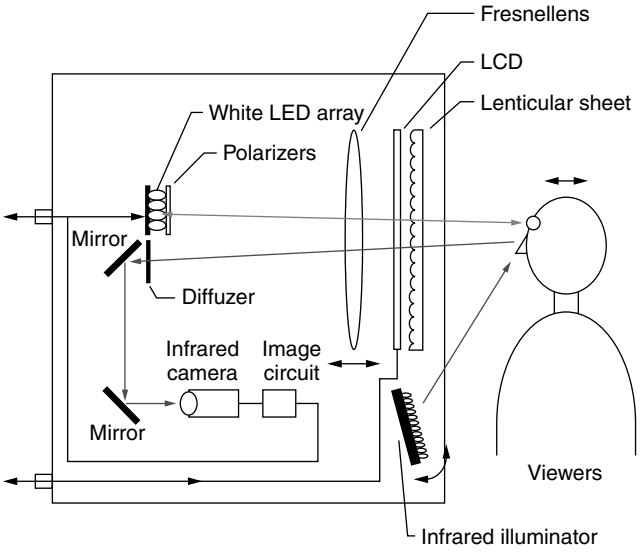


Figure 23. A schematic of the Seaphone display.

sheet creates vertical optical scattering. Horizontal strips of two types of micropolarizers that have orthogonal polarization axes are transmitted on odd versus even lines of the LCD. The backlight unit consists of a large format convex lens and a white LED array filtered by the polarizers whose axes of polarization are the same as those of the μPol array. The large format convex lens is arranged so that an image of the viewers is focused on the white LED array.

The light from the white LED array illuminates the right half face of the viewer using the odd (or even) field of the LCD, when the geometrical condition is as indicated in Fig. 25. The viewers' right eyes perceive the large convex lens as a full-size bright light source, and the viewers' left eyes perceive it as dark one; similarly for the left eye (see Fig. 24).

On the head tracking system, the viewers' infrared image is focused on the diffuser by the large format convex lens and is captured by the infrared camera. An image circuit modulates the infrared image and produces binary half right and left face images of each viewer. The binary half face images are displayed on the appropriate cells of the white LED array. The infrared image is captured by using the large convex format lens. There is no parallax

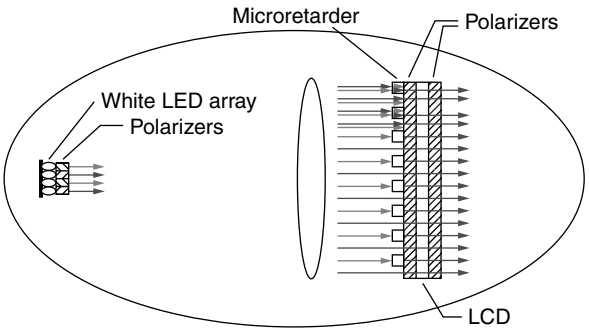


Figure 24. Each perspective backlight.

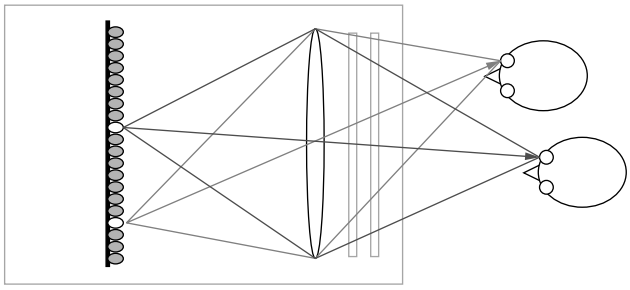


Figure 25. Plan view of a backlight unit.

in the captured infrared image when the image is focused on the white LED array. Hence, the displayed infrared viewers' binary half right face image (the appropriate cells) and the viewers' image that is focused by the large format convex lens are automatically superimposed on the surface of the white LED array. The bright areas of the binary half face images (the appropriate cells) are distributed to the correct eye of the viewers.

On the Seaphone display, several viewers can perceive a stereo pair simultaneously, and they can move independently without special attachments. The display currently costs 1,492,000 yen.

The Sanyo Display

The Sanyo display uses LC technology for both image presentation and a parallax barrier (17). Because the thermal expansion coefficients are the same, registration is maintained under different operating conditions. They call the parallax barrier part of the display the "image splitter." They use two image splitters, one on each side of the LC (image presentation) panel (Fig. 26).

The splitter on the backlight side is two-layer thin films of evaporated aluminum and chromium oxide. The vertical stripes are produced by etching. The stripe pitch is slightly larger than twice the dot pitch of the LC panel. The viewer-side splitter is a low-reflection layer. The stripe pitch is slightly smaller than twice the dot pitch on the LC image presentation panel. Each slit corresponds to a column of the LC panel. They claim that the technique produces no

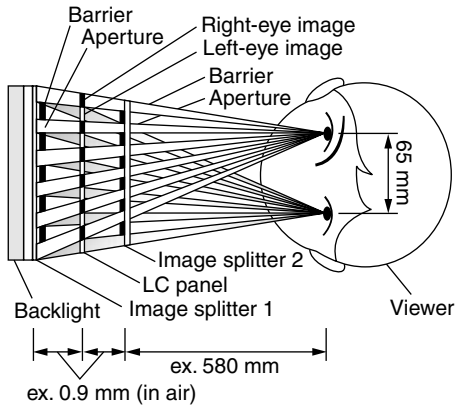


Figure 26. A double image splitter.

ghosting. They also have a head-tracking system in which the viewer does not have to wear any attachments.

The HinesLab Display

An autostereoscopic display using motion parallax (18–20) has been developed by HinesLab, Inc. (www.hineslab.com) of Glendale, California. The display uses live or recorded camera images, or computer graphics, and displays multiple views simultaneously (Fig. 27). The viewer stands or sits in front of the display where the eyes fall naturally into two of multiple viewing positions. If the viewer shifts positions, the eyes move out of the two original viewing positions into two different positions where views that have the appropriate parallax are prepositioned. This gives a natural feeling of motion parallax as the viewer moves laterally. An advantage of this approach is that multiple viewers can use the display simultaneously.

The technology provides from 3 to 21 eye positions that give lateral head freedom and look-around ability, confirming the positions and shapes of objects. The device is NTSC compatible, and all images can be projected on a screen simultaneously in full color without flicker. The display is built around a single liquid-crystal panel, from which multiple images are projected to a screen where they form the 3-D image. The general approach used to create the autostereo display was to divide the overall area of the display source into horizontal rows. The rows were then filled by the maximum number of images, while maintaining the conventional 3 : 4 aspect ratio; no two images have the same lateral position (Fig. 28).

The optical design for these configurations is very straightforward. Identical projection lenses are mounted on a common surface in the display housing, and they project each image to the back of a viewing screen from unique lateral angles. Working in conjunction with a Fresnel field lens at the viewing screen, multiple exit pupils, or viewing positions, are formed at a comfortable viewing distance in front of the display. Figure 29 shows an arrangement of seven images displayed in three horizontal rows on the LCD panel.



Figure 27. HinesLab autostereoscopic computer display — video arcade games.

VOLUMETRIC DISPLAYS

A representation technique used in computer visualization to represent a 3-D object uses parallel planar cross sections of the object, for example, CAT scans in medical imaging. We call such a representation a *multiplanar* image. *Volumetric* or multiplanar 3-D displays normally depend on moving mirrors, rotating LEDs, or other optical techniques to project or reflect light at points in space. Indeed, aquariums full of Jell-O that have images drawn in ink inside the Jell-O have also been used for such displays. A survey of such methods can be found in (2,21).

A few techniques are worth mentioning. First, we discuss the principle of the oscillating mirror.

Oscillating Planar Mirror

Imagine a planar mirror which can vibrate or move back and forth rapidly along a track perpendicular to the face of a CRT, and assume that we can flash a point (pixel) on the CRT that decays very rapidly (Fig. 30).

Let the observer be on the same side of the mirror as the CRT, so that the image in the CRT can be seen reflected by the mirror. If a point is rendered on the surface of the CRT when the mirror reaches a given location in its vibration and the rate of vibration of the mirror is at least fusion frequency (30 Hz), the point will appear continuously in the same position in space. In fact, the point would produce a solid image in the sense that, as we changed our position, our view of the point would also change accordingly. If the point is not extinguished as the mirror vibrates, then the mirror would reflect the point at all positions on its track, and the viewer would see a line in space perpendicular to the face of the CRT. Any point plotted on the surface of the CRT would appear at a depth depending on the position of the mirror at the instant the point appears on the CRT. The space that contains all possible positions of points appearing on the CRT defines what is called the *view volume*.

All depth cues would be consistent, and there would be no "disconnection" of accommodation and vergence as for stereo pairs. The optics of the planar mirror produce a view volume depth twice that of the mirror excursion or

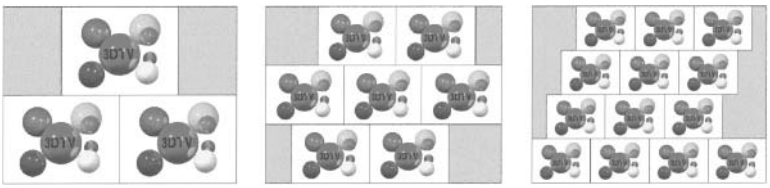


Figure 28. Possible image arrangements on the liquid-crystal projection panel.

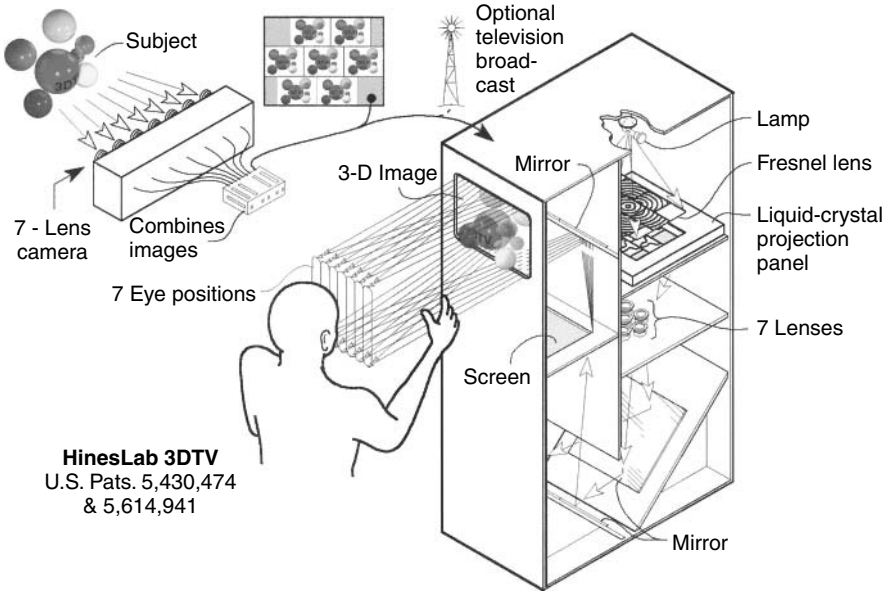


Figure 29. The seven-lens autostereo display.

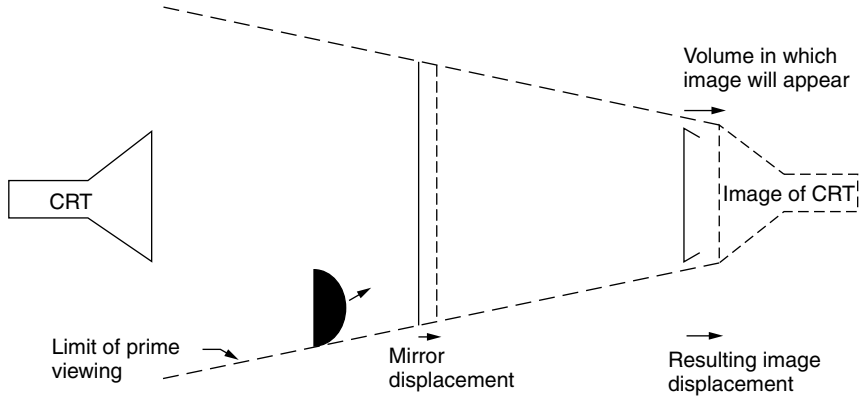


Figure 30. Vibrating mirror.

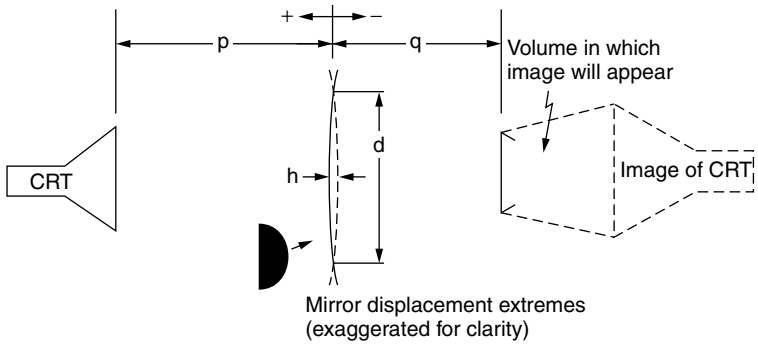


Figure 31. A varifocal mirror.

displacement depth. If the focal length of the mirror is also changed during the oscillation, a dramatic improvement in view volume depth can be obtained.

Varifocal Mirror

The varifocal mirror was a commercially available multiplanar display for several years. The technique uses a flexible circular mirror anchored at the edges (Fig. 31). A common woofer driven at 30 Hz is used to change the focal length of the mirror. A 3-D scene is divided into hundreds of planes, and a point-plotting electrostatic CRT plots a single point from each. The mirror reflects these points, and the change in the focal length of the mirror affects their apparent distance from the viewer. A software

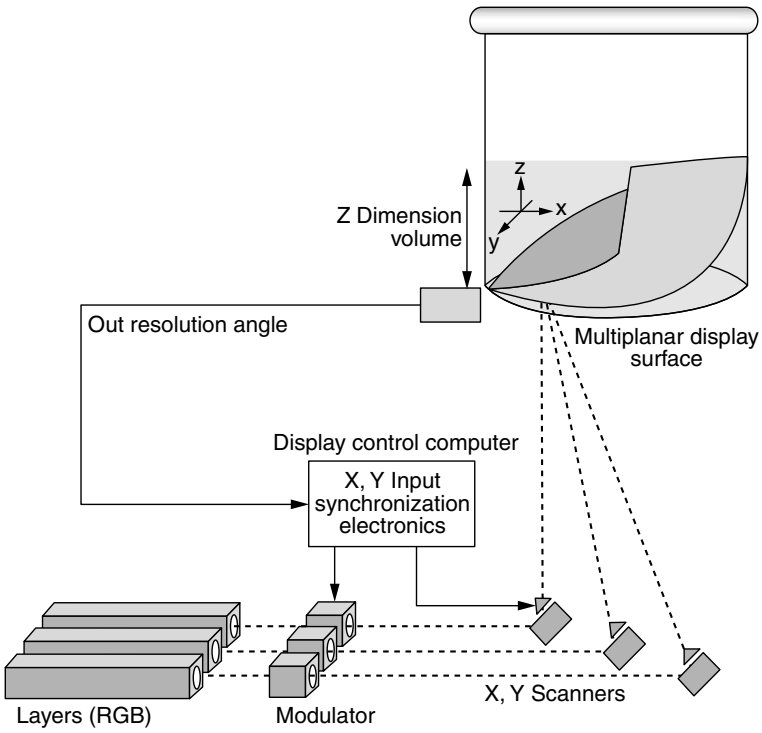


Figure 32. Omniview volumetric display.

program determines which point from each plane is to be rendered, so that lines appear to be continuous and uniform in thickness and brightness. The resulting image is solid. The view volume depth is approximately 72 times the mirror displacement depth at its center.

The images produced by the CRT would have to be warped to handle the varying focal length of the mirror. Such a mirror was produced by several companies in the past. At that time, only a green phosphor existed which had a sufficiently fast decay rate to prevent image smear.

Rotating Mirror

A variant of this approach developed by Texas Instruments using RGB lasers for point plotting and a double helix mirror rotating at 600 rpm as a reflecting device was also commercially available for a time under the name of Omniview (Fig. 32).

Some recent efforts have included LCD displays, but the switching times are currently too slow to produce useful images.

Problems and Advantages

A major advantage of multiplanar displays is that they are "solid." Accommodation and convergence are not disconnected, as in viewing stereo pairs where the visual system always focuses at the same distance. Users who are stereo-blind can see the depth, and the image is viewable by several people at once.

The primary problem that these mirror-oriented technologies have is that the images they produce are transparent. The amount of information they can represent before the user becomes confused is small because of the absence of hidden surface elimination, although head trackers could be implemented for single view use. In addition, they are limited to showing computer-generated images.

Another major disadvantage of multiplanar displays has been that the electro-optics and point-plotting devices used to produce the image are not sufficiently fast to produce more than a few points at a time on a 3-D object, and laser grids are far too expensive to generate good raster displays. Hence, multiplanar or volumetric displays have been limited to wire frame renderings.

Acknowledgments

The author thanks the following individuals who contributed to this article:

Marc Highbloom, Denise MacKay, VREX; Shihoko Kajiwara, Seaphone, Inc.; Jesse Eichenlaub, Dimension Technologies, Inc.; Jeff Wuopio, StereoGraphics, Inc.; Richard Steenblik, Chromatek; David McConville, Elumens Corporation; Vivian Walworth, The Rowland Institute for Science; Shunichi Kishimoto, Sanyo Corporation; Jeff Brum, Fakespace Systems, Inc.; Michael Starks, 3-DTV Corp.; David McConville, Elumens Corp.; Stephen Hines, HinesLab, Inc.; and Mark Holzbach, Zebra Imaging, Inc.

BIBLIOGRAPHY

1. T. Okoshi, *Three-Dimensional Imaging Techniques*, Academic Press, NY, 1976.
2. D. F. McAllister, ed., *Stereo Computer Graphics and Other True 3-D Technologies*, Princeton University Press, Princeton, NJ, 1993.
3. H. Morgan and D. Symmes, *Amazing 3-D*, Little, Brown, Boston, 1982.
4. J. Lipscomb, *Proc. SPIE: Non-Holographic True 3-D Display Techniques*, 1989, Vol. 1083, pp. 28–34, LA.

5. *Multi-pieced, portable projection dome and method of assembling the same*. US Pat. 5,724,775, March 10, 1998, R.W. Zobel Jr. et al.
6. *Tiltable hemispherical optical projection systems and methods having constant angular separation of projected pixels*. US Pat. 5,762,413, June 9, 1998, D. Colucci et al.
7. *Systems, methods and computer program products for converting image data to nonplanar image data*. US Pat. 6,104,405, August 15, 2000, R.L. Idaszak et al.
8. *Visually seamless projection screen and methods of making same*. US Pat. 6,128,130, October 3, 2000, R.W. Zobel Jr. et al.
9. C. W. Tyler and M. B. Clarke, *Proc SPIE: Stereoscopic Displays and Applications*, 1990, Vol. 1256, pp. 187, Santa clara.
10. D. Bar-Natan, *Mathematica.* **1**(3), pp. 69–75 (1991).
11. *Autostereoscopic display with illuminating lines and light valve*, US Pat. 4,717,949, January 5, 1988, J. Eichenlaub.
12. *Autostereoscopic display illumination system allowing viewing zones to follow the observer's head*, US Pat. 5,349,379, September 20, 1994, J. Eichenlaub.
13. *Stroboscopic illumination system for video displays*, US Pat. 5,410,345, April 25, 1995, J. Eichenlaub.
14. T. Hattori, T. Ishigaki et al., *Proc. SPIE*, 1999, Vol. 3639, pp. 66–75, San Jose.
15. D. Swinbanks, *Nature*, **385**(6,616), 476 Feb. (1997).
16. *Stereoscopic display*, US Pat. 6,069,649, May 30, 2000, T. Hattori, San Jose.
17. K. Mashitani, M. Inoue, R. Amano, S. Yamashita, and G. Hamagishi, *Asia Display* **98**, 151–156 (1998).
18. S. Hines, *J. Soc. Inf. Display* **7**(3), 187–192 (1999).
19. *Autostereoscopic imaging system*, US Pat. 5,430,474, July 4, 1995, S. Hines.
20. *Multi-image autostereoscopic imaging system*, US Pat. 5,614,941, March 25, 1997, S. Hines.
21. B. Blundell and A. Schwarz, *Volumetric Three Dimensional Display Systems*, Wiley, NY, 2000.

STILL PHOTOGRAPHY

Russell Kraus
Rochester Institute of Technology
Rochester NY

INTRODUCTION

Camera still imaging: The use of a lighttight device that holds a light-sensitive detector and permits controlled exposure to light by using a lens that has diaphragm control and a shutter, a device that controls the length of exposure. Controlled exposure is simply taken as an amount of light during a continuous period of time that produces a desired amount of density on the film after development. The desired amount of exposure is typically determined by either experimental methods know as sensitometry or through trial and error.

The shutter range for exposure can be between several hours and 1/8000th of a second, excluding the use of a stroboscopic flash that permits exposure times shorter than a millionth of a second.

The image has a size or format of standard dimensions: 16-mm (subminiature), 35-mm(miniature), and 60-mm (medium format). These are commonly referred to as roll film formats. Four inches by five inches, 5×7 inches, and 8×10 inches are three standard sheet film sizes. A view camera, monorail camera, or folding type is used to expose each sheet one at a time. Folding type, technical cameras that have a baseboard can be single sheet exposure or can adapt to a roll film back of reduced dimensions. All current camera systems can either replace their film detectors with digital detectors or themselves are replaced entirely by a digital version.

The nature of current still photography can be seen in its applied aspects: Documentary, reportage, scientific recording, commercial/ advertising, and fine art shooting are the primary realms of the professional photographer. In each activity, care is given to the materials, equipment, and processes by which a specific end is achieved. Scientific and technical photographers use the photographic process as a data collection tool where accuracy in time and space is of paramount importance. Photography for the commercial, documentary, and fine arts photographers has never been an objective and simple recording of an event or subject. Documentary photography before the Farm Security Administration attempts at reporting the Dust Bowl of the 1930s illustrated a point of view of the photographer. Photography was not simply a moment of captured time, an opening of a window blind to let in the world through a frame, but rather the photograph was the result of a complex social and political view held by the image-maker. The documentary/reportage photography of Hine, Evans, and Peress represent their unique vision and understanding of their times, not a naïve recording of events. This type of photography has as much purpose and artifice as commercial/advertising shooting. The visual selling of product by imagery designed to elicit an emotional response has been an integral part of advertising for more than 100 years. The psychology of photography has remained relatively the same during this time, albeit photocriticism has had many incarnations, but the technology that has made photography the tool it is has changed very rapidly and innovatively during the last century. The current digital evolution in photography will further advance the tools available and alter the way images are captured and displayed.

A BRIEF HISTORY

The history of photography traces the coincidence of two major technical tracks. The first track is the optical track. It includes the history of the *camera obscura.* Mentioned by Aristotle in the fourth century B.C., the *camera obscura* is basically a dark chamber useful for projecting the world through a pinhole into a closed, dark space. Described for the viewing of a solar eclipse, this pinhole device remained the basis for "camera imaging" for more than 1000 years. By the thirteenth century A.D. a lens had been added to the

device, at least in theory. Three hundred years later at the height of the Italian Renaissance a number of scientifico-artisans (Cardano and della Porto) mention or describe the *camera obscura* using a lens. By the mid sixteenth century, Barbaro describes the use of a lens in conjunction with a diaphragm. The diaphragm allows the camera to project a finely sharpened image by virtue of stopping down to minimize aberrations and to control the intensity of the light passing through the lens. The diaphragm or stop and the lens limit the width of the beam of light passing through the lens. The changeover from double convex lens to a meniscus type lens at the end of the eighteenth century limited the aberrations produced by the *camera obscura*. The prototype of the modern camera was born.

There are other modifications that gradually came into existence. Optical achievements in the areas of astronomy and microscopy contributed toward contemporary photographic equipment. Sturm, a mathematician, produced the forerunner of the modern single-lens reflex camera by the later part of the seventeenth century. Photography had to wait approximately 150 years before the idea of permanently capturing an image became a practicality, and modern photography began its journey into the twenty-first century.

The idea of permanently fixing an image captured by the camera obscura must have been in the ether for hundreds of years. Successful fixing took the independent efforts of Louis Daguerre and Joseph Niepce who approached the problem of permanent imaging from different vantage points. Both Daguerre and Niepce followed the trail broken by others to permanent pictures. The chemist, Johann Heinrich Schulze, is credited with the discovery of the light sensitivity of silver nitrate ($AgNO_3$), and a decade later in England, Sir Josiah Wedgewood of Wedgewood China fame employed $AgNo_3$ to make temporary photograms. These silhouettes were made by exposing silver nitrate coated paper to sunlight. Silver nitrate darkened upon exposure to sunlight, and those areas of the coated paper covered by an object remained white. Unfortunately, Wedgewood could not prevent the $AgNO_3$ paper from darkening *in toto* after time. While Niepce and Daguerre worked in France, Sir John Herschel and Wm. Henry Fox Talbot contributed to the emerging science of photography from across the channel. Talbot is credited with the invention of the negative-positive process resulting in the Calotype in photography. A coating of silver iodide and potassium iodide is applied to heavy paper and dried. Before exposure, a solution of silver nitrate, acetic acid, and gallic acid is applied. Immediately after exposure to bright sunlight, the paper is developed in a silver nitrate and gallic acid solution. Washing, fixing in sodium thiosulfate, and drying completed the processing of a Calotype paper negative.

Herschel's chemical tinkering led to his formulation of "hyposulfite." "Hypo" was a misnomer for sodium thiosulfate; however, the term hypo has remained a photographer's shortspeak for fixer. The invention of this chemistry permitted fixing the image. The removal of light-sensitive silver nitrate by this solution prevented the image from darkening totally under continued exposure to light.

Daguerre continued to proceed with his investigations into the direct positive process that was soon to be known by his name, the daguerreotype, and by 1837, he had produced permanent direct positive images.

The daguerreotype depended on exposing silvered metal plates, generally copper, that had previously been in close proximity to iodine vapors. The sensitized plate is loaded into a camera for a long exposure, up to one hour. Development was managed by subsequent fuming of the plate in hot mercury vapors (60°C). The plate is finally washed in a solution of hot salts that remove excess iodine, permanently fixing the image (Fig. 1).

For 15 years, the daguerreotype achieved world-wide recognition. During these years, major improvements were made in the process. Higher sensitivity was achieved through by using faster lenses and bromide-chlorine fuming. Images that formerly required 30 minutes or more of exposure could be made in 2 minutes. This process produced an amazing interest in photography, and an untold number of one of a kind images were produced, but the daguerreotype eventually gave way to the negative-positive process of Talbot and Herschel. Modern still photography had been born. Contemporary still photography begins where the daguerreotype ends. The ascendancy of the negative-positive process is the fundamental basis of film-based photography.

IMAGE RECORDING

Photography is distinguished by the idea that the image that will be recorded is of a real object. The recording of the image is the result of the optics involved in collecting the reflected light, and the recorded density range is a relative recording of the tones of the subject created by the source light falling, illuminance (luminous flux per unit area), on the subject. The establishment of a viewpoint controls perspective. The photographer must first choose

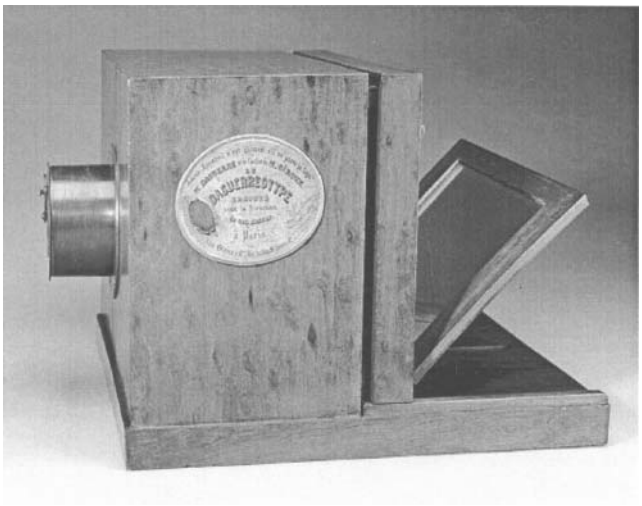


Figure 1. A Giroux camera explicitly designed for the daguerreotype.

the location from which the photograph will be taken. This achievement of perspective is essential to creating an appearance of a three-dimensional image on a two dimensional surface. The image projected by the lens is seen on ground glass placed at the focal plane or reflected by a mirror arrangement as in a single-reflex camera. The ground glass organizes or frames the image. In miniature and medium format photography, the photographer is looking at the image through a viewfinder. The image produced in the viewfinder shows the relative positions of objects in the scene as images produced by the lens in the same relative relationship. Distance is suggested in the image by the relative image sizes of objects in the field of view.

PERSPECTIVE

Photographic perspective is also controlled by the choice of optics. Using short focal length lenses, a wide angle view of the scene is imaged in the viewfinder. The angle of view determines how much of the image will be projected by the lens onto the frame. In practice, the frame is considered to the film size, horizontal × vertical dimensions. The lens focal length and the film size determine the angle of view (Fig. 2).

The wider angle permits projecting a greater area, and objects in the scene will appear smaller in image size. A change to a longer than normal focal length lens limits the area projected due to a narrower angle of view, but the relative size of the image is greater than normal. In both cases, the change in relative size of the images depends solely on the distance of the objects from the camera. Both the position of the camera and the lens focal length control perspective. Image size is directly proportional to focal length and inversely proportion to the distance from the viewpoint and camera position. For a given lens, a change in film size alters the angle of view, for example, a 4×5 inch reducing back used in place of a 5×7 inch film plane decreases the angle of view.

Perspective creates the illusion of depth on the two-dimensional surface of a print. Still photography employs several tools to strengthen or weaken perspective. Control of parallel lines can be considered part of the issue of perspective. Linear perspective is illustrated by the convergence of parallel lines that give visual cues to the presentation of implied distance in the photograph. This is shown through the classic example of the illusion of converging railroad tracks that seem to come together at a vanishing point in the distance. The image size of the cross ties becomes smaller as the their distance from the camera increases. We are more used to seeing horizontal receding lines appear to converge than vertical lines. In photographing a building from a working distance that requires the camera to be tilted upward to capture the entire building, the parallel lines of the building converge to a point beyond the film frame. When the film format is presented to the scene in portrait fashion, the vertical dimension of the film is perpendicular to the lens axis, parallel lines in an architectural structure will cause the image of the building to fall away from the viewer. The resultant image is known as keystone distortion. This can be controlled by the photographer by using of a view or field camera whose front and rear standards can tilt or rise.

On a smaller camera, 35-mm or 6-cm, a perspective control lens can be used for correction. This type of lens allows an 8 to 11-mm shift (maximum advantage occurs when the film format is rotated so that the longer dimension is in the vertical direction) and a rotation of 360°. Small and medium format camera lenses permit tilting the lens relative to the film plane. In a relatively short object distance, the tilt-shift or perspective control (PC) lens is very helpful. The PC lens has a greater angle of view than the standard lens of the same focal length. This larger angle of view allows the photographer to photograph at a closer working distance. The angle of view can be calculated as follows: $2 \tan^{-1}(d/2/f)$: where d is equal to the dimension of the film format, and f is the focal length of the lens. This larger angle of view is referred to in relation to its diagonal coverage. These lenses have a greater circle of good definition. This larger circle of good definition permits shifting the lens by an amount equal to the difference between the standard circle of good definition (which is usually equal to the diagonal of the image frame) and the circle of good definition for the PC lens that has the greater angle of view (Fig 3).

Further, assume that the image size of a building has a 0.001 scale of reproduction for a 35-mm lens at a working distance of 35 m. Therefore, every 1-mm upward shift of the PC lens causes a 1-m downward movement of the object. The effect generated by the perspective control lens is the same as the rising front on a view/field camera. Objects far a away subtend a smaller angle and create an image that is so small that the details of the object are unresolvable. The photographer can also use this lack

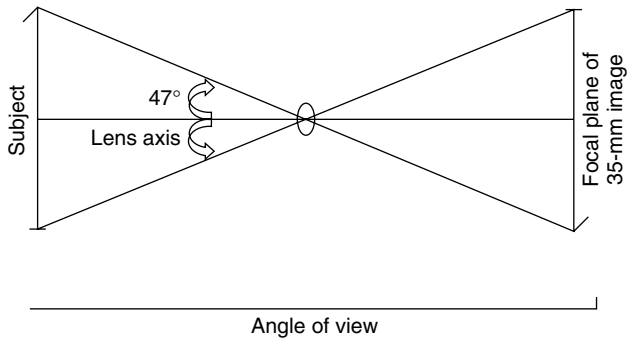


Figure 2. Angle of view of a lens–film format combination.

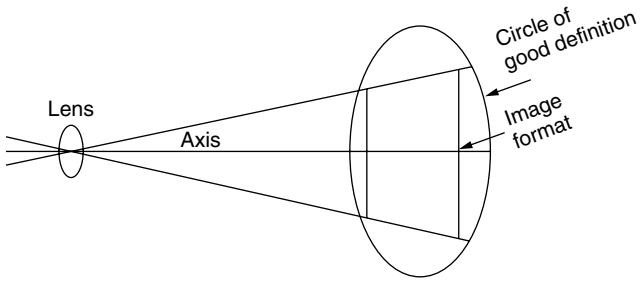


Figure 3. Format and circle of good definition.

of detail by the controlling the depth of field (DOF). A narrow DOF causes objects beyond a certain point of focus to appear unsharp, thus fostering a sense of distance. Camera lenses focus on only one object plane but can render other objects acceptably sharp. These other objects in front of and behind the plane of focus are not as sharp. These zones of sharpness are referred to as depth of field.

Acceptable sharpness depends on the ability of the eye to accept a certain amount of blur. The lens does not image spots of light outside the focused zone as sharply as objects that are focused. These somewhat out of focused spots are known as circles of confusion. If small enough, they are acceptable to the average viewer in terms of perceived sharpness. When the sharpness of objects in the zones before and after the zone of focus cannot be made any sharper, then the circles of confusion are commonly referred to as permissible circles of confusion. In 35-mm format, these permissible circles of confusion typically have a diameter of 0.03 mm in the negative. This size permits magnification of the negative to 8 to 10× and maintains acceptable sharpness when the print is viewed at a normal viewing distance. When the lens renders permissible circles of confusion in front of and behind the focused zone, these zones are referred to as the near distance sharp and the far distance sharp. At long working camera-to-subject distances, near and far sharpness determine depth of field (distance far sharp minus distance near sharp). In practice, the hyperfocal distance, the near distance rendered sharp when the lens is focused at infinity, for a given $f\#$ is used to determine the near distance sharp and the far distance sharp. Hyperfocal distance is focal length squared divided by $f\#$ times the circle of confusion, $H = fl^2/(f\# \times cn)$. Therefore, the near and far distances sharp can be calculated as follows: $D\mathrm{n} = Hu/[H + (u - fl)]$ where u is the distance from camera lens to subject and $D\mathrm{f} = Hu/[H - (u - fl)]$.

In studio photography, space is limited, a sense of depth can be fostered by the placement of lighting. The casting of foreground and background shadows and the creation of tonality on curved surfaces suggest greater depth than is actually there. In outdoor landscape photography, the inclusion of some near camera objects such as a tree branch in relationship to a distance scene cues the viewer to the distances recorded.

LENS CHOICE AND PERSPECTIVE

Photographs have either a strong or weak perspective that is permanently fixed. In three-dimensional viewing of a scene, the perspective and image size change in response to a change in viewing distance. This does not occur in photographs of two-dimensional objects. But in photographs containing multiple objects at different image planes or converging parallel lines, viewing distance can have an effect. Viewing distance changes influence the sense of perspective. Viewing distance is equal to the focal length of the camera's lens when the print is made from contact exposure with the negative. When an enlarger is used, the focal length of the camera lens must be multiplied by the amount of magnification. Thus, if a 20-mm lens is used on a 35-mm camera and the negative is magnified by projection to an image size of 4 × 6 inches, the correct viewing distance is 80 mm. This has the effect of changing the perspective from a strong to a normal perspective. People tend to view photographs from a distance that is approximately the diagonal measurement of the print being viewed, thereby accepting the perspective determined by the photographer. In the previous example, 80 mm is too close for people to view a photograph. Seven to 8 inches is most likely to be chosen as the correct viewing distance. This maintains the photographer's point of view and the strong perspective chosen.

The choice of a wide-angle lens to convey strong perspective carries certain image limitations. Wide-angle close-up photography of a person's face can present a distorted image of the person. The nose will appear unduly large in relation to the rest of the face. This would occur if the photographer wishes to fill the film frame with the face and does so by using a short object distance and a wide-angle lens. To maintain a more naturalistic representation, an experienced photographer will use a telephoto lens from a greater distance. Other distortions arise when using wide-angle lenses. In large group portraits where the wide-angle lens has been chosen because of the need to include a large group and object distance is limited by physical constraints, the film receives light rays at an oblique angle. Spherical objects such as the heads of the group portrait will be stretched. The amount of stretching is determined by the angle of the light from the lens axis that forms the image. The image size of the heads changes relative to the reciprocal of the cosine of the angle of deviation from the lens axis. This type of distortion occurs in using normal focal length lenses, out the amount of stretch is much less, and because the final print is viewed at a normal viewing distance, the viewer's angle of view approximates the original lens' angle of view and corrects for the distortion.

NEGATIVE-POSITIVE PROCESS

The process of exposing silver halide light-sensitive materials, in camera, relies on the range of density formed on a developed negative. Light reflected, luminance (intensity per unit area), from a subject is collected by a lens and focused as an image at a given distance behind the lens, the focal plane. The purposeful act of exposure and the concomitant development given to the exposed film achieves a tonal response or range of density that captures the relative relationships of tones in the original scene, albeit in reverse or negative state. Whites of uniform brightness in a scene, considered highlights, appear as a uniform area of darkness or relatively greater density in the negative. Blacks of uniform darkness in a scene, considered shadows, appear as uniform areas of lightness or a lesser density in the negative. Other tones between whites and blacks are rendered as relative densities between the extremes of black and white. This is the continuous tone or gray scale of photography. The negative's density range must be exposed to other silver sensitized material to convert the reversed tones into a positive. The negative is made on a transparent base

that permits its subsequent exposure to a positive, either by contact or projection. There are obvious exceptions to the general practice of the negative-positive photographic approach in image making. Positives can be made directly on film for projection, direct duplication, or scanning. These transparencies may either be black-and-white or color. Likewise, direct positives can be made on paper for specific purposes. This process of representing real world objects by a negative image on film and then by a positive image on paper is basic to traditional still photography.

CAMERAS

Comments on the breadth of available camera styles, formats, features and advantages and disadvantages are beyond the scope of this article. However, certain specific cameras will be mentioned in terms of their professional photographic capabilities. The miniature camera is the 35-mm format. This format is represented by two types: compact and ultracompact. There are fixed-lens and interchangeable-lens type 35-mm cameras. The latter type, unlike the former, constitutes the basis of an imaging system that can be further subdivided into two basic types: rangefinder and single-lens reflex cameras. Both single-lens reflex and rangefinder style cameras are available as medium and special format cameras as well (Fig. 4).

Professional 35-mm camera systems are characterized by their interchangeable lens systems, focal plane shutters, motorized film advance and rewinds, electronic flash systems, extra length film magazines, and highly developed autofocusing and complex exposure metering systems.

Lenses from ultrawide-angle focal lengths, fish-eye (6-mm), to extremely long focal lengths (2,000-mm) can replace the typical normal (50-mm) focal length lens. Special purpose lenses such as zoom lenses of various focal lengths and macro lenses that permit close-up photography resulting in image to object ratios greater than 1 : 1 are two of the most common specialty lenses.

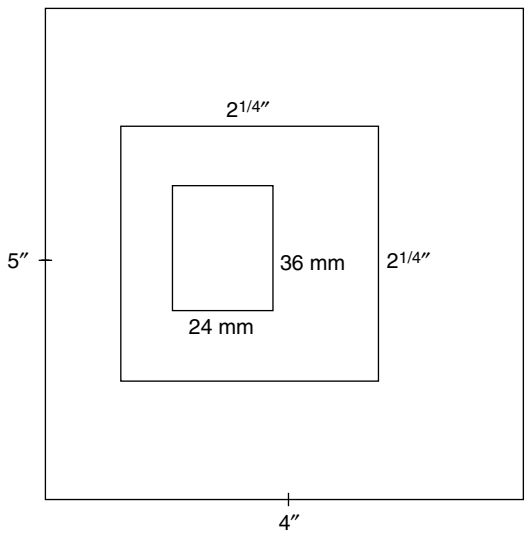


Figure 4. Film/camera formats showing relative differences among them.

Among the specialty lenses that are of interest is the fish-eye lens. A wide-angle lens can be designed to give a greater angle of view if the diagonal format covered is reduced relative to the standard format. This unique lens projects a circular image whose diameter is 21 to 23-mm onto the 43-mm diagonal dimension of the 35 mm film format. Hence, it provides a circular image of the scene. The true fish-eye angle of view is between 180 and 220°. A normal wide-angle lens that has a rectilinear projection can achieve a very short focal length (15-mm) until its peripheral illumination noticeably decreases by the $\cos^4$ law. A change in lens design to one of retrofocus does not alter this loss of illumination. Other projection geometry is used to permit recording angles of view greater than 180°. Equidistant, orthographic, and equisolid angle projections are used to increase angles of view. The image projection formula for a rectilinear projection is $y = f \tan \theta$, for equidistant projection, $y = f\theta$, and for equisolid angle projection, $y = 2f \sin(\theta/2)$.

Motor Drive

A motor drive permits exposing of individual frames of film at relatively high film advance speeds. Unlike motion picture film cameras that expose at a rate of 24 frames per second continuously, motor drives advance a film frame at typically from three to eight frames per second. An entire roll of 35-mm 36-exposure film may be rewound by the motor in approximately 4 seconds. State-of-the-art 35-mm camera systems provide built-in film advance systems and do not require separate add-on hardware. When coupled with advanced fast autofocusing, the photographer has an extraordinary tool for recording sports action, nature, and surveillance situations.

Electronic Flash

Specialized flash systems are another feature of the advanced camera system. The fundamental components are power supply, electrolytic capacitor, reflector design, triggering system, and flash tube. Some triggering systems control the electronic flash exposure through the lens and are referred to as TTL. TTL cameras permit evaluating light that reaches the film plane. Variations of the basic TTL approach are the A and E versions of TTL. The A-TTL approach uses a sensor on the flash unit to evaluate the flash in conjunction with the aperture and shutter settings of the camera. E-TTL uses the camera's internal sensor to evaluate the flash and to set the camera aperture. Another variant uses the camera-to-subject distance determined by the camera's focusing system to control the flash duration. Modern camera systems communicate with the camera via a hot shoe and or cable (PC) connection. Professional cameras can program electronic flash to allow for high-speed flash at a shutter speed of 1/300 of a second. This is achieved by dividing the current discharged by the capacitor into a series of very short pulses.

There are occasions for flash exposure at the end of the shutter travel. The rear-curtain synchronization allows continuous illumination of a moving subject that would produce any associated movement blur trailing the subject. X-synchronization flash would occur when the shutter is

first fully opened. In X synchronization, the flash occurs when the first curtain is fully open and the following curtain has not begun to move; the blur of a moving image occurs in front of the image. Electronic flash systems produce high intensity, ultrashort bursts of light from the discharge of an internal capacitor. This discharge occurs in a quartz envelope filled with xenon gas. Typical effective flash excitation occurs within 1 millisecond. When used with a thyristor that controls exposure, the flash duration can be as short as 0.02 milliseconds (Fig. 5).

The thyristor switches off the flash when a fast photodiode indicates that sufficient exposure has occurred. This system allows fast recharging of the capacitor to the appropriate current if an energy efficient design is used. This type of design uses only the exact level of charge in the capacitor needed for a specific exposure. Fast recycling times are available, albeit at power levels well below the maximum available. The quenching tube design permits dumping excess current to a secondary, low-resistance flash tube after the primary flash tube has been fired. While the accuracy of the exposure is controlled, the current is completely drained from the capacitor. Recycling times and battery life are fixed.

The output of an electronic flash is measured in lumens per second, a measure of the rate of flux. A lumen is the amount of light falling on a uniform surface of 1 ft^2. The source is one candela at a distance of 1 foot from the surface. Because an electronic flash emits light in a defined direction, a lumen can be considered the amount of light given off in a solid angle. A more photographic measure is beam-candle-power seconds (BCPS), a measure of the flash's output at the beam position of the tube. Beam-candle-power seconds is used to determine the guide number for flash effectiveness. An electronic flash's guide number can be expressed as $GN = (ISO \times BCPS)\ 0.5\ K$, where K is a constant, 0.25 if distance is in feet or 0.075 if distance is measured in meters. Guide numbers express the relationship between object distance and aperture number. If a guide number is 88 for a given film speed/flash combination, the photographer can use an aperture of $f/8$ at the camera-to subject-distance of 11 ft or $f/11$ at 8 feet. At 16 feet, the aperture would need to be set two stops wider or $f/5.6$. The inverse square law governs this relationship. The law states that illumination increases or decrease as a function of the square of the distance ($E = I/d^2$) from a subject to the point source of light. Guide numbers are approximations or initial exposure recommendations. The environment of the subject can alter the guide number, plus or minus, by as much as 30%.

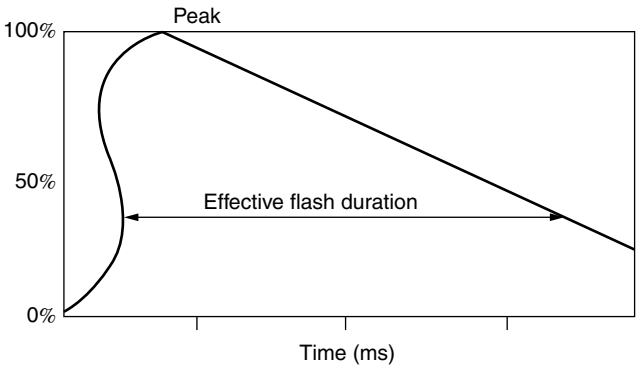


Figure 5. Flash curve and time.

The type of electronic flash and the placement of the flash have a significant effect on the photographic image. Flash units mounted on the reflex prism of the camera are also on-axis (2° or less) with the lens. The combination of this location and the camera-to-subject distance is the source of "red-eye", the term given to the reflection of the flash by the subject's retina. The reddish hue is caused by the blood vessels. This particularly unpleasing photographic effect is also accompanied by flat and hard lighting. Moving the flash head well off center and bouncing the flash off a reflector or nearby wall can provide more pleasing lighting. When shadowless lighting is desired, a ring flash can be used. A ring flash uses a circular flash tube and reflector that fits around the front of the lens. Ring flash systems generally have modeling lighting ability to assist in critical focusing. Scientific, technical, and forensic applications use this device.

Single-lens Reflex Camera

The single-lens reflex (SLR) camera is the most widely used miniature format, professional, camera system. Having supplanted the rangefinder as the dominant 35-mm system by the late 1970s, its distinctive characteristic is direct viewing and focusing of the image by a mirror located behind the lens and in front of the film gate. This mirror reflects the image formed by the lens to a ground glass viewing screen. The mirror is constructed of silver- or aluminum-coated thin glass or metal. The coatings prevent optical distortions and increase the illuminance reflected to the viewing screen. The mirror is characterized by its fast return to its rest position after each exposure. The viewing screen is viewed through a viewfinder or pentaprism that laterally reverses the image. The image is in view in correct vertical and horizontal relationship to the subject. Mirror shape and length are two important design considerations that impact performance. Maximum reflectance depends on the intersection of the light path exiting the lens and mirror location. Mirrors can be trapezoidal, rectangular, or square and are designed to intersect the exiting cone of light best. Mirror length can impact the overall size of the camera housing and/or the viewing of the image. When the mirror is too short, images from telephoto lenses are noticeably darker at the top and bottom of the viewing screen. Overly long mirrors necessitate deep camera bodies or a lens system that uses a retrofocus design. Some manufacture's use a hinged mirror that permits upward and rearward movement during exposure. Secondary mirrors that are hinged to the main mirror are used to permit autofocusing and exposure metering through the lens (Fig. 6).

Almost all professional SLR systems have a mirror lockup feature that eliminates vibrations during exposure. This is very useful for high magnification photography where slight vibration can cause a loss of image quality, and for very long telephoto use at slow shutter speeds where a similar loss of image quality can occur. It is noteworthy to mention that long focal lengths (telephoto)

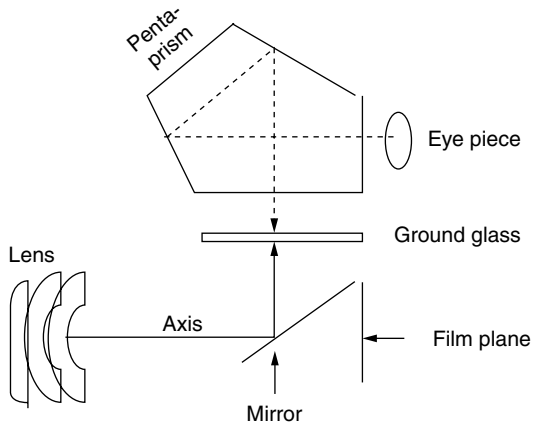


Figure 6. Lens, mirror, and pentaprism arrangement.

from 150 to 2,000 mm are more accurately focused in a SLR system than in a rangefinder system.

The Rangefinder

In 1925, Leica introduced a personal camera system that established the 35-mm roll format and the rangefinder type camera as a professional photographic tool. The modern-day camera is compact and easy to focus, even in low light. It is compact and quiet because it does not require a reflex mirror mechanism, and it permits using a wide-angle lens of nonretrofocusing design.

The rangefinder establishes the focal distance to the subject by viewing the subject from two separate lines of sight. These lines converge at different angles, depending on the working distance. At near distances, the angle is greater than at far distances. The subject is viewed through one viewing window and through a mirror that sees the subject through a second window. The distance between the two windows is the base length. By virtue of a sliding mirror or rotating mirror, the subject is viewed as two images that coincide on each other. The mirror or mirrors may be silvered or half silver and can be made to image the subject in two halves, one image half (upper) above the other image half (lower), or one image half alongside the other image half. This vertical or horizontal coincidence is the basis for focusing the camera lens. There are several design variations for constructing a rangefinder system. One variation maintains both mirrors in fixed position, and the appropriate deviation is achieved by the inserting a sliding prism in the light between the two mirrors (Fig. 7).

Given that $\tan x = b/d$, where b = base length of the rangefinder, d = distance of the object, and $\tan x$ is the angle of rotation of the mirror, it is obvious that a minimum rotation can accommodate focusing from near to distant. When the rangefinding system is mechanically coupled to the lens (cams and gears), visual focusing through the eyepiece of the viewfinder and lens focusing of the subject now are in unison. When the superimposed images of the rangefinder are coincident, the lens is correctly focused on the subject.

Focusing accuracy depends on a number of factors: base length, focusing error tolerance, mechanical couplings, and

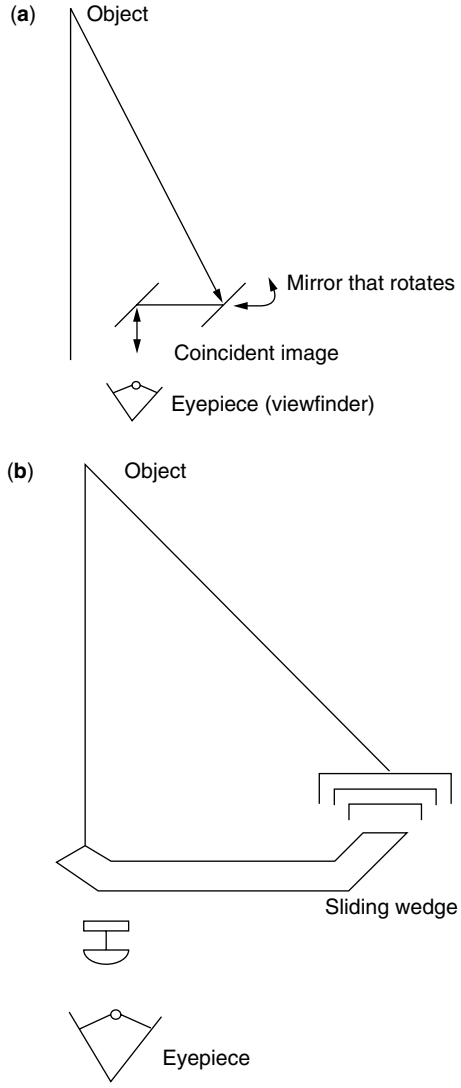


Figure 7. Drawing of rangefinder types.

image scale of reproduction in the viewfinder. Accuracy of focus is defined by the limits of the eye (using a standard of 1′ of arc) as an angle. Because the acuity of the eye can be influenced by a magnifier in the eyepiece of the optical viewfinder, rangefinder error is described as $2D^2a/Rb$, where Rb is the scale of reproduction × base length, D is distance, and a is the angle of 1′ of arc. Rb is usually referred to as true base length. Therefore, if the base length of a rangefinder camera is 150 mm and the scale of reproduction is 0.8, the true base length is 120-mm. Note that the scale of reproduction in all miniature cameras and smaller formats is less than one. This is necessary to permit the rangefinder image in the viewfinder window. There are some systems that permit attaching a scale of reproduction larger than one. This extends the true base length without expanding the physical dimensions of the camera. Telephoto lenses can focus on a subject less than 5 feet from the camera. The coupled rangefinder permits focusing a 50-mm lens on a 35-mm camera from infinity to 30 inches (0.75 m). However, if the focusing error is too

great, it will exceed the DOF for a given lens, aperture, and distance. The appropriate base length has a practical limit because there is a maximum focal length for a given base length. The base length becomes unwieldy, or the camera would require large magnification viewfinders for focal lengths in excess of 150 mm in a 35-mm system. This is expressed in the formula Rb = focal length squared/f-number times C, where C is the permissible circle of confusion in the negative.

At the end of the 1900s, the rangefinder camera had resurgence in the issuance of several new 35-mm systems, in the marketing of a large number of new APS (Advanced Photo System, a sub 35 mm Format) roll film cameras, and in the establishment of "prosumer" digital cameras.

Medium Format Cameras

These cameras are chiefly considered professional by virtue of their extensive accessories, specialized attachments, and their larger film format. The larger film format has distinct advantages in that the magnification used to reach final print output is generally less than that for the smaller format. Consequently, microimage characteristics of the film that may detract from overall print quality (definition) are limited by the reduced magnification required. The film format is generally unperforated roll type in sizes of 120 or 220, but specialized cameras in this classification using 120 roll film can produce images that are $2\frac{1}{4} \times 6\frac{3}{4}$ inches to 6×4.5 cm. Other film formats derived from 120 or 220 film are 6×6 cm, 6×7 cm, 6×8 cm, 6×9 cm and 6×12 cm. Seventy-millimeter perforated film stock used in long roll film-magazines is considered medium format. Two-twenty film essentially provides for double the number of exposures as the 120 film and has us backing paper. Nonperforated 120 roll film is backed by yellow and black paper. Exposure numbers and guide markings printed on the yellow side are visible through a viewing area (red safety window) on the camera back. When there is no viewing window, the guide marks are aligned with reference marks on the cameras' film magazine, the film back is shut, and the film is advanced forward by a crank until the advancing mechanism locks into the first frame exposure position. The film is acetate approximately 3.6 mils thickness and must be held flat in the film gate by a pressure plate similar to that employed in 35-mm cameras. Medium format cameras are generally SLR types, but rangefinder types and the twin lens reflex are also available. Many of the available medium format camera offer only a few lenses, viewfinders, and interchangeable back, and a few offer complete and unique systems that provide all of the advantages of a large format camera within the smaller and more transportable design of the medium format camera. More recently introduced in the $2\frac{1}{4}$ $s \times 2\frac{1}{4}$-inch format are systems that provide tilt and shift perspective control, a modified bellows connection between the front lens board and the rear film magazine, and interchangeable backs for different film formats and for Polaroid film. In some cameras, the backs can rotate to either portrait or landscape mode. These backs protect the film by a metal slide that is placed between the film gate and the camera body. The slide acts as an interlock, thereby preventing exposure or accidental fogging of the film when in place. This feature also permits exchanging film backs in mid roll. Other forms of interlocks prevent multiple accidental exposures.

The twin-lens reflex camera is a unique $2\frac{1}{4}$-inch square, medium format camera design. This design permits viewing the image at actual film format size. The image is in constant view and it is unaffected by the action of the shutter or advancing the film. However, its capabilities are affected by parallax error when it is used for close-up photography. The viewing lens of the camera is directly above the picture-taking lens. The mirror that reflects the image for viewing to a ground glass focusing screen is fixed. The camera is designed to frame the subject from a low viewpoint, generally at the photographer's waist level. The shutter is located within the picture-taking lens elements and is typical of other medium format shutter mechanisms.

Shutters

All types of shutter mechanisms may be found in all format cameras. Typically, one type of shutter may be mostly associated with a specific format camera. Focal plane shutters are found mostly in miniature format cameras, although some medium format cameras use this design. Likewise, leaf shutters, so-called between the lens shutters because of the leaflike design of their blades, are used mostly in medium and large format cameras, although they are found in some fixed-lens 35-mm cameras and in smaller format cameras as well (Fig. 8).

Specialized shutters such as revolving shutters that are generally considered typical in aerial technical cameras have been adapted to 35-mm systems.

Leaf shutters are located between lens elements or behind the lens itself. Ideally, the shutter should be at the optical center of the lens, parallel with the diaphragm. The blades of the shutter can be made to function as an iris diaphragm as well. The shutter and aperture are mostly part of one combined mechanism, the compound shutter, exemplified by manufacturers such as Compur, Copal, and Prontor. These shutters are numbered from 00 to 3 and reflect an increasing size that is necessitated by the increasing exit pupil diameter of the lens. For the most part, these lenses are limited to a top shutter speed of 1/500th of a second. X synchronization is available at all shutter speeds. When fired, the shutter opens from the center to the outer edges of the lens. Because the shutter does not move infinitely fast, the center of the aperture is uncovered first and stays uncovered the longest. The illuminance changes as the shutter opens and closes. Consequently, if exposure times were based on the full travel path of the shutter blades, the

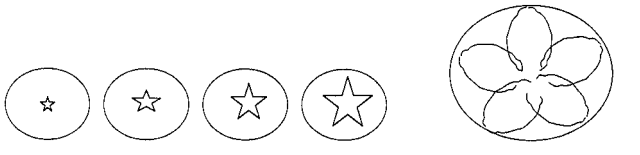


Figure 8. Illustration of leaf design.

resulting image would be underexposed. To correct for this, the shutter speed is measured from the one-half open position to the one-half closed position. This is known as the effective exposure time. Note that small apertures are exposed sooner and longer than larger apertures. Effective exposure time is longer for smaller apertures. This poses no problem for exposure at slower shutter speeds; however, as shutter speeds increase and apertures decrease, exposures move in the direction of overexposure. In a test of a Copal shutter, a set shutter speed of 1/500th of a second and an aperture number of *f*/2.8 produced a measured exposure time of 1/400th of a second. This difference between the set exposure time and the measured exposure time was minimal, less than one-third of a stop. However, when the diaphragm was stopped down to *f*/32 and the 1/500th of a second shutter speed was left untouched, the measured shutter speed was 1/250th of a second. This difference is attributed solely to the fact that the smaller aperture is left uncovered for a longer time. This one stop difference is significant (Fig. 9).

Between the lens shutters increase the cost of lenses for medium and large format cameras due to the built-in shutter for each lens. These shutters may be fully mechanical and have clockwork type gears and springs or may be electromechanical, using a battery, resistor-based timing circuits, and even quartz crystals timers. Resistors can be coupled in circuit with a photocell, thereby creating an autoexposure system where the shutter speed is determined by the amount of light sensed by the photocell relative to the aperture setting, that is, aperture priority.

Focal plane shutters have a distinct advantage because they are part of the camera. The cost of lenses does not reflect the need for shutters to be incorporated. Higher shutter speeds are possible, 1/8,000th of a second, and advanced electronics can synchronize flash to shutter speeds as fast as 1/300th of a second. The heart of modern focal plane shutter design is the Copal vertical travel shutter. Horizontal travel shutters are still used. Focal plane shutters derive their name from the fact that they are located close to the focal plane at the film or detector. This closeness avoids effective exposure time problems for small apertures as happens with between the lens shutters. Historically, the shutter was a slit in a curtain that was comprised of a leading and trailing edge; the film was exposed by the action of the slit scanning the film. The curtain traveled horizontally. Exposure time is the result of the width of the slit in the curtain divided by the velocity of the traveling curtain. The slit used for the exposure may in fact not be a slit in a curtain, but rather a set of titanium or metal alloy blades or the travel difference between two metallic curtains or blinds. High-speed shutters require blades made of durable and lightweight materials such as carbon fibers, and they require complex systems to control vibrations (shutter brakes and balancers) and to prevent damage to the blades. State-of -the-art shutters may include a self-monitoring system to ensure accuracy of speed and reliability. These systems are made even more complex by electronic components that are used to control exposure. The Hasselblad 200 series camera systems offer electronically controlled focal plane shutters that permit exposure times from 34 minutes to 1/2,000th of a second and flash synchronization up to 1/90th of a second.

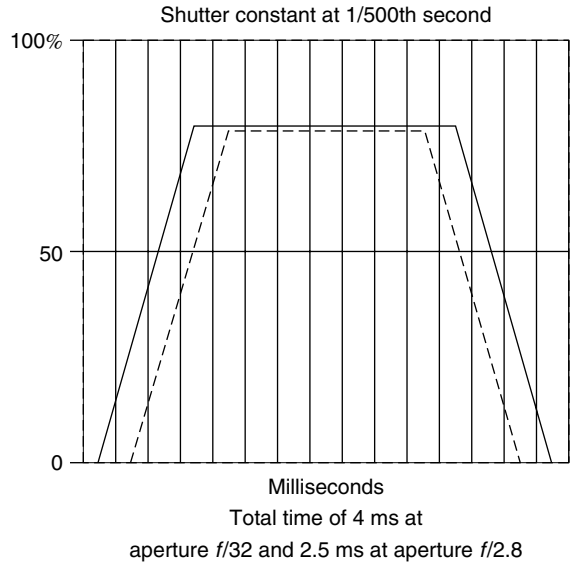


Figure 9. Oscilloscope trace of two apertures and fast shutter speed and an example of effective exposure time.

These electronic shutters permit aperture priority mode exposures that couple an atypical shutter speed (e.g., 1/325th of a second) to a selected aperture. Normal shutter speeds are established on a ratio scale whereby the speeds increase or decrease by a constant factor of 2. Unique to focal plane shutters is the capability of metering changing illuminance values for exposure control off-the-film plane (OTF). Focal plane shutters are identified mostly with a professional, 35-mm system and in rarer medium format systems. They are still rarer in large format camera systems but are found in barrel lens camera systems.

Large Format Cameras

The view camera, a large format camera, is the most direct approach for producing photographic exposures. A view camera is essentially a lens mounted on a board connected by a bellows to a ground glass for composing the image. The frame is further supported by using a monorail or for a field or technical camera, by folding flatbed guide rails. These field cameras may use rangefinder focusing in addition to direct viewing or an optical viewfinder. The view camera is capable of a series of movements. The lens plane and film plane can move independently. These movements are horizontal and vertical shifts for the lens and film planes, forward and rearward tilting of the lens and film planes, clockwise and counterclockwise swinging of the lens plane about its vertical axis, and clockwise and counterclockwise swinging of the film plane about its vertical axis. These movements control image shape, sharpness of focus, and location of the image on the film plane. Simply shifting the film plane or lens plane places the image appropriately on the film. This movement can avoid the need to tilt the camera (upward/downward) to include the full object.

Because the film and lens planes can be shifted independently, a shift of the lens plane in one direction

is equivalent to shifting the film plane an equal distance in the opposite direction. This type of shift, as well as the other movements, functions purposefully as long as the diagonal of the image formed by the lens covers the diagonal of the film plane format. The projection of the image circle in which sharpness is maintained is known as the circle of good definition. The circle of good definition increases in size when the lens to film-plane distance is increased to maintain good focus. Stopping down, that is, reducing the aperture, can increase the circle of good definition. The larger the circle of good definition, the greater the latitude of view camera movement. Often the angle of coverage is used as another measure of the covering power of the lens. The angle of coverage is unaffected by changes in the image distance. Movements can be done in tandem. When shifting the lens plane to include the full object being photographed, the shift may not be sufficient. Tilting the lens plane back and tilting the film plane forward can position the image within the area of good definition. Swinging the film plane around its vertical axis, so that it parallels the object being photographed, can eliminate the convergence of horizontal lines. This has the effect of making the scale of reproduction equal at the ends of all horizontal lines. This image effect is explained by the formula, $R = V/U$, where R is the scale of reproduction, V is the image distance, and U is the object distance. Therefore, given an object, where the two ends of its horizontal lines are at an object distance of 20 feet(left side) and 10 feet (right side), the image distance V must be adjusted by swinging the film plane so that V_r/U_l is equal to V_l/U_r. Swinging the lens plane will not produce the same effect because U_r is increased and V_l is increased; consequently, the ratio V/U is the same, and R is constant (Fig. 10).

Film plane movements can be used to improve the plane of sharp focus. However, when using film-plane swing capabilities to control image shape, the lens plane swing must be used to control sharp focus because adjusting the film plane around its vertical axis affects sharpness and shape simultaneously. Swinging the lens plane does not affect image shape. Focus control by the lens plane is limited by the covering power of the lens. Movement of the lens plane moves the image location relative to the film plane. Excessive movement can adjust the image beyond the circle of good definition and beyond the circle of illumination, that is, optical vignetting. An oblique pencil of light is reduced in illumination compared to an axial beam from the same source.

Physical features of the lens, such as lens length, can further impact optical vignetting. Vignetting can be reduced by using smaller apertures, but some loss of illumination will occur as a matter of course, natural vignetting, because the illumination falls off as the distance from the lens to the film increases. Illumination is inversely proportional to the square of the lens to film distance, $\cos^4$ law. It has been calculated that the $\cos^4$ law affects a normal focal length lens that has an angle of view of 60°, so that there is a 40 to 50% loss of illumination at the edges of the film plane. For lenses of greater angle of view such as wide angle lenses, a 90° lens could have as much as a 75% loss of illumination at the edges. Reverse telephoto lens, a design that permits a greater lens to film plane distance than a normal lens design for the same focal length.

The swings, shifts, and tilts achieved by the view camera provide a powerful tool for capturing a sharp image. When these tools are unavailable, the photographer can focus only at a distance where DOF can be employed to achieve overall acceptable sharpness in the negative. The relationship of the lens to film-plane distance expressed by the formula $1/f = (1/U) + (1/V)$ determines that objects at varying distances U are brought into focus as the lens to film-plane distance V is adjusted. The view camera that has independent front and back movements may use either front focusing when the lens to film-plane distance is adjusted by its movement or rear focusing when the film plane is moved closer to or further from the lens. Back focusing controls image focus, and front focusing alters image size as well as focus. It can be seen from the previous formula that U and V are conjugates and vary inversely. Back focus is required when copying to form an image of exact size. The plane of lens, the plane of the film, and the plane of the object are related by the Scheimpflug rule (Fig. 11). When the object is at an oblique angle to the film plane, the lens plane can be swung about its vertical axis or tilted around its horizontal axis, so that the three planes, object, film, and lens planes, meet at a common line. When the lens plane is parallel to the object plane, the film plane is swung or tilted in the

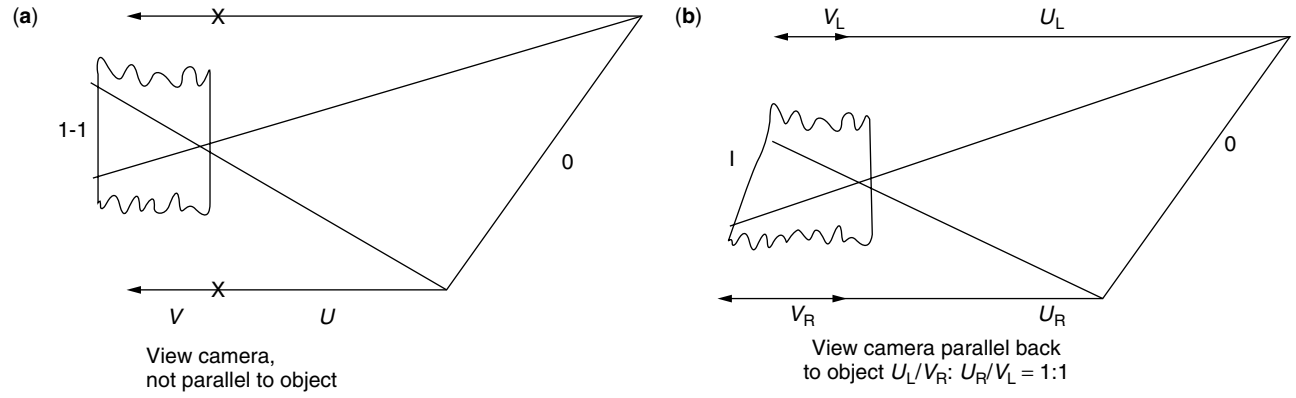


Figure 10. Diagram of the film plane swing.

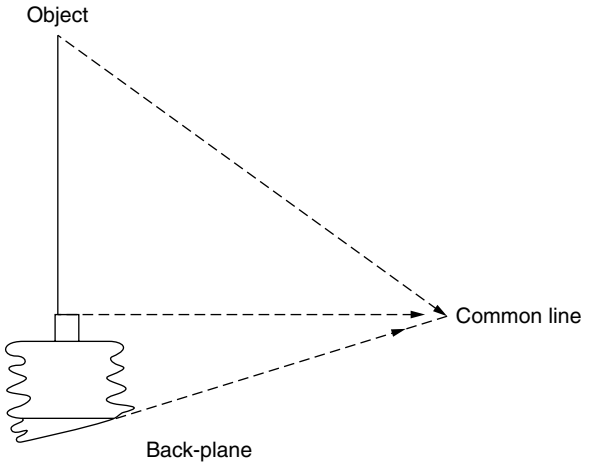


Figure 11. Illustration of the Scheimpflug rule.

opposite direction to achieve the three-plane convergence. The correct order of action is to adjust the back-plane to ensure correct image shape and then to adjust the lens plane to ensure sharpness. It is apparent that the image plane has different foci, left to right, when swung, and top to bottom when tilted. DOF must be used to further the creation of an overall sharp image.

There are limitations in using DOF. DOF calculations depend on the formula $C = f^2/(\mathrm{NH})$. Both f, focal length, and N, aperture number, alter the DOF. Doubling the size of the permissible circle of confusion would require a change of one stop of aperture. Depth of field is directly proportional to f-number. DOF increases as the object distance U increases, and it is expressed as $D_1/D_2 = (U_1)^2/(U_2)^2$; this is conditional on the caveat that the hyperfocal distance does not exceed the object distance. DOF increases as focal length decreases for a given image format. When comparing the DOF for two lenses for the same image format, the DOF ratio is equal to the focal length ratio squared.

Lenses and Image Forming Principles

Focal length is defined by the basic formula $1/f = (1/U) + (1/V)$. This is the foundation for a series of equations that describe basic image formation. I (image size) is equal to O (object size) $\times\, V/U$, except for near objects less than twice the focal length. Practically, $I = O \times f/U, I/O = f/U$, and the scale of reproduction R is equal to f/U. R is determined by focal length for any given distance, and for a specific focal length, R is determined by distance. Focal length determines the size of the image for an object located at infinity for any given film/detector size. The measured distance between the lens and the focused image is expressed as focal length. In camera systems where the lens is focused by varying the location of an element within the lens, the focal length is dynamic. The ability of the lens to project a cone of light of differing brightness is a function of the aperture control or iris diaphragm. The ratio of the focal length to the maximum diameter of the diaphragm (entrance pupil) is the lens' f-number (f#). f# = focal length / D, where D is the diameter of the entrance pupil. f-numbers are a ratio scale, 1 : 1.4, and the image illuminance changes by a factor of 2. In photographic parlance, this factor of change is referred to as a stop. The smaller the f#, the relatively brighter the image. The intensity of the image will be less than the intensity of the light falling on the lens. The transmission of the reflected light depends on a number of lens properties; absorption and reflection factors. When the focal length of the lens equals the diagonal of the image format of the camera, the focal length is considered "normal." The normal focal length lens has an angle of view of approximately 47 to 53° that is akin to the angle of view of the human eye. The 50-mm lens is the "normal" standard for 35-mm format photography. Medium format lenses have been standardized at 75 or 80 mm, and 4×5-inch cameras have a standardization range between 180 and 210 mm. There is a difference between the calculated normal lens determined by the diagonal of the film format and those that are actually found on cameras. The actual normal lenses are those of a longer focal length than would be required by the length of the diagonal.

Wide-angle lenses are characterized by the larger angle of view. The focal length of these lenses is much less than the diagonal of the image format that they cover. Because of the short lens to focal plane distance, problems of camera function may occur. In the SLR camera, the mirror arrangement may be impeded by using short focal lengths, and in view cameras, camera movement may be hindered. Reverse-telephoto wide-angle designed lenses (retrofocus) overcome such problems. The design requires placing a negative element in front of a positive element, thereby spacing the lens at a greater distance from the image plane. When the focal length of a lens is much greater than the diagonal of the film, the term telephoto is applied. The angle of view is narrower than the normal focal length. The telephoto design lens should not be confused with a long focus lens (Fig. 12).

In the telephoto design, the placement of a negative element/group behind the positive objective brings the cone of light to a focus as if it had been from a positive objective of greater focal length. The back nodal plane is now located to give a shorter lens to film distance than that of a lens of normal design for the same focal length. The opposite of this is short focal length wide-angle designs by which the distance is increased when a negative element is placed in front of the objective. This greater lens to film distance permits full use of the SLR mirror and more compact 35-mm camera designs.

Macro lenses primarily used in scientific and medical photography have found their way into other photographic venues. This has been made possible by the availability of "telemacro" lenses. These are not true macro lenses;

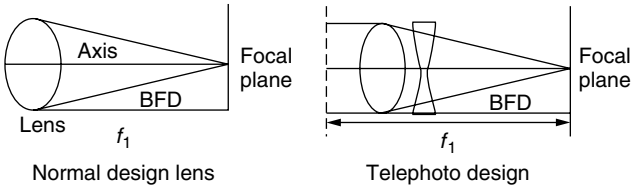


Figure 12

although they allow a close working distance, the image delivered is generally of the order of 0.5 magnification. True photomacrography ranges from 1 to 40×. The "telemacros" and some normal macros permit close-up photography at a working distance of approximately 3 to 0.5 feet and at a magnification of 0.10 to 1.0. These are close-up photography lenses, although they are typically misnamed macro-zoom lenses. Close-up photography requires racking out the lens to achieve a unit of magnification. The helical focusing mount limits the near focusing distance, so that other alternatives must be found to extend the lens to film distance, increase magnification, and shorten working distance. Many of these close-up lenses require supplemental lenses or extension tubes to achieve magnification beyond 0.5×. The classic Micro Nikkor 55-mm lens for the 35-mm format can magnify by 0.5×. An extension tube permits the lens to render 1× magnification. The 60-mm, 105-mm, and 200-mm Micro Nikkor achieve 1 : 1 reproduction without extension tubes. Long focus length macro lenses for miniature and medium format cameras require supplemental lenses to achieve magnification up to 3×. Positive supplemental lens focal lengths are designated in diopters. Diopter (D) power can be converted into focal length by the formula $f = 1$ (meter) $/D$. It is a practice to add supplemental lenses to each other to increase their power and increase magnification. The useful focal length is now the sum of all of the focal lengths in the optical system. The formula $1/f = (1/f_1) + (1/f_2) + (1/f_n)$ expresses this. When used with rangefinder cameras or TLL reflex cameras, an optional viewfinder that corrects for parallax must be used. The working distance relationship to the focused distance is determined by the formula $u_c = uf_s/(u + f_s)$; where u_c is the close-up working distance, f_s is the focal length of the system, and u is the focused distance of the main lens. Extension tubes are placed between the lens and the camera body and may be replaced by a bellows that provides variable magnification. The bellows system offers continuous magnification, an option to attach different lenses to the bellows to achieve different results, and the ability to use a reversing ring that reverses the lens position so that the front element faces toward the camera. As for lenses that use an internal floating element/group to achieve increased magnification, the bellows attachment makes good use of TTL metering for optimum exposure control.

Autoexposure

Automated systems for exposure and focusing are the hallmarks of modern camera systems. Photographic exposure is defined as $H = E \times T$, where H is meter candle seconds (log10), E is illuminance in meter candles, and T is time in seconds. Autoexposure systems are designed to determine the optimum H, range of apertures, and choice of shutter speed for a given film speed. When the aperture is predetermined by the photographer, the camera's autoexposure system will select the appropriate shutter speed. When the image at a selected speed may show camera shake, a warning signal may occur, or a flash is activated in those cameras that incorporate built-in-flash. This aperture priority system is found in many autoexposure cameras. The nature of autoexposure depends on the calculated relationship between the luminance of the subject and the sensitivity of the film. Film sensitivity is defined by the International Standards Organization (ISO). ISO has two parts to its designation, or Deutsche Industrie Normale (DIN) and Arithmetic (ASA). Both designations are used to represent the same minimum exposure necessary to produce a density level above the base + fog of the developed film for a given range of exposure. The relationship between the two components is described by the formula $\log(\text{ASA}) \times 10 + 1$. Thus an ASA of 100 is also a DIN of 21, $[\log(100) = 2 \times 10 + 1 = 21]$. The most advanced autoexposure systems measure the subject luminances passed by the lens aperture and determine the shutter speed. Conversely, a shutter speed may be set, and the aperture of the camera would be automatically determined. Illuminance measured by a photometer located within the light path projected by the lens. Because focusing and viewing are done through a wide open aperture and metering is for a working aperture, a stop-down method is required, if a full aperture method is not offered. Off-the-film metering is a very useful stop-down approach. A number of devices from secondary reflex mirrors, beam splitters, prisms, multiple (segmented) photocells, and specially designed reflex mirrors allow a photocell to measure the light that reaches the film plane. The ideal location of the cell for measurement is the film plane.

Photocell choices for measurement have specific advantages and certain disadvantages. For example, a selenium photocell does not require a battery, is slow, and has low sensitivity, but its spectral sensitivity matches that of the eye. A gallium arsenic phosphide cell requires a battery and an amplifier but is fast and very sensitive to low light. Its spectral sensitivity is limited to the visual spectrum. The calculations of exposure depend on the assumption that the luminance transmitted by the lens is 18%, the integrated value of an average scene. This is not always the case. Consequently, in camera metering systems apply certain patterns that vary the nature of the calculation for exposure. The patterns are selectable and can cover the range from a 3° spot to an overall weighted average.

These metering approaches are found in miniature and medium format cameras; however, view cameras can use direct measurement by a special fiber optic or other types of probes directly on the ground glass. Without such a device, large format camera users must resort to handheld meters. Unique to the handheld meter is the incident or illuminance meter. An integrating, hemispheric diffuser covers the photocell. The meter is

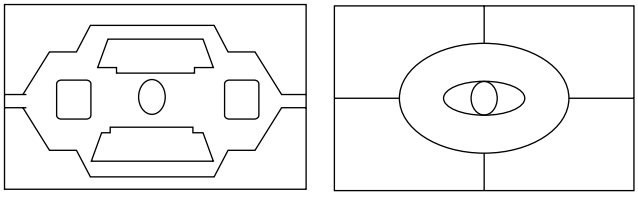


Figure 13. Illustration of metering patterns.

held at the subject and aimed at the camera. It is assumed that the subject is not very dark or light. The meter is designed on the assumption that the subject approximates a normal range of tones. When the subject is very dark or very light, exposure must be adjusted by one-half to one stop. The location of the light source is also of importance. Studio lighting close to the subject requires that the photographer compensate for any loss of luminance that results from a increase in the subject to light source distance, following the inverse square law. Handheld meters do not compensate for other photographic exposure influences such as reciprocity law failure or the use of filters for contrast or color control; nor can handheld illuminance meters be used for emitting sources. The relationship among the various factors, film speed, shutter speed, f#, and illuminance is expressed in the formula, foot candles = 25 (a constant) × $f\#^2$/(arithmetic film speed × shutter speed). It is obvious that a measurement of the illuminance in foot candles can be used in the previous formula to solve for $f\#$ or shutter speed.

Autofocusing

Coupled with autoexposure systems are advanced autofocusing systems. Autofocusing can be applied to a number of photographic systems. Likewise, a number of approaches can be used for autofocusing. Phase detection, sound ranging, and image contrast comparisons have been used in different camera systems. Electromechanical coupling racks the lens forward or rearward for correct exposure. By using a linear array containing up to 900 photosites, adjacent photosites on the array are compared. Distance may be measured by correlation based on the angle subtended between the two zones. Multiple zone focusing in which a number of fixed zones are preset for focusing distances from infinity to less than 2 ft are found in a number of current prosumer digital cameras and nonprofessional miniature cameras. Other similar camera systems offer infrared (IR) focusing. IR focusing involves scanning an IR beam emitted through the lens by using a beam splitter. The return IR reflection is read by a separate photo array through a nearby lens. This array sets the focus. IR beam-splitting focusing fails for subjects that have very low or very high IR reflectance. An IR lock can be set when one is photographing through glass. If the camera is equipped with autozooming, the IR diode detector can drive the zoom and maintain constant focus at a given magnification. The use of CCD arrays, photodiodes, and electronic focusing controls (micromotors) is made possible by incorporating of high-quality, miniaturized analog–digital circuitry and attendant CPUs and ROM chips. Such advanced technology permits focusing by measuring image contrast or phase difference. The state-of-the-art Nikon F-5 and its digital equivalent use a phase detection system. This system has a specifically designed array. Phase or image shift is the measured illuminance exiting from the pupil in two distinct zones. Two images are projected to the focal plane, and their displacement is measured. This is very much like a rangefinder system, but instead of a visual split image being seen by the photographer, a digital equivalent is detected by the CCD array (Fig. 14).

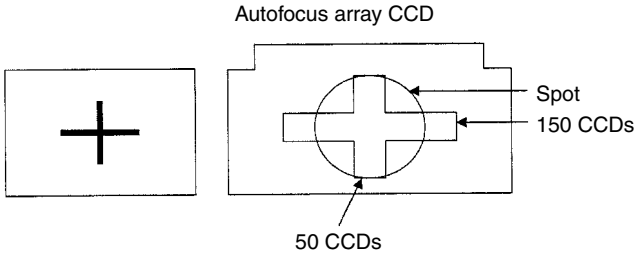


Figure 14. Nikon autofocus arrays.

CAMERA FILM

A light-sensitive material that upon exposure creates a latent image whose susceptibility to development (amplification) is proportional to the exposure received. Camera film is made of a colloidal suspension commonly referred to as an emulsion. A polyester, acetate base, or other substrate is coated with a suspension of a compound of silver and one or more halides, as well as other addenda such as spectral sensitizing dyes. A film may have several different emulsions coated on it. The difference in emulsion gives rise to the characteristics of the film. The pictorial contrast of a film is the result of coating the substrate with a number of emulsions that have varying sizes of silver halide grains. Silver halides, denoted AgX, can be made of silver (Ag) and any or all of the three halides bromide, iodine, and chloride. The combination of three halides extends the spectral response of the film beyond the film's inherent UV–blue sensitivity. The other two halides, astatine and fluoride, are not used because of either radioactivity or water solubility. The size of the grain, a microscopic speck of AgX, is the primary determinant of speed or the sensitivity of the film: the larger the grain, the greater the response to a photon. Other addenda are added to increase the speed of the film further. Photographic speed (black-and-white) is determined by the formula ASA = $1/H_m \times 0.8$, where H_M is the exposure in lux-seconds that produces a density of 0.1 above base + fog and 0.8 is a safety factor to guard against underexposure. Overall, films exhibit a variety of properties such as speed, spectral sensitivity, exposure latitude, and a unique characteristic curve response (nonlinear) to exposure and development combinations. The overall film's image quality is referred to as definition. Definition is comprised of three components: resolution, graininess, and sharpness. Black-and-white films (panchromatic) that are used in pictorial photography can be made of three to nine emulsion layers, have an exposure latitude of 1,000 : 1, and a spectral sensitivity from UV to deep red. Black-and-white pan films are sensitive to all visible wavelengths as various tones of gray. They record color information from the object in terms of luminance and chromaticity. Pan films can be designed to limit their spectral response only to the UV and IR portions of the spectrum (Fig. 15).

Color films are essentially monochrome emulsions that record the blue, green, and red record of a object or scene on three discrete emulsions that are superimposed, a "tripack." Development causes these layers to hold the

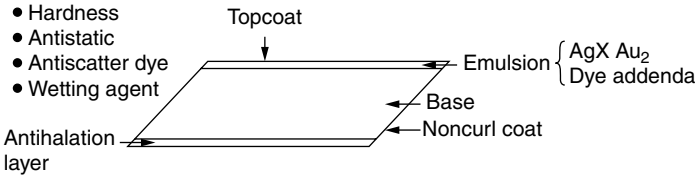


Figure 15. Cross section of panchromatic film.

image as layers of yellow, magenta, and cyan dyes from which a positive or full color print may be made. Negative color film requires reexposing the negative to a similar negative emulsion coating on paper. This negative-to-negative process produces a full color positive image. If the film selected for exposure is transparent, then the image captured after processing will be positive, that is, the colors of the image will be the same as the colors of the object.

Density

The most useful approach to determining the effect of exposure and development on a given film is measuring the film's density. The nature of film and light produce three interactive effects during exposure: scattering, reflection, and absorption. Film density is the result of these three exposure outcomes, and it is defined as $-\log(1/T)$. T is transmittance, the ratio of transmitted light to incident light. Density can be measured and provides a logarithmic value. The relationship between density, as an outcome of photographic exposure and development, is graphically expressed in the $D-\log H$ curve. The curve typically illustrates four zones of interest to the photographer. The base + fog region, the toe, the straight line, and the shoulder region are the four sections that provide useful information about exposure and development. The aim of exposure is to produce a shadow detail of the subject as a density in the negative that has a value of 0.1 above the density of the film's base+fog. The midtone and highlight reflectances of the scene follow the placement of the shadow on the curve. This correct placement or exposure permits proper transfer of these densities in the negative to paper during the printing process. Because a film's density range can greatly exceed the capacity of paper's density range, the correct placement of shadow (exact exposure) is the fundamental first step in tone reproduction. Underexposure results in a lower than useful density in the shadow detail. Graphically, this would place the shadow density below 0.1 and possibly into the base+fog of the film. Consequently, no tonal detail would be captured. In a severe overexposure, the shadow detail would be placed further to the right, up the curve. Though shadow detail would still exist further up the curve and greater in density, it is quite likely that the highlights would move to the shoulder zone. Loss of highlight detail would result. The permissible error range between the shadow–highlight shift is known as exposure latitude. Consider that a subject's luminance range is 1 : 160. Its log luminance is 2.2. If the film-developer's exposure range is log 2.8, the difference of 0.6 is the exposure latitude or two stops. This margin for adjusting exposure or for error is not equidistant on the curve. The two stops of adjustment are primarily in favor of overexposure because the distance up the curve is greater than that toward the base+fog region. The exposure latitude of negative films is much greater than that of transparent film. The characteristic curve that illustrates these relationships is not fixed but can exhibit different slopes or measures of contrast, depending on development factors such as time, agitation, dilution, and type of developer. The sensitometric studies and densitometric plottings graphically illustrate for the photographer the possible outcomes of film and developer combinations and their exposure latitude.

MEASURES OF IMAGE QUALITY

The overall measure of image quality is termed definition. It is comprised of three major components: graininess, sharpness, and resolution. The transfer of tonal information from the original scene or subject to the negative and through to the final print is of great importance, but the lack of definition in whole or part can contribute to an overall bad image. Sharpness and resolution are attributes of the optical system and the detector. Graininess (noise) is a characteristic of the detector. The quality of an edge can be described as a subjective evaluation of sharpness. When measured on a microlevel as a change of density across an edge, it is known as acutance. The rate of change of density across the edge, or slope, determines the image's appearance of sharpness. Many factors can influence sharpness. Imprecise focusing, low contrast, poor optics, camera vibrations, and bad developing techniques can result in loss of sharpness.

After exposure, silver halides are transformed during development into silver grains whose structure and size change. The increase in size causes the grains to overlap and clump into an irregular pattern that is detectable at higher magnifications, such as enlargement printing. This noticeable pattern is referred to as graininess. It is not objectionable in its own right, but it can obfuscate detail. As density increases in the negative, the perception of graininess decreases. Graininess is most visible in the midtone region. It is inherent in the material and cannot be eliminated simply. Options to minimize graininess are to use film sizes that require minimum magnification and print on lower contrast or matte paper. A microdensitometer can measure graininess and provide a trace of density fluctuations across distance. This granularity is considered a measure of standard deviation around the mean or average density measured. Because the standard deviation is the root mean square, this measure is known as rms granularity. Manufacturers' measures of rms correlate well with perceptional graininess, but these measures of granularity do not correlate well among various manufacturers.

Resolution is the ability of the optical and detection systems to reproduce fine detail. All of the components in the imaging system combine to produce an overall measure of resolution known as resolving power. Resolving power is expressed as $1/R_S = 1/R_L + 1/R_D$. Greater accuracy can be achieved by taking the second moment, that is, $1/(R_S)^2 = 1/(R_L)^2 + 1/(R_D)^2$. Every component in the system contributes to the overall measure, and this measure cannot be higher than the lowest component. Overall photographic definition describes the total image and can consist of many isolated measures that affect the image. Such measures are the point-spread function, as indicated by the size of the Airy disk or diffraction effects, the emulsion spread function of a particular film, and the line spread function that measures the ability of the image to separate adjacent lines in the image. It would be onerous for the photographer to collect various measures and attempt to correlate them. There is an overall measure made available by manufacturers that eliminates such a task. The modulation transfer function (MTF) represents the overall contrast transfer of the object to the image (Fig. 16). If the contrast of the object were to be matched totally by the image, the transfer would be 100%. All detail or frequencies of the object would be maintained at a 1 : 1 contrast ratio regardless of a change in the finest details or frequencies.

Modulation is determined as $M_O = E_{max} - E_{min}/(E_{max} + E_{min})$ and $M_I = E_{max} - E_{min}/(E_{max} + E_{min})$, therefore, M_{image}/M_{object}. Individual MTF measures for various imaging components can be multiplied to produce one MTF factor for the system.

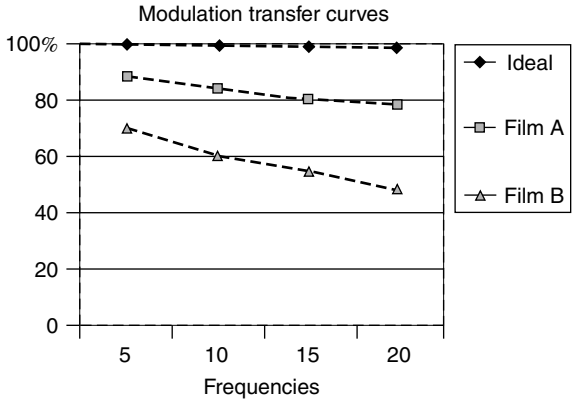


Figure 16. Modulation transfer function.

DIGITAL PHOTOGRAPHY

The application of CCD or CMOS detectors in place of film at the plane of focus has quickly changed photography. As the resolving power of the detectors has improved and the inherent firmware in the digital camera/back has improved its algorithms for image reconstruction, the usefulness and availability of digital hardware has increased as well. The basic principles of image formation, lens types, and image quality also hold true for digital imaging. The creation of binary image data that are easily manipulated by computer-based software can take enormous advantages of digital pictures. Postexposure photographic data can be eliminated, improved, edited, or added. Images may be sent directly to video monitors, satellites, and remote sites or may be printed on hard copy via a number of devices that do not require any darkroom or projection device. Ink-jet printers, thermal dye imagers, and other devices can produce images that are virtually indistinguishable from traditional photographic images. The crossover to digital from analog in the professional arena began with the photojournalist and soon extended into catalog photography. This was driven by improved quality, ease of application, and cost effectiveness compared to film. Publication (matrix reproduction) of images in magazines, newspapers, journals, and other media have become more digital than analog. Hybrid approaches that use film as the capture medium and scanners that convert the image to digital data have almost totally ended the use of the process camera in the printing industry. Large format scanning backs are readily available for the view camera. Medium format camera manufacturers provide digital back options for all major systems. Surveys of digital camera sales at the consumer and professional level show a steadily upward trend. The indications are that digital photography will not disappear and may be the preferred method of basic still photography.

BIBLIOGRAPHY

1. C. R. Arnold, P. J. Rolls, and J. C. J. Stuart, in D. A. Spencer, ed., *Applied Photography*, Focal Press, London, 1971.
2. M. J. Langford, *Advanced Photography*, Focal Press, London, 1972.
3. S. Ray, *Applied Photographic Optics*, Focal Press, Boston, 1994.
4. S. Ray, *Camera Systems*, Focal Press, Boston, 1983.
5. L. Stroebel, J. Compton, I. Current, and R. Zakia, *Basic Photographic Materials and Processes*, 2 ed., Focal Press, Boston, 1998.
6. L. Stroebel, *View Camera Techniques*, Focal Press, Boston, 1992.
7. *The Encyclopedia of Photography*, Eastman Kodak Co., Rochester, 1981.

T

TELEVISION BROADCAST TRANSMISSION STANDARDS

ALAN S. GODGER
JAMES R. REDFORD
ABC Engineering Lab 30 W/7
New York, NY

Since the invention of television, the images and sound have been captured, processed, transmitted, received, and displayed using analog technology, where the picture and sound elements are represented by signals that are proportional to the image amplitude and sound volume.

More recently, as solid-state technology has developed, spurred primarily by the development of computers, digital technology has gradually been introduced into handling the television signal, both for image and sound. The digital electric signal representing the various elements of the image and sound is composed of binary numbers that represent the image intensity, color, and so on, and the sound characteristics.

Many portions of television systems are now hybrid combinations of analog and digital, and it is expected that eventually all television equipment will be fully digital, except for the transducers, cameras, and microphones (whose inputs are analog) and the television displays and loudspeakers (whose outputs are analog).

The currently used broadcast television transmission standards [National Television Systems Committee (NTSC), phase alternate line (PAL) and sequential and memory (SECAM)] for 525- and 625-line systems were designed around analog technology, and although significant portions of those broadcast systems are now hybrid analog/digital or digital, the "over the air" transmission system is still analog. Furthermore, other than for "component" processed portions of the system, the video signals take the same "encoded" form from studio camera to receiver and conform to the same standard.

The recently developed ATSC Digital Television Standard, however, uses digital technology for "over the air" transmission, and the digital signals used from the studio camera to the receiver represent the same image and sound, but differ in form in portions of the transmission system. This variation is such that in the studio, maximum image and sound information is coded digitally, but during recording, special effects processing, distribution around a broadcast facility, and transmission, the digital signal is "compressed" to an increasing extent as it approaches its final destination at the home. This permits practical and economical handling of the signal.

ANALOG TELEVISION SYSTEMS

Black-and-White Television

The purpose of all conventional broadcast television systems is to provide instantaneous vision beyond human sight, a window into which the viewer may peer to see activity at another place. Not surprisingly, all of the modern systems evolved to have similar characteristics. Basically, a sampling structure is used to convert a three-dimensional image (horizontal, vertical, and temporal variations) into a continuous time-varying broadband electrical signal. This modulates a high-frequency carrier with the accompanying sound, and it is broadcast over the airwaves. Reasonably inexpensive consumer television sets recover the picture and sound in the viewer's home.

Image Representation. The sampling structure first divides the motion into a series of still pictures to be sequenced rapidly enough to restore an illusion of movement. Next, each individual picture is divided vertically into sufficient segments so that enough definition can be retrieved in this dimension at the receiver. This process is called *scanning*. The individual pictures generated are known as *frames*; each contains *scanning lines* from top to bottom.

The number of scanning lines necessary was derived from typical room dimensions and practical display size. Based on the acuity of human vision, a viewing distance of four to six picture heights is intended. The scanning lines must be capable of enough transitions to resolve comparable definition horizontally. The image *aspect ratio* (width/height) of all conventional systems is 4 : 3, from the motion picture industry "academy aperture." All systems sample the picture from the top left to bottom right.

In professional cinema, the projection rate of 48 Hz is sufficient to make flicker practically invisible. Long-distance electric power distribution networks throughout the world use slightly higher rates of 50–60 Hz alternating current. To minimize the movement of vertical "hum" in the picture caused by marginal filtering in direct current power supplies, the picture repetition rate was made equal to the power line frequency.

A variation of this process used by all conventional systems is *interlaced scanning*, whereby every other line is scanned to produce a picture of half the vertical resolution, known as a *field*. The following field "fills in" the missing lines to form the complete frame. Each field illuminates a sufficient portion of the display so that flicker is practically invisible, yet only half the information is being generated. This conserves bandwidth in transmission. For both fields to start and stop at the same point vertically, one field must have a half scanning line at the top, and the other field must have a half scanning line at the bottom of the picture. This results in an odd number of scanning lines for the entire frame.

Mechanical systems using rotating disks that have spiral holes to scan the image were investigated in the 1920s and 1930s, but these efforts gave way to "all electronic" television. Prior to World War II, developers in the United States experimented with 343-line and 441-line systems. Developers in Great Britain began a 405-line service, and after the war, the French developed an 819-line system, but these are no longer in use.

Synchronization. In most of North and South America and the Far East, where the power line frequency is 60 Hz, a 525-line system became the norm. This results in an interlaced scanning line rate of 15.750 kHz. The development of color television in Europe led to standardization of 625 lines in much of the rest of the world. The resulting line frequency of a 50 Hz field rate is 15.625 kHz. The similar line and field rates enable the use of similar picture tube deflection circuitry and components. Horizontal and vertical frequencies must be synchronous and phase-locked, so they are derived from a common oscillator.

Synchronization pulses are inserted between each scanning line (Fig. 1) and between each field to enable the television receiver to present the picture details that have the same spatial orientation as that of the camera. The *sync* pulses are of opposite polarity from the picture information, permitting easy differentiation in the receiver. The line sync pulses, occurring at a faster rate, are narrower than the field sync pulses, which typically are the duration of several lines. Sync separation circuitry in the receiver discriminates between the two time constants. Sync pulses cause the scanning to *retrace* rapidly from right to left and from bottom to top.

Blanking. To provide time for the scanning circuits to reposition and stabilize at the start of a line or field, the picture signal is *blanked*, or turned off. This occurs just before (front porch) and for a short time after (back porch) the horizontal sync pulse, as well as for several lines before and after vertical sync. During vertical sync, *serrations* are inserted to maintain horizontal synchronization. Shorter *equalizing pulses* are added in the several blanked lines before and after vertical sync (Fig. 2). All of these pulses occur at twice the rate of normal sync pulses, so that the vertical interval of both fields (which are offset by one-half line) can be identical, simplifying circuit design.

Additional scanning lines are blanked before the active picture begins; typically there is a total of 25 blanked lines per field in 625-line systems and 21 blanked lines per field for the 525-line system M. Modern television receivers complete the vertical retrace very soon after the vertical sync pulse is received. The extra blanked lines now contain various ancillary signals, such as for short-time and line-time distortion and noise measurement, ghost cancellation, source identification, closed captioning, and teletext.

Fields and lines of each frame are numbered for technical convenience. In the 625-line systems, field 1 is that which begins the active picture with a half line of video. In the 525-line system M, the active picture of field 1 begins with a full line of video. Lines are numbered sequentially throughout the frame, beginning at the vertical sync pulse in the 625-line systems. For the 525-line system M, the line numbering begins at the first complete line of blanking for each field. Field 1 continues halfway through line 263, at which point field 2 begins, containing through line 262.

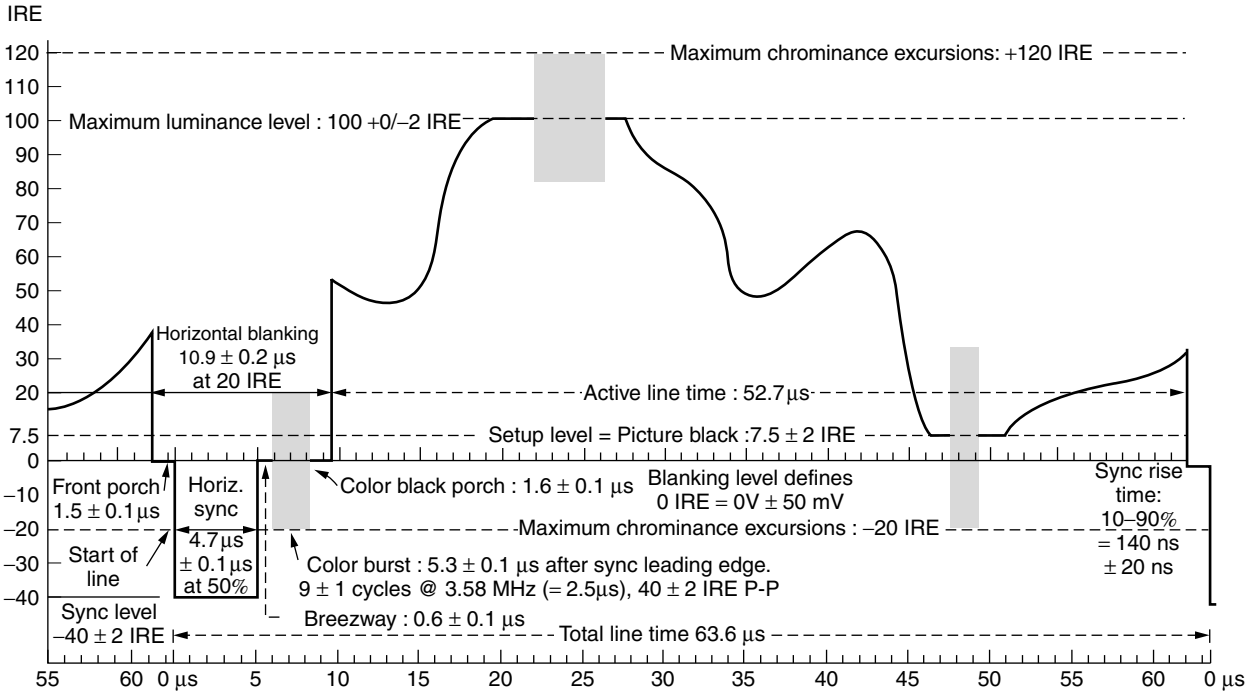


Figure 1. The 525-line system M: Line-time signal specifications.

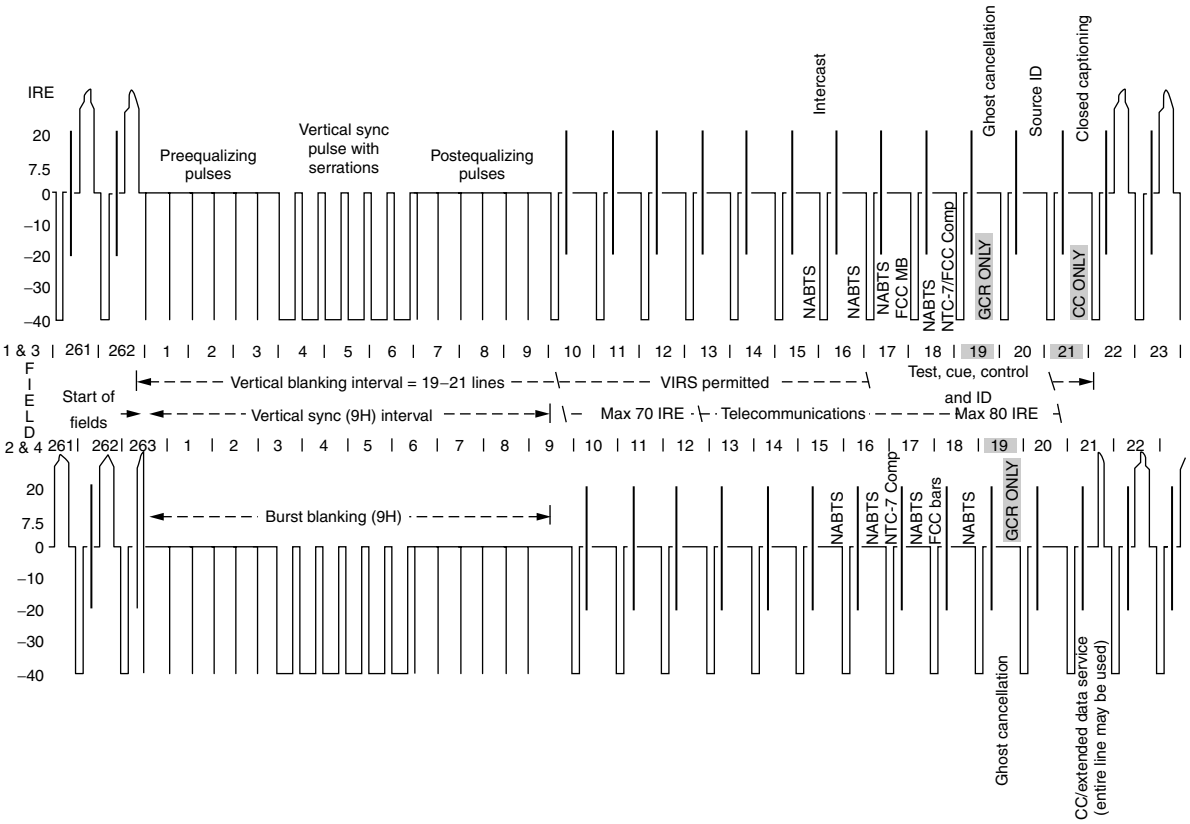


Figure 2. The 525-line system M: Field-time and vertical interval signal specifications.

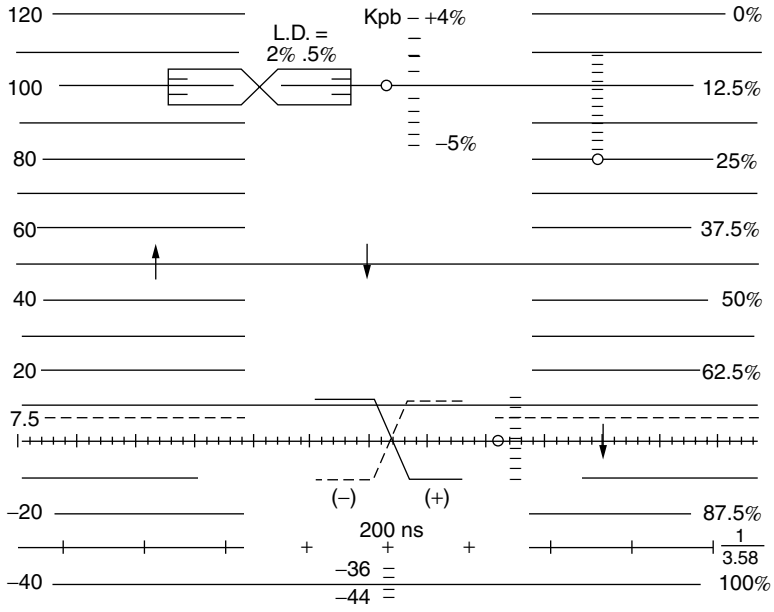


Figure 3. Typical NTSC-M waveform monitor graticule. A number of additional markings are for measuring various signal distortions.

Signal Levels. During blanking, the video signal is at 0 V, the reference used to measure picture (positive-going) and sync (negative-going) levels. Signal amplitudes are measured directly in millivolts, except that, because of changes made during the conversion to color, the 525-line system uses *IRE units*. A specialized oscilloscope is used to monitor the characteristics of the signal amplitude and period. The *waveform monitor* has its voltage scale calibrated in millivolts (or IRE units for 525-line applications), and its time base is calibrated to scanning line and picture field rates, as well as in microseconds (Fig. 3).

Originally, the 525-line system used an amplitude of 1 V peak-to-peak (p−p) for the picture information, and it used 0.4 V for sync. So that color information (modulated onto a *subcarrier* which can extend above peak white level) could be accommodated within the same dynamic range of existing equipment, the 1.4 V p−p scaling was compressed in amplitude to 1 V p−p. This created fractional voltage levels for peak white (714.3 mV) and sync (−286.7 mV), so a 1 V scale of 140 IRE units was adopted to simplify measurement. The 625-line standards did not have this historical complication. The peak white level is 700 mV, and sync level is −300 mV.

Another anachronism of the 525-line system is the use of a direct-current (dc) offset of the picture black from the blanking level. This was done to ensure that during retrace, the electron beam in the display tube was completely cut off, so *retrace lines* did not appear in the picture. This *setup level* originally varied between 5 and 10 IRE units above blanking but was standardized at 7.5 IRE for color TV, although it has been discarded altogether in Japan. Setup, or "lift," was used to some extent in earlier systems, but abandoned by the advent of 625-line services.

The electrical-to-optical transfer characteristic (*gamma*) of the cathode-ray picture tube is nonlinear. Doubling the video signal level applied to the control grid of the picture tube does not cause the light output to double; rather, it follows a power law of approximately 2.5. To correct for this, the video signal itself is made nonlinear, and an opposite transfer characteristic is about 0.4. This correction is applied at the camera in all systems.

Resolution. Resolution in the vertical direction is determined by taking the total number of scanning lines and subtracting those used for vertical blanking. This is multiplied by 0.7, a combination of the *Kell factor*, (a correction for slight overlap between adjacent lines) and an additional correction for imperfect interlace. By convention, television resolution is expressed in *television (TV) lines per picture height*, in contrast to photographic "line pairs per millimeter." Because the resolution is fixed by the scanning system, picture size is immaterial. Note that a vertical "line pair" in television requires two scanning lines.

To compute the bandwidth necessary for equal horizontal resolution, the vertical resolution is multiplied by the aspect ratio of 4 : 3 and is divided by the ratio of total scanning line time to active picture (unblanked) line time. This number is halved because an electric cycle defines a line pair, whereas a "TV line of resolution" is really only one transition. Multiplying the number of cycles per scanning line by the total number of scanning lines in a field and then multiplying the number of fields per second gives the bandwidth of the baseband video signal.

Broadcasting Standards. The various systems have been assigned letter designations by the International Telecommunications Union (ITU). The letters were assigned as the systems were registered, so alphabetical order bears no relation to system differences (Table 1), but a rearrangement highlights similarities (Table 2). Only scanning parameters and radio-frequency (RF) characteristics are defined; color encoding is not specified. Systems A, C, E, and F are no longer used.

Portions of the very high frequency (VHF) and ultrahigh-frequency (UHF) RF spectrum are divided into *channels* for television broadcast. Modern channel assignments are 6 MHz wide in the Americas and the Far East. Elsewhere, they are generally 7 MHz in VHF and 8 MHz in UHF, and the carrier is a fixed distance from one edge. Because the picture carrier in most systems is near the lower edge and the audio signals are at the upper end, when the opposite is true, the channels are called *inverted*.

As a bandwidth-saving technique, the amplitude-modulated RF signal is filtered so that only one sideband is fully emitted; the other sideband is *vestigial*, or partially suppressed, which aids in fine-tuning to the correct carrier frequency at the receiver. The full sideband, which represents the video bandwidth, extends in the direction of the audio carrier(s), but sufficient guard band is included to prevent interference.

The bandwidth of the vestigial sideband varies among systems as does the placement of the audio carrier in relation to the picture carrier (Fig. 4). These factors complicate receiver design in areas where signals of two or more systems may exist. The main audio signal is sent via an amplitude-modulated or, more commonly, frequency-modulated carrier. Peak deviation in frequency modulation (FM) is ±50 kHz with 50 μm preemphasis, except ±25 kHz and 75 μs for systems M and N. Preemphasis for improving the signal-to-noise ratio is common in FM systems; it was used in some amplitude-modulation (AM) systems to simplify receivers that could accommodate both modulation schemes.

Amplitude modulation is used in all systems for the video waveform, which, unlike audio, is not sinusoidal. The majority of systems employ a *negative sense of modulation*, such that negative excursions of the baseband signal produce an increase in the amplitude of the modulated carrier. This allows the constant amplitude sync pulses to serve as an indication of the received RF signal level for automatic gain control. Interfering electrical energy also, tends to produce less noticeable black flashes in the received picture, and the duty cycle of the signal is reduced, which consumes less power at the transmitter.

Multichannel Sound

Early attempts to provide stereo sound for special TV events involved *simulcasting*, whereby an FM radio station in the same coverage area broadcast the stereo audio for the program. Due to the high profitability of FM radio today, this scheme is becoming impractical. For the 525-line system M that has channels of only 6 MHz bandwidth, a multiplexing scheme is used on the existing single audio carrier. Due to the wider channel bandwidths in 625-line systems, multiple sound carriers emerged as the solution.

Multiplex Sound Systems. In the United States, Zenith Electronics developed multichannel television sound (MTS), a pilot-tone system in which the sum of the

Table 1. Principal Characteristics of World Television Systems

Standard	A United Kingdom	B/G Western Europe	C Luxembourg	D/K Eastern Europe	E France	F Belgium	H Belgium	I United Kingdom	K′ French Overseas Territories	L France	L′ France	M North America/ Far East	N South America
Frequency band	VHF	V/U	VHF	V/U	VHF	VHF	UHF	V/U	V/U	V-3/U	VHF-1	V/U	V/U
Channel B/W (MHz)	5	7/8	7	8	14	7	8	8	8	8	8	6	6
Visual/Edge separation (MHz)	−1.25	+1.25	+1.25	+1.25	±2.83	+1.25	+1.25	+1.25	+1.25	+1.25	−1.25	+1.25	+1.25
Video B/W (MHz)	3	5	5	6	10	5	5	5.5	6	6	6	4.2	4.2
Vestigial sideband (MHz)	+0.75	−0.75	−0.75	−0.75	±2.0	−0.75	−1.25	−1.25	−1.25	−1.25	+1.25	−0.75	−0.75
Visual modulation polarity (Wht =)	Pos.	Neg.	Pos.	Neg.	Pos.	Pos.	Neg.	Neg.	Neg.	Pos.	Pos.	Neg.	Neg.
Picture/synchronization ratio	7/3	7/3	7/3	7/3	7/3	7/3	7/3	7/3	7/3	7/3	7/3	10/4	7/3
Black pedestal(%)	0	0	0	0	0	0	0	0	0	0	0	7.5	0
Visual/aural separation (MHz)	−3.5	+5.5	+5.5	+6.5	±11.15	+5.5	+5.5	+6.0	+6.5	+6.5	−6.5	+4.5	+4.5
Aural modulation	AM	FM	AM	FM	AM	AM	FM	FM	FM	AM	AM	FM	FM
Aural peak deviation (kHz)		±50		±50			±50	±50	±50			±25	±25
Aural preemphasis (μs)		50	50	50		50	50	50	50			75	75
Lines per frame	405	625	625	625	819	819	625	625	625	625	625	525	625
Field blanking (lines)	13	25	25	25	33	33	25	25	25	25	25	21	25
Field synchronization (lines)	4	2.5	2.5	2.5	20 μs	3.5	2.5	2.5	2.5	2.5	2.5	3	2.5
Equalization pulses (lines)	0	2.5	2.5	2.5	0	3.5	2.5	2.5	2.5	2.5	2.5	3	2.5
Vertical resolution (L/ph)	254	413	413	413	516	516	413	413	413	413	413	350	413
Horizontal resolution (L/ph)	270	400	400	450	635	400	400	430	450	450	450	320	320
Field rate (Hz)	50	50	50	50	50	50	50	50	50	50	50	59.94	50
Line rate (Hz)	10,125	15,625	15,625	15,625	20,475	20,475	15,625	15,625	15,625	15,625	15,625	15,734	15,625
Line blanking (μs)	18	12	12	12	9.5	9.2	12	12	12	12	12	10.9	12
Front porch (μs)	1.75	1.5	1.4	1.5	0.6	1.1	1.5	1.5	1.5	1.5	1.5	1.5	1.5
Line synchronization (μs)	9.0	4.7	5.0	4.7	2.5	3.6	4.7	4.7	4.7	4.7	4.7	4.7	4.7

Table 2. Survey of World Television Systems

Standard	Field	Lines	Bandwidth: Channel	Bandwidth: Video	Vestigial Sideband	Polarity	Audio	Standard	VHF	UHF	NTSC	PAL	SECAM
M	60	525						M	√	√	√	√	
N			6	4.2				N	√	√		√	
D					−0.75	Neg.	FM	D	√			√	√
K								K		√			√
L′				6.0	+1.25			L′	√				√
L			8			Pos.	AM	L	√	√			√
K′		625			−1.25			K′	√				√
I	50			5.5				I	√	√		√	
H						Neg.	FM	H		√		√	
G								G		√		√	√
B				5.0	−0.75			B	√			√	√
C			7					C	√				
F		819				Pos.	AM	F	√		Obsolute standards		
E			14	10.0	±2.0			E	√				
A		405	5	3.0	+0.75			A	√				

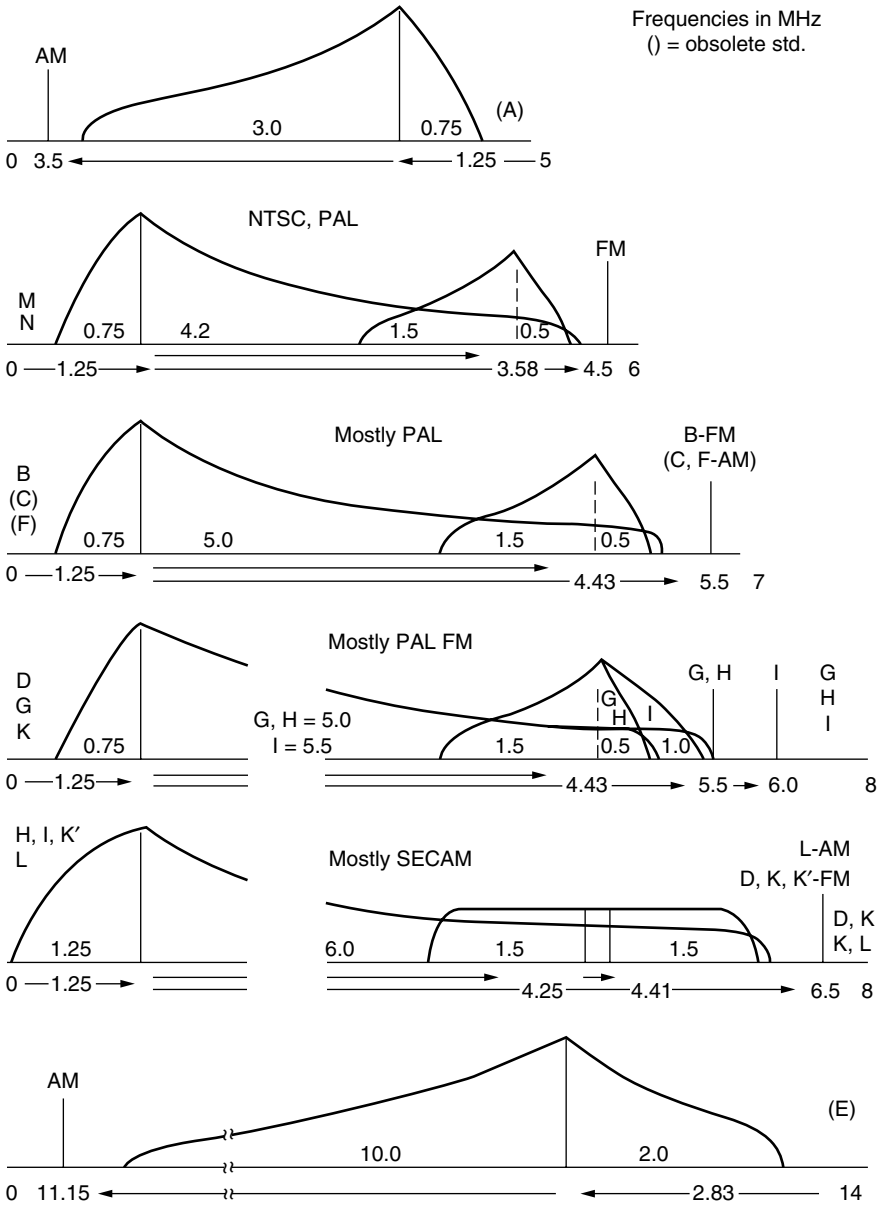


Figure 4. Television RF channel spectra for the various world television systems. For simplicity, only a single illustration of the lower and upper portions of the channel spectra is shown for the 8 MHz wide channels. Therefore, for systems D, H, I, and K, the lower and upper illustrations are not adjacent to each other.

two stereo channels (L + R) modulates the main TV audio FM carrier and provides a monophonic signal to conventional receivers. A difference signal (L − R) is dbx-TV suppressed-carrier amplitude-modulated onto an audio subcarrier at twice the line scanning frequency ($2f_H$), and a pilot is sent at the line frequency as a reference for demodulation. A second audio program (SAP) may be frequency modulated at $5f_H$, and nonpublic voice or data may be included at $6.5f_H$.

Japan's developed a similar *FM/FM* stereo system using FM of the L − R subchannel at $2f_H$. A control subcarrier at $3.5f_H$ is tone-modulated to indicate whether stereo or second audio programming is being broadcast.

Multiple Carrier Systems. In Germany, IRT introduced *Zweiton*, a dual carrier system for transmission standards B and G. In both standards, the main carrier is frequency-modulated at 5.5 MHz above the vision carrier. For stereo, this carrier conveys the sum of the two channels (L + R). The second frequency-modulated sound carrier is placed 15.5 times the line scanning frequency above the main carrier, or 5.742 MHz above the vision carrier. For transmission system D, a similar relationship exists to the 6.0 MHz main channel. For stereo, this carrier conveys only the right audio channel.

A reference carrier at $3f_H$ is tone-modulated, and the particular tone frequency switches receivers for stereo or second audio channel sound. A variant is used in Korea where the second carrier is placed at 4.72 MHz above the vision carrier and conveys L − R information.

The British Broadcasting Corporation (BBC) in Britain developed near-instantaneous companded audio multiplex (NICAM), a digital sound carrier system. Both audio channels are sampled at 32 kHz at 14-bit resolution. Each sample is compressed to 10 bits, then arranged into frame packages of 728-bit length. These are rearranged and then the data are scrambled to disperse the effect of noise at the receiver. Finally, two bits at a time are fed to a QPSK modulator for transmission.

Either stereo or two independent sound signals may be transmitted. Other possible combinations are (1) one sound and one data channel or (2) two data channels. The original analog-modulated single-channel sound carrier is retained for older receivers. The digital carrier is 6.552 MHz above the vision carrier for transmission system I or 5.85 MHz for systems B, G, and L.

NTSC Color Television

It has long been recognized that color vision in humans results from three types of photoreceptors in the eye, each sensitive to a different portion of the visible spectrum. The ratio of excitation creates the perception of hue and saturation, and the aggregate evokes the sensation of brightness. Stimulating the three types of photoreceptors using three wavelengths of light can produce the impression of a wide gamut of colors. For television, the image is optically divided into three primary color images, and this information is delivered to the receiver, which spatially integrates the three pictures, something like tripack color film and printing.

Sequential Color Systems. The first attempts at commercial color TV involved transmitting the three color images *sequentially*. Compatibility with existing transmitters was essential from an economic standpoint. Because line-sequential transmission caused crawl patterning, field-sequential was preferred. However, there was no separate *luminance* signal for existing black-and-white receivers to use.

To reduce flicker caused by the apparent brightness differences among the three primary colors, the field rate had to be increased, and maintaining an equivalent channel bandwidth required that the number of scanning lines be reduced. These changes aggravated the compatibility problem with existing sets. A field-sequential system developed by the Columbia Broadcasting System (CBS) network was briefly commissioned in the United States during 1951.

To be *compatible*, a color television system must have the same channel bandwidth as existing monochrome transmitters and receivers, use equivalent scanning parameters, and supply the same luminance signal, as if the picture were black-and-white. An all-industry body, the National Television Systems Committee (NTSC), was set up in the United States to devise such a color TV system.

Separate Luminance and Mixed Highs. The human visual system senses shapes and edges from brightness variations. Color fills in only the larger areas, much like a child's coloring book. At the suggestion of Hazeltine Electronics, the existing wide-bandwidth luminance signal of black-and-white television was retained. The color information is limited to a much narrower bandwidth of the order of 1 MHz and restricts its resolution in the horizontal direction.

This first led to a dot-sequential system that sampled the three colors many times along each scanning line to form a high-frequency chrominance signal. The frequency of sampling may be likened to a *subcarrier* signal whose amplitude and phase are changing according to color variations along the line. At the receiver, the "dot" patterns of each primary resulting from sampling are averaged in low-pass frequency filters. The result is a continuous but low-resolution full-color signal. Equal amounts of their higher frequency components are summed to form a *mixed-highs* signal for fine luminance detail ($Y = 1/3R + 1/3G + 1/3B$), an idea from Philco.

Quadrature Modulation. The dot-sequential concept formed the basis for a more sophisticated simultaneous system. The luminance signal contains both high- and low-frequency components. Only two lower resolution color signals are needed (the third can be derived by subtracting their sum from the low-frequency portion of luminance). The spectral composition of green is nearest to that of luminance, so transmitting the red and blue signals improves the signal-to-noise performance. These low-frequency color signals are sampled by using a time-multiplexing technique proposed by Philco, known as *quadrature modulation*.

The chrominance signal is formed by the addition of two subcarriers, which are locked at the same frequency

but differ in phase by 90°. The two subcarriers are modulated by separate baseband signals such that each is sampled when the other carrier is at a null. This results in modulating the subcarrier in both amplitude and phase. The amplitude relates to the saturation of the color, and the phase component corresponds to the hue (Fig. 5).

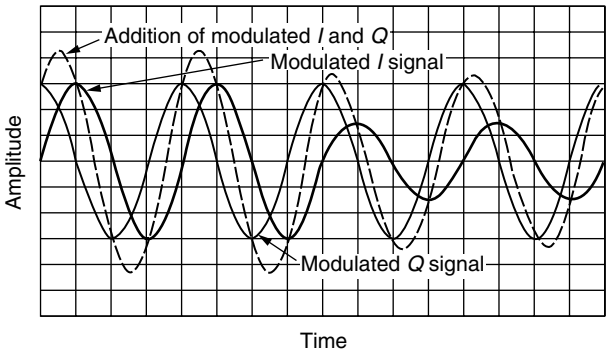


Figure 5. Quadrature modulation. Addition of two amplitude-modulated signals whose carrier frequencies are in phase quadrature (same frequency but offset in phase by 90°) produces an output signal whose carrier is modulated in both amplitude (AM) and phase (PM) simultaneously. This method of combining two baseband signals onto a single carrier is called quadrature modulation. In the case of NTSC or PAL encoding for color television, the two baseband components represent the two chrominance signals (I and Q for NTSC, U and V for PAL). The resulting amplitude of the subcarrier relates to the saturation, and the phase conveys the hue information. The frequency of the subcarrier is unchanged.

Frequency Multiplexing. The sampling rate is more than twice the highest frequency of the color signals after low-pass filtering, so the chrominance information shares the upper part of the video spectrum with luminance. This frequency-multiplexing scheme was put forward by General Electric. The scanning process involves sampling the image at line and field rates; therefore, energy in the video signal is concentrated at intervals of the line and field frequencies. These *sidebands* leave pockets between them where very little energy exists.

The exact subcarrier frequency was made an odd multiple of one-half the line scanning frequency. This causes sidebands containing the color information to fall likewise in between those of the existing luminance signal (Fig. 6). Therefore, the phase of the subcarrier signal is opposite line-to-line. This prevents the positive and negative excursions of the subcarrier from lining up vertically in the picture, and it results in a less objectionable "dot" interference pattern between the subcarrier and luminance signal. Comb filtering to separate luminance and chrominance may be employed by examining the phase of information around the subcarrier frequency on adjacent lines. The dot pattern is further concealed because the subcarrier phase is also opposite frame-to-frame.

A *four-field sequence* is established whereby the two interlaced picture fields, together with the alternating phase of subcarrier on sequential frames, requires maintaining the proper sequence. Sources to be intercut or mixed must be properly timed, and editing points must be chosen to preserve the sequence of the four color fields.

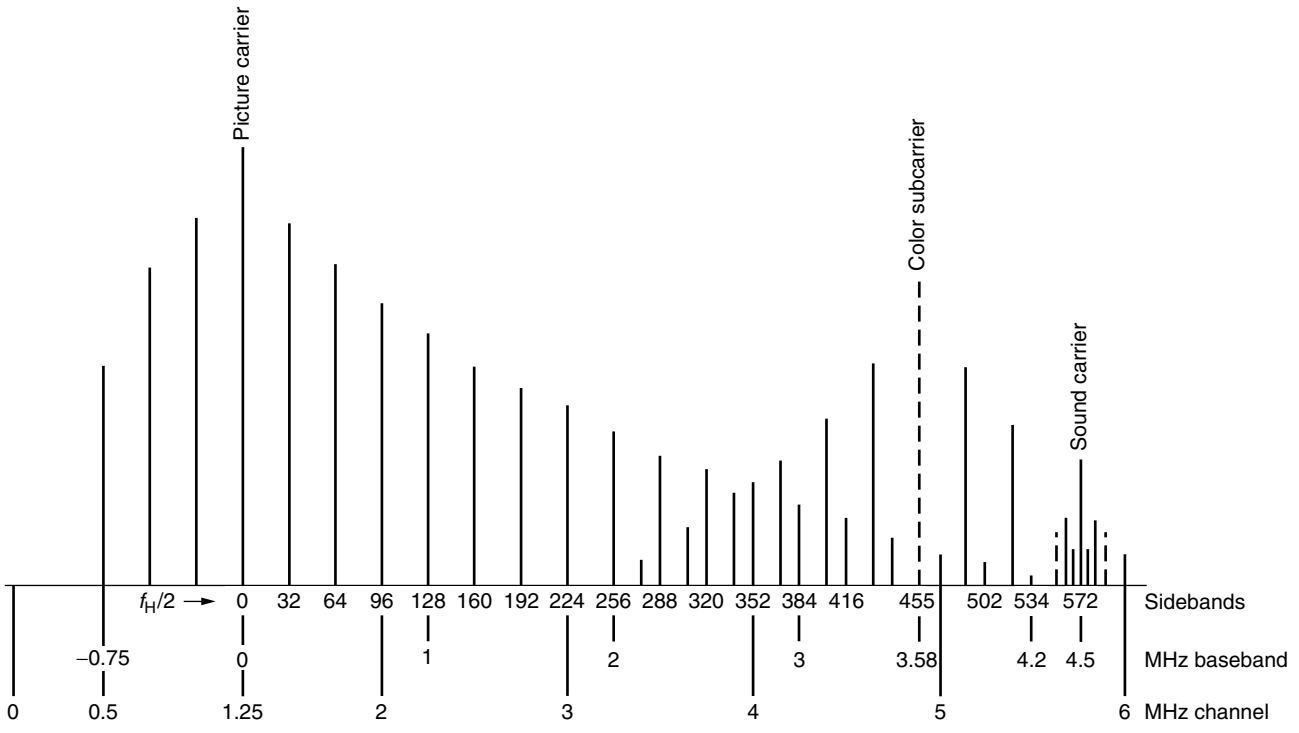


Figure 6. Frequency spectrum of composite NTSC-M color television signal showing relationships between the baseband and channel spectra and between sidebands of the picture carrier and color subcarrier.

A slight modification of the line and field scanning frequencies was necessary because one of the sidebands of the new color subcarrier fell right at 4.5 MHz, the rest frequency of the FM sound carrier for system M. Existing black-and-white receivers did not have adequate filtering to prevent an annoying buzz when the program sound was low and color saturation was high. By reducing the scanning frequencies by a mere 0.1%, the sidebands of luminance and chrominance remained interleaved, but shifted to eliminate the problem. Hence, the field frequency became 59.94 Hz, and the line frequency became 15.734 kHz.

Color-Difference Signals. Another suggestion came from Philco: Interference with the luminance signal is minimized by forming the two color signals as the difference between their respective primary and luminance (i.e., $R-Y$, $B-Y$). This makes the color-difference signals smaller in amplitude because most scenes have predominantly pale colors.

The subcarrier itself is suppressed, so that only the sidebands are formed. When there is no color in the picture, the subcarrier vanishes. This necessitates a local oscillator at the receiver. A *color-burst* reference is inserted on the back porch of the horizontal sync pulse that synchronizes the reference oscillator and provides an amplitude reference for color saturation automatic gain control.

Constant Luminance. In the constant-amplitude formulation ($Y = 1/3R + 1/3G + 1/3B$), the luminance signal does not represent the exact scene brightness. Part of the brightness information is carried by the chrominance channels, so unwanted irregularities in them, such as noise and interference, produce brightness variations. Also, the gray-scale rendition of a color broadcast on a black-and-white receiver is not correct.

Hazeltine Electronics suggested weighting the contributions of the primaries to the luminance signal according to their actual addition to the displayed brightness. The color-difference signals will then represent only variations in hue and saturation because they are "minus" the true brightness ($R-Y$, $B-Y$). A design based on this principle is called a *constant-luminance* system. For the display phosphors and white point originally specified, the luminance composition is $Y = 30\%R + 59\%G + 11\%B$.

Scaling Factors. The two low-bandwidth color-difference signals modulate a relatively high-frequency subcarrier superimposed onto the signal level representing luminance. However, the peak subcarrier excursions for some hues could reach far beyond the original black-and-white limits, where the complete picture signal is restricted between levels representing blanking and peak white picture information. Overmodulation at the transmitter may produce periodic suppression of the RF carrier and/or interference with the synchronizing signals. If the overall amplitude of the composite (luminance level plus superimposed subcarrier amplitude) signal were simply lowered, the effective power of the transmitted signal would be significantly reduced.

A better solution was to reduce the overall amplitude of only the modulated subcarrier signal. However, such an arbitrary reduction would severely impair the signal-to-noise ratio of the chrominance information. The best solution proved to be selective reduction of each of the baseband $R-Y$ and $B-Y$ signal amplitudes to restrict the resulting modulated subcarrier excursions to $\pm 4/3$ of the luminance signal levels. The $R-Y$ signal is divided by 1.14, and $B-Y$ is divided by 2.03. It was found that the resulting 33.3% overmodulation beyond both peak white and blanking levels was an acceptable compromise because the incidence of highly saturated colors is slight (Fig. 7).

Proportioned Bandwidths. RCA proposed shifting the axes of modulation from $R-Y$, $B-Y$ to conform to the greater and lesser acuity of human vision for certain colors. The new coordinates, called I and Q, are along the orange/cyan and purple/yellow-green axes. This was done so that the bandwidths of the two color signals could be proportioned to minimize cross talk (Fig. 8).

Early receivers used the wider bandwidth of the I signal; however, it became evident that a very acceptable color picture could be reproduced when the I bandwidth is restricted to the same as that of the Q channel. Virtually all NTSC receivers now employ "narrowband" I channel decoding. A block diagram of NTSC color encoding is shown in Fig. 9. These recommendations were adopted by the U.S. Federal Communications Commission in late 1953, and commercial color broadcasting began in early 1954.

Sequential and Memory (SECAM)

Economic devastation of World War II delayed the introduction of color television to Europe and other regions. Because differences between 525- and 625-line scanning standards made video tapes incompatible anyway and satellite transmission was unheard of, there seemed little reason not to explore possible improvements to the NTSC process.

Sequential Frequency Modulation. The most tenuous characteristic of NTSC proved to be its sensitivity to distortion of the phase component of the modulated subcarrier. Because the phase component imparts color hue information, errors are quite noticeable, especially in skin tones. Also of concern were variations in the subcarrier amplitude, which affect color saturation. Most long-distance transmission circuits in Europe did not have the phase and gain linearity to cope with the added color subcarrier requirements.

A solution to these drawbacks was devised by the Campagnie Française de Tèlèvision in Paris. By employing a one-line delay in the receiver, quadrature modulation of the subcarrier could be discarded, and the color-difference signals (called D_R and D_B in SECAM) sent sequentially on alternate lines. This reduces vertical resolution in color by half; however, it is sufficient to provide only coarse detail vertically, as is already the case horizontally.

In early development, AM was contemplated; however, the use of FM also eliminated the effects of subcarrier

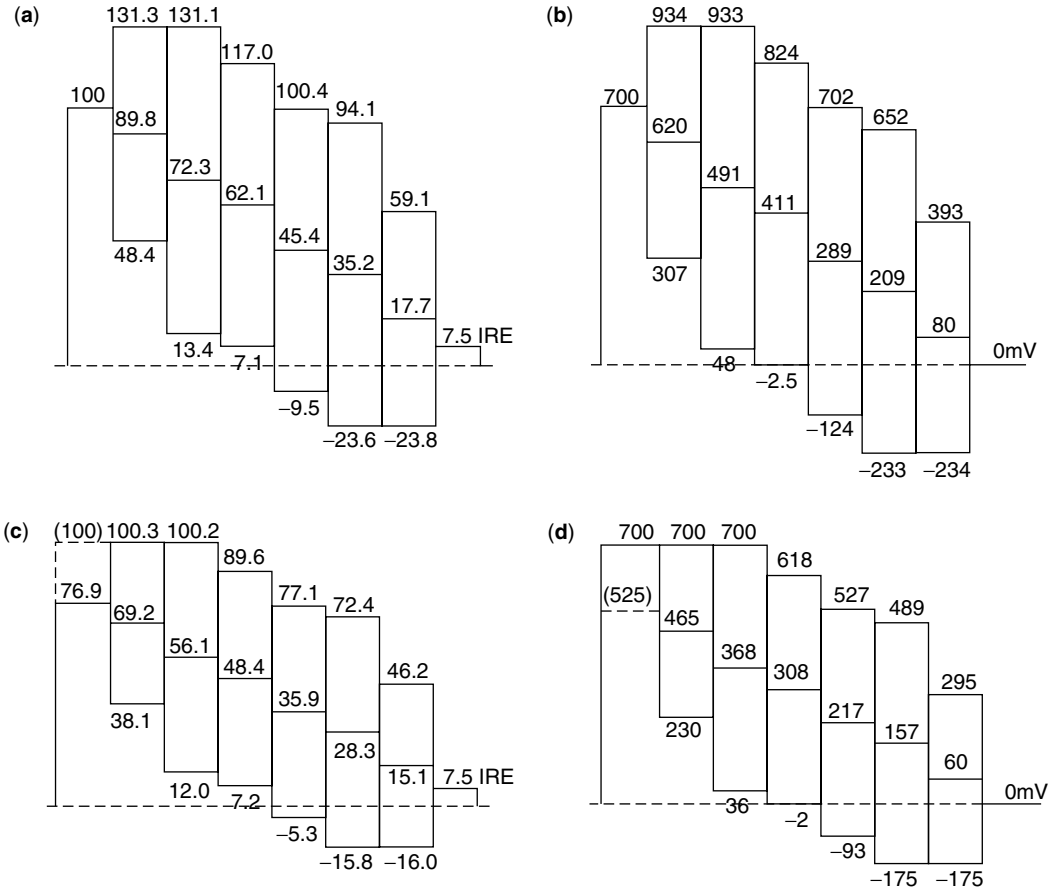


Figure 7. (**a**) 100% NTSC color bars (100/7.5/100/7.5). (**b**) 100% PAL color bars (100/0/100/0). (**c**) Standard 75% "EIA" color bars (75/7.5/75/7.5). (**d**) Standard 75% "EBU" color bars (100/0/75/0).

amplitude distortion. In addition, FM allowed recording the composite signal on conventional black-and-white tape machines because precise time base correction was not necessary.

Compatibility. In FM, the subcarrier is always present, superimposed on the luminance signal at constant amplitude (unlike NTSC, in which the subcarrier produces noticeable interference with the luminance only on highly saturated colors). To reduce its visibility, a number of techniques are employed.

First, preemphasis is applied to the baseband color-difference signals to lessen their amplitudes at lower saturation, but preserve adequate signal-to-noise ratio (low-level preemphasis; see Fig. 10). Second, different subcarrier frequencies are employed that are integral multiples of the scanning line frequency; f_{oB} is 4.25 MHz (272 H), and f_{oR} is 4.40625 MHz (282 H). The f_{oR} signal is inverted before modulation so that the maximum deviation is toward a lower frequency, reducing the bandwidth required for the dual subcarriers.

Third, another level of preemphasis is applied to the modulated subcarrier around a point between the two rest frequencies, known as the "cloche" frequency of 4.286 MHz (high-level preemphasis, the so-called "antibell" shaping shown in Fig. 11). Finally, the phase of the modulated subcarrier is reversed on consecutive fields and, additionally, on every third scanning line, or, alternately, every three lines.

Line Identification. Synchronizing the receiver to the alternating lines of color-difference signals is provided in one of two ways. Earlier specifications called for nine lines of vertical blanking to contain a field identification sequence formed by truncated sawteeth of the color-difference signals from the white point to the limiting frequency (so-called "bottles"; see Fig. 12). This method is referred to as "SECAM-V."

As use of the vertical interval increased for ancillary signals, receiver demodulators were fashioned to sample the unblanked subcarrier immediately following the horizontal sync pulse, providing an indication of line sequence from the rest frequency. Where this method is employed, it is called "SECAM-H." An advantage of this method is near-instantaneous recovery from momentary color field sequence errors, whereas SECAM-V receivers must wait until the next vertical interval.

Issues in Program Production. High-level preemphasis causes the chrominance subcarrier envelope to increase in amplitude at horizontal transitions, as can be seen on a waveform monitor (Fig. 13). Unlike NTSC, the subcarrier amplitude bears no relation to saturation, so, except for testing purposes, a luminance low-pass filter is employed

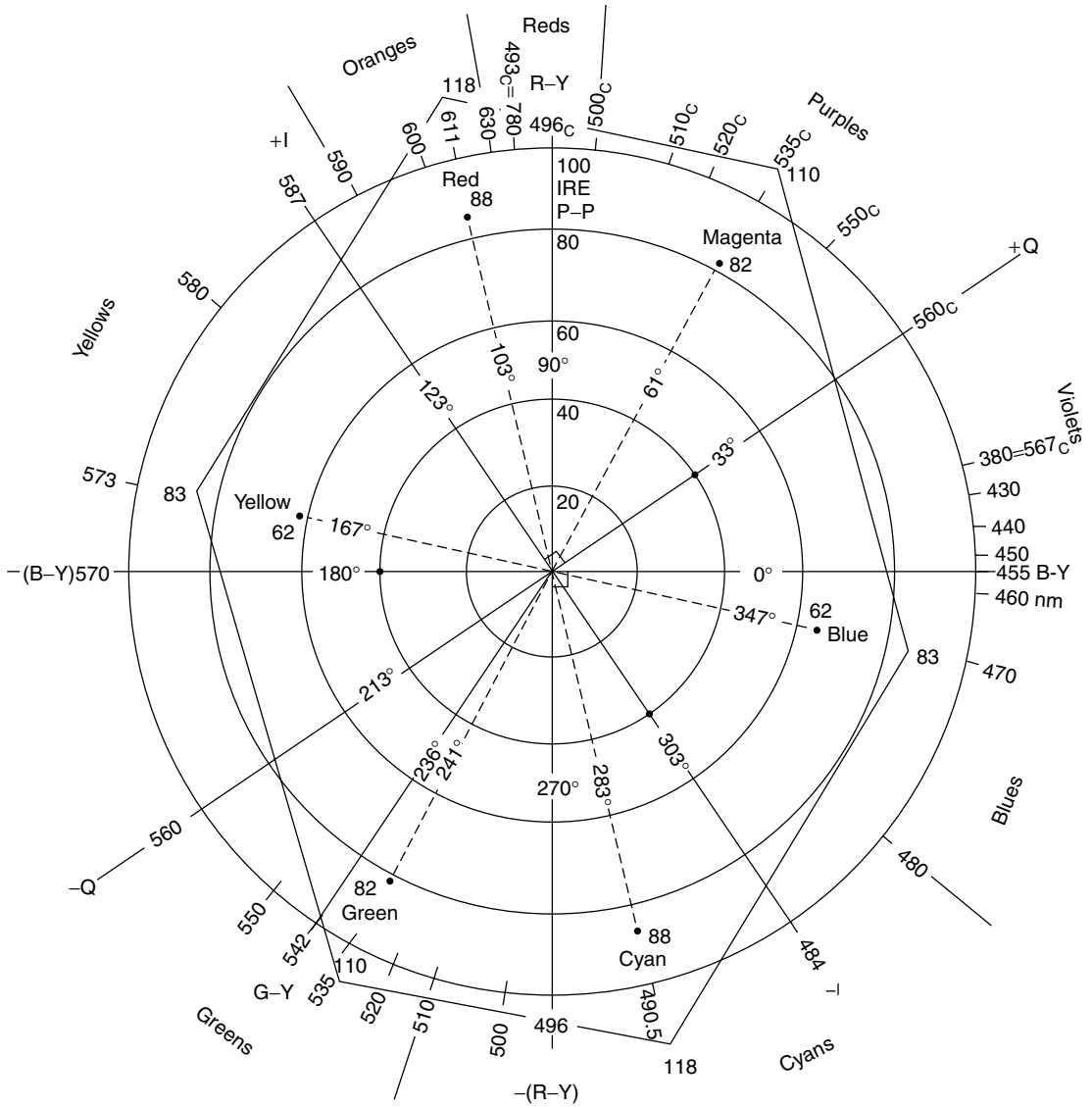


Figure 8. Vector relationship among chrominance components and corresponding dominant wavelengths. 75% color bars with 7.5% setup. Linear NTSC system, NTSC luminophors, illuminant C. Hexagon defines maximum chrominance subcarrier amplitudes as defined by 100% color bars with 7.5% setup. *Caution*: The outer calibration circle on vectorscopes does *not* represent exactly 100 IRE P–P.

on the waveform monitor. A vectorscope presentation of the saturation and hue is implemented by decoding the FM subcarrier into baseband color-difference signals and applying them to an X, Y display.

Unfortunately, the choice of FM for the color subcarrier means that conventional studio production switchers cannot be employed for effects such as mixing or fading from one scene to another because reducing the amplitude of the subcarrier does not reduce the saturation. This necessitates using a component switcher and then using encoding afterward. When the signal has already been encoded to SECAM (such as from a prerecorded video tape), it must be decoded before the component switcher and then reencoded.

Like NTSC, program editing must be done in two-frame increments. Although the subcarrier phase is reversed on a field-and-line basis, establishing a "12-field sequence," it is the instantaneous frequency — not phase — that defines the hue. However, the line-by-line sequence of the color-difference signals must be maintained. The odd number of scanning lines means that each successive frame begins with the opposite color-difference signal. As described before, mixes or fades are never done by using composite signals.

Because the instantaneous frequency of the subcarrier is not related to the line scanning frequency, it is impossible to employ modern comb-filtering techniques to separate the chrominance and luminance in decoding. Increasingly, special effects devices rely on decoding the composite TV signal to components for manipulation, then reencoding. Every operation of this sort impairs luminance resolution because a notch filter must be used around

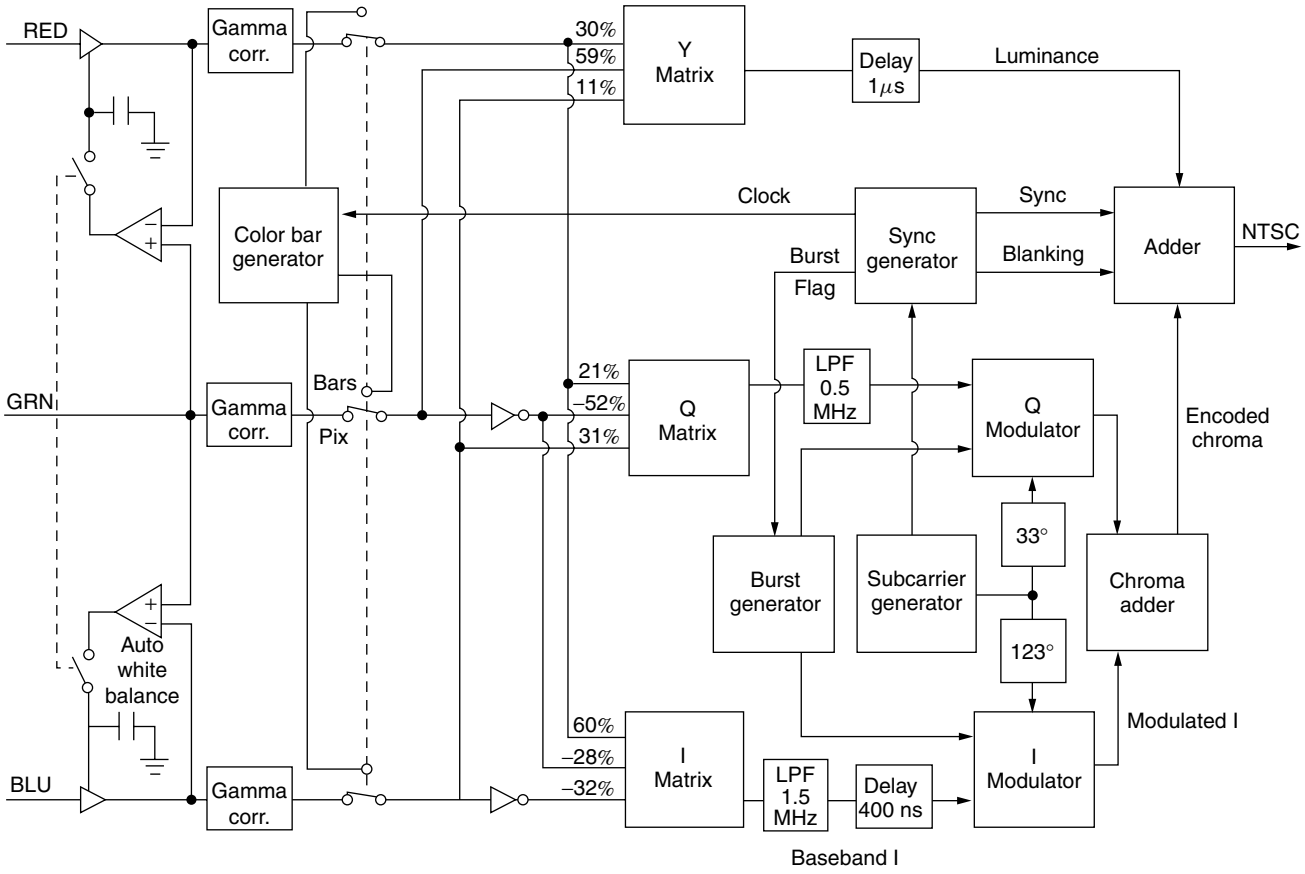


Figure 9. Block diagram of RGB to NTSC encoding (and related circuitry).

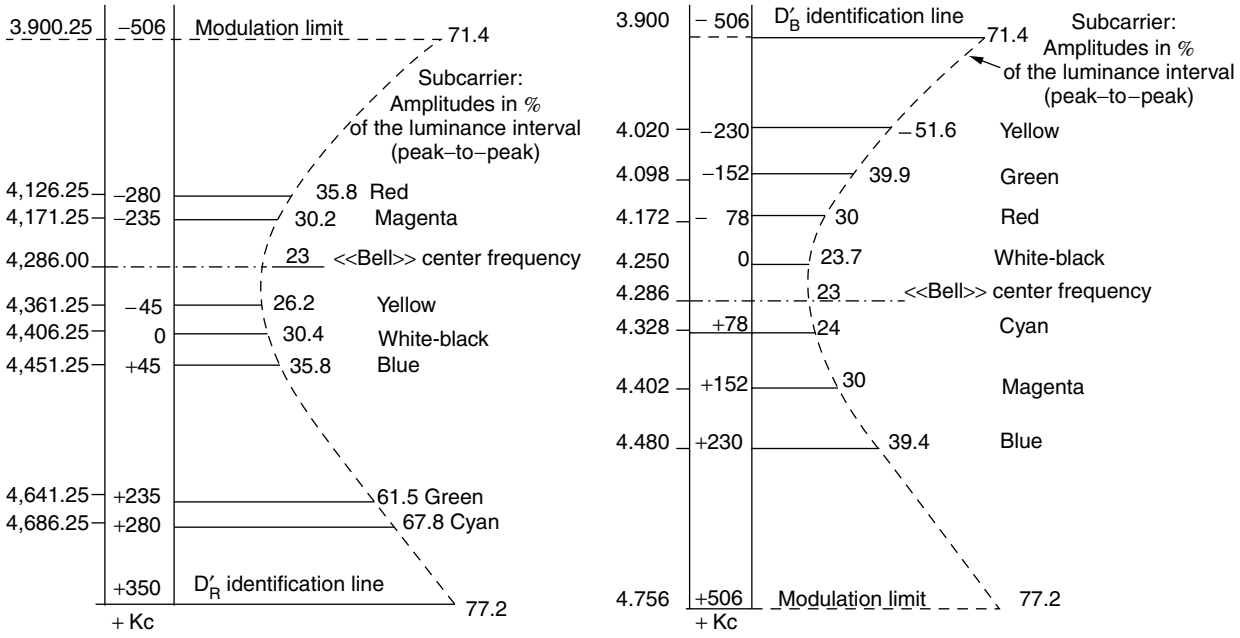


Figure 10. SECAM baseband (low-level) preemphasis.

the subcarrier frequency. These concerns have led many countries that formerly adopted SECAM to switch to PAL for program production and transcode to SECAM only for RF broadcasting.

Phase Alternate Line (PAL)

To retain the ease of the NTSC in program production, yet correct for phase errors, the German Telefunken Company developed a system more comparable to the NTSC that retains quadrature modulation. Because of the wider channel bandwidth of 625-line systems, the color subcarrier could be positioned so that the sidebands from both color-difference signals have the same bandwidth. This means that $R-Y$ and $B-Y$ signals could be used directly, rather than I and Q as in the NTSC. Identical scaling factors are used, and the signals are known as V and U, respectively.

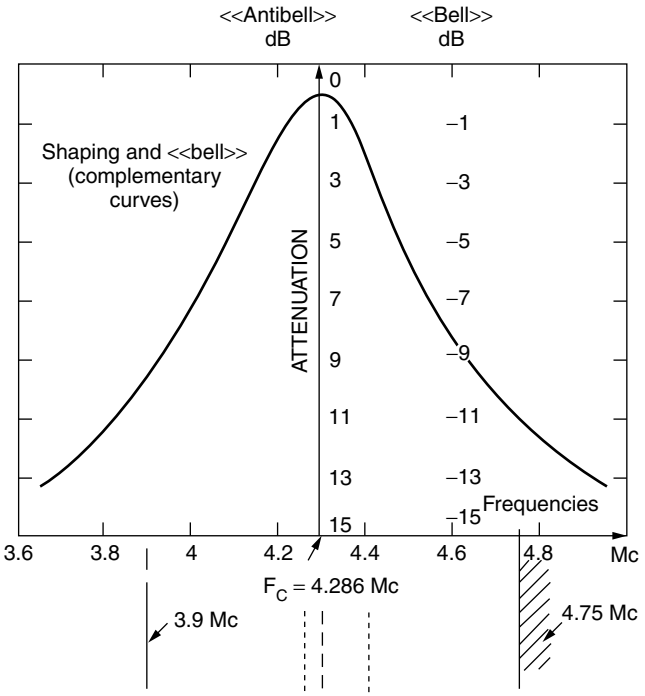


Figure 11. SECAM RF (high-level) preemphasis.

Color Phase Alternation. In the PAL System, the phase of the modulated V component of the chrominance signal is reversed on alternate lines to cancel chrominance phase distortion acquired in equipment or transmission. A reference is provided to indicate which lines have $+V$ or $-V$ phase by also shifting the phase of the color burst signal by $\pm 45^\circ$ on alternate lines. Any phase shift encountered will have the opposite effect on the displayed hue on adjacent lines in the picture. If the phase error is limited to just a few degrees, the eye integrates the error, because, more chrominance detail is provided in the vertical direction than can be perceived at normal viewing distances. Receivers based on this principle are said to have "simple PAL" decoders.

If the phase error is more than a few degrees, the difference in hue produces a venetian-blind effect, called *Hanover bars*. Adding a one-line delay permits integrating chrominance information from adjacent scanning lines electrically, and there is a slight reduction in saturation for large errors. Color resolution is reduced by half in the vertical direction but more closely matches horizontal resolution due to band-limiting in the encoder. This technique of decoding is called "deluxe PAL."

Compatibility. The line-by-line phase reversal results in an identical phase on alternate lines for hues on or near the V axis. To sustain a low-visibility interference pattern in PAL, the subcarrier frequency is made an odd multiple of one-quarter of the line frequency (creating eight distinct color fields). This effectively offsets the excursions of the V component by 90° line-to-line and offsets those of the U component by 270°. Because this 180° difference would cause the excursions of one component to line up vertically with those of the other in the next frame, an additional 25 Hz offset ($f_V/2$) is added to the PAL subcarrier frequency to reduce its visibility further.

In early subcarrier oscillator designs, the reference was derived from the mean phase of the alternating burst signal. Interlaced scanning causes a half-line offset between fields with respect to the vertical position, so that

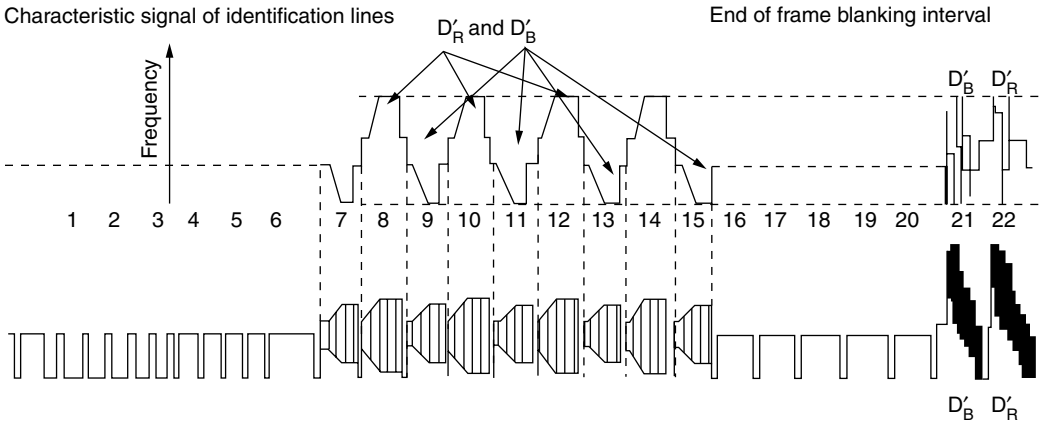


Figure 12. SECAM field identification "bottles."

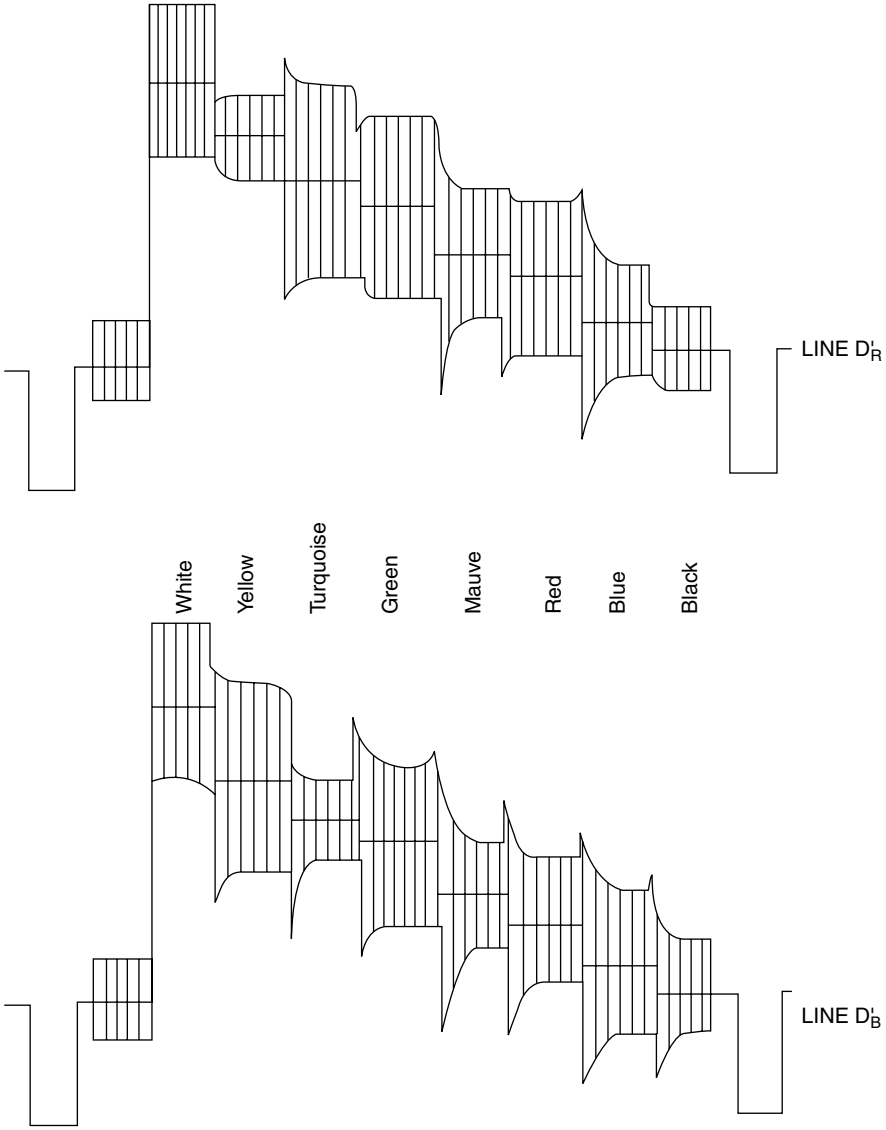


Figure 13. SECAM color bar waveforms.

the number of bursts actually blanked during the 7(1/2) H vertical interval would be different for the odd versus even fields. Because the phase of burst alternates line-to-line, the mean phase would then appear to vary in this region, causing disturbances at the top of the picture.

This is remedied by a technique known as "Bruch blanking." The burst blanking is increased to a total of nine lines and repositioned in a four-field progression to include the 7(1/2) H interval, such that the first and last burst of every field has a phase corresponding to $(-U + V)$ or $+135°$. The burst signal is said to "meander," so that color fields 3 and 7 have the earliest bursts.

Issues in Program Production. Because the subcarrier frequency in PAL is an odd multiple of one-quarter the line frequency, each line ends on a quarter-cycle. This, coupled with the whole number plus one-half lines per field, causes the phase of subcarrier to be offset in each field by $45°$. Thus, in PAL, the subcarrier phase repeats only every eight fields, creating an "eight-field sequence." This complicates program editing because edit points occur only every four frames, which is slightly less than 1/10 s.

Comb filtering to separate chrominance and luminance in decoding is somewhat more complicated in PAL; however, it has become essential for special picture effects. On a waveform monitor, the composite PAL signal looks very much like NTSC, except that the 25 Hz offset causes a slight phase shift from line to line, so that when viewing the entire field, the sine wave pattern is blurred. Because of the reversal in the phase of the V component on alternate lines, the vectorscope presentation has a mirror image about the U axis (Fig. 14).

Variations of PAL. The differences between most 625-line transmission standards involve only RF parameters (such as sound-to-picture carrier spacing). For 625-line PAL program production, a common set of technical specifications may be used. These standards are routinely referred to in the production environment as "PAL-B," although the baseband signals may be used with any 625-line transmission standard.

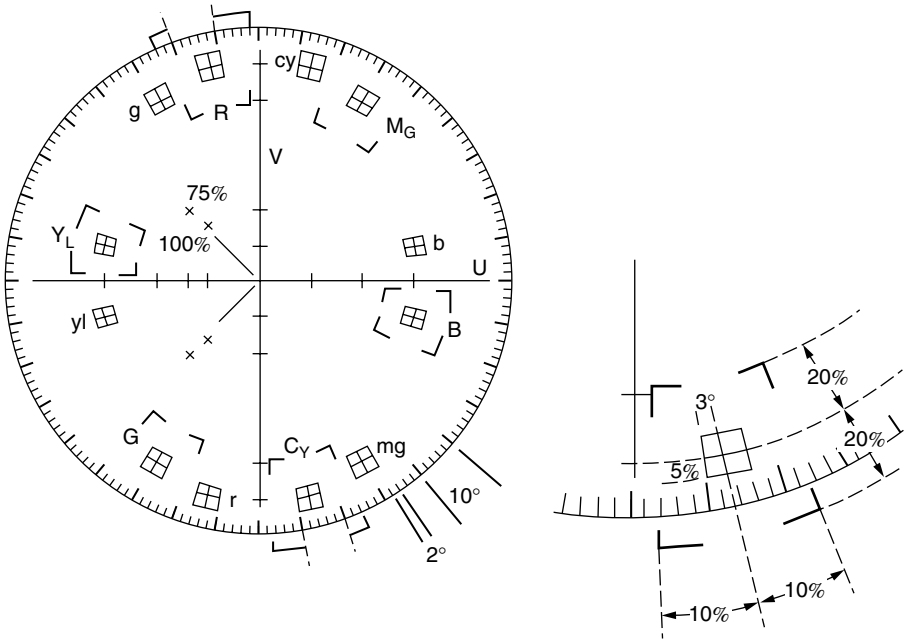


Figure 14. Typical PAL vectorscope graticule.

Several countries in South America have adopted the PAL system. The 6 MHz channel allocations in that region meant that the color subcarrier frequency had to be suitably located, about 1 MHz lower in frequency than for 7 MHz or 8 MHz channels. The exact frequencies are close to, but not the same as, those for the NTSC. The 625-line system is known as PAL-N. Studio production for this standard is done in conventional "PAL-B," then converted to PAL-N at the transmitter.

The 525-line PAL is known as PAL-M, and it requires studio equipment unique to this standard, although the trend is toward using conventional NTSC-M equipment and also transcoding to PAL-M at the transmitter. PAL-M does not employ a 25 Hz offset of the subcarrier frequency, as in all other PAL systems.

Similarities of Color Encoding Systems

The similarities of the three basic color television encoding systems are notable (see Table 3). They all rely on the concept of a separate luminance signal that provides compatibility with black-and-white television receivers. The "mixed-highs" principle combines high-frequency information from the three color primaries into luminance, where the eye is sensitive to fine detail; only the relatively low-frequency information is used for the chrominance channels. All three systems use the concept of a subcarrier, located in the upper frequency spectrum of luminance, to convey the chrominance information (Fig. 15).

All systems use color-difference signals, rather than the color primaries directly, to minimize cross talk with the luminance signal, and all derive the third color signal by subtracting the other two from luminance. The constant luminance principle is applied in all systems, based on the original NTSC picture tube phosphors, so the matrix formulations for luminance and color-difference signals are identical (some recent NTSC encoders use equal-bandwidth $R-Y$ and $B-Y$ signals, instead of proportioned-bandwidth I and Q signals).

All of the systems use scaling factors to limit excessive subcarrier amplitude (NTSC/PAL) or deviation (SECAM) excursions. Finally, all three systems use an unmodulated subcarrier sample on the back porch of the horizontal sync pulse for reference information in the decoding process.

Because of these similarities, conversion of signals between standards for international program distribution is possible. Early *standards converters* were optical, essentially using a camera of the target standard focused on a picture tube operating at the source standard. Later, especially for color, electronic conversion became practical.

The most serious issue in standards conversion involves motion artifacts due to the different field rates between 525- and 625-line systems. Simply dropping or repeating fields and lines creates disturbing discontinuities, so interpolation must be done. In modern units, the composite signals are decoded into components, using up to three-dimensional adaptive comb filtering, converted using motion prediction, then reencoded to the new standard. Table 4 lists the transmission and color standards used in various territories throughout the world.

Component Analog Video (CAV)

The advent of small-format video tape machines that recorded luminance and chrominance on separate tracks led to interest in component interconnection. Increasingly, new equipment decoded and reencoded the composite signal to perform manipulations that would be impossible or would cause significant distortions if done in the composite environment. It was reasoned that if component signals (Y, $R-Y$, $B-Y$) could be taken from the camera and if encoding to NTSC, PAL, or SECAM could be done just before transmission, then technical quality would be greatly improved.

Table 3. Principal Characteristics of Color Television Encoding Systems

System	NTSC	PAL	SECAM
Display primaries	FCC	EBU	EBU
White reference	CIE III C	CIE III D_{65}	CIE III C
Display gamma	2.2	2.8	2.8
Luminance	$E'_Y = +0.30E'_R - 0.59E'_G + 0.11E'_B$	$E'_Y = +0.299E'_R + 0.587E'_G + 0.114E'_B$	
Chrominance signals	$Q = +0.41(B-Y) + 0.48(R-Y)$ $I = -0.27(B-Y) + 0.74(R-Y)$	$U = 0.493(B-Y)$ $V = 0.877(R-Y)$	$D_B = +1.505(B-Y)$ $D_R = -1.902(R-Y)$
Chrominance baseband video preemphasis (kHz)	—	—	$D^*_B = A \times D_R$ $D^*_R = A \times D_R$ $A = \left\vert \frac{1 + j \times \frac{f_B/f_R}{85}}{1 + j \times \frac{f_B/f_R}{255}} \right\vert$
Modulation method	Amplitude modulation of two suppressed subcarriers in quadrature		Frequency modulation of two sequential subcarriers
Axes of modulation	$Q = 33°, I = 123°$	$U = 0°, V = \pm 90°$	—
Chroma BW/Deviation (kHz)	$Q = 620, I = 1, 300$	$U + V = 1, 300$	$\Delta f_{oB} = \pm 230 + 276/-120$, $\Delta f_{oR} = \pm 280 + 70/-226$
Vestigial sideband (kHz)	+620	+570 (PAL-B,G,H), +1, 070 (PAL-I), +620 (PAL-M,N)	—
Composite color video signal (CCVS)	$E_M = E'_Y + E'_Q(\sin \omega t + 33°) + E'_I(\cos \omega t + 33°)$	$E_M = E'_Y + E'_U \sin \omega t \pm E'_V \cos \omega t$	$E_M = E'_Y + G_{SC} \times \cos 2\pi(f_{oB} + D^*_B \Delta f_{oB})t + G_{SC} \times \cos 2\pi(f_{oR} + D^*_B \Delta f_{oR})t$
Modulated subcarrier amplitude/preemphasis	$G_{SC} = \sqrt{E'^2_Q + E'^2_I}$	$G_{SC} = \sqrt{E'^2_U + E'^2_V}$	$G_{SC} = D^*_B/D^*_R \times 0.115E'_Y(P - P) \times \left\vert \frac{1 + j(16)F}{1 + j(1.26)F} \right\vert$ $F = \frac{f_B/f_R}{f_0} - \frac{f_0}{f_B/f_R}$ ($f_0 = 4.286 \pm 0.02$MHz)
SC/H frequency relationship	$f_{SC} = (455/2)f_H$	$f_{SC} = (1, 135/4)f_H + f_V/2$ (PAL-B, G, H, I) = $(909/4)f_H$(PAL-M) = $(917/4)f_H + f_V/2$ (PAL-N)	$f_{oB} = 272\, f_H / f_{oR} = 282\, f_H$
Subcarrier frequency (MHz)	3.579545 ± 10 Hz	4.43361875 ± 5 Hz(PAL-B, G, H); ± 1 Hz(PAL-I) 3.57561149 ± 10 Hz(PAL-M) 3.58205625 ± 5 Hz(PAL-N)	$f_{oB} = 4.250000 \pm 2$ kHz, $f_{oR} = 4.406250 \pm 2$ kHz
Phase/Deviation of SC reference	180°	$+V = +135°, -V = -135°$	$D_B = -350$ kHz, $D_R = +350$ kHz
Start of SC reference (μs)	5.3 ± 0.1	5.6 ± 0.1 (PAL-B, G, H, I, N); 5.2 ± 0.5 (PAL-M)	5.7 ± 0.3
SC reference width (cycles)	9 ± 1	10 ± 1 (PAL-B, G, H, I); 9 ± 1 (PAL-M, N)	—
SC reference amplitude (mV)	286 (40 IRE $\pm$ 4)	300 ± 30	$D_B = 167, D_R = 215$

However, component signal distribution required that some equipment, such as switchers and distribution amplifiers, have three times the circuitry, and interconnection required three times the cable and connections as those of composite systems. This brought about consideration of *multiplexed analog component* (MAC) standards, whereby the luminance and chrominance signals are time-multiplexed into a single, higher bandwidth signal. No single standard for component signal levels emerged (Table 5), and the idea was not widely popular. Interest soon shifted to the possibility of digital signal distribution.

Digital Video

Devices such as time-base correctors, frame synchronizers, and standards converters process the TV signal in the digital realm but use analog interfaces. The advent of digital video tape recording set standards for signal sampling and quantization to the extent that digital interconnection became practical.

Component Digital. The European Broadcasting Union (EBU) and Society of Motion Picture and Television Engineers (SMPTE) coordinated research and conducted

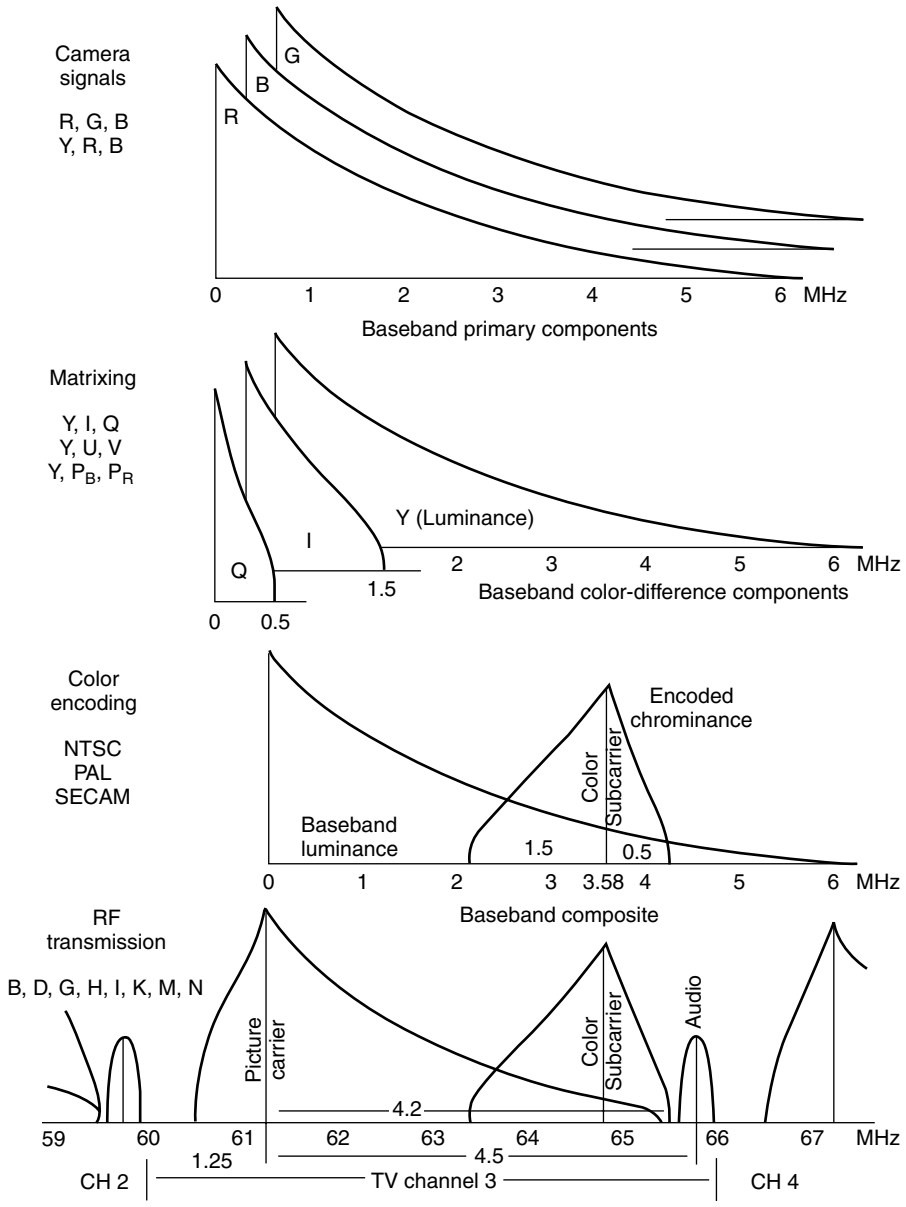


Figure 15. Four-stage color television frequency spectrum, showing the compression of three wideband color-separation signals from the camera through bandwidth limiting and frequency multiplexing into the same channel bandwidth used for black-and-white television.

demonstrations in search of a component digital video standard that would lend itself to the exchange of programs worldwide. A *common data rate* of 13.5 Mbps based on line-locked sampling of both 525- and 625-line standards was chosen. This allows analog video frequencies of better than 5.5 MHz to be recovered and is an exact harmonic of the scanning line rate for both standards, enabling great commonality in equipment.

A *common image format* of static orthogonal shape is also employed, whereby the sampling instants on every line coincide with those on previous lines and fields and also overlay the samples from previous frames. There are 858 total luminance samples per line for the 525-line system, 864 samples for the 625-line system, but 720 samples during the picture portion for both systems. This image structure facilitates filter design, special effects, compression, and conversion between standards.

For studio applications, the color-difference signals are sampled at half the rate of luminance, or 6.75 MHz, cosited with every odd luminance sample, yielding a total data rate of 27 Mbps. This provides additional resolution for the chrominance signals, enabling good special effects keying from color detail. The sampling ratio for luminance and the two chrominance channels is designated "4 : 2 : 2." Other related ratios are possible (Table 6).

Quantization is uniform (not logarithmic) for both luminance and color-difference channels. Eight-bit quantization, providing 256 discrete levels, provides adequate signal-to-noise ratio for videotape applications. However, the 25-pin parallel interface selected can accommodate two extra bits, because 10-bit quantization was foreseen as desirable in the future. Only the active picture information is sampled and quantized, allowing better resolution of the signal amplitude. Sync and blanking are coded by special signals (Figs. 16 and 17).

Table 4. National Television Transmission Standards

Territory	VHF	UHF	Color
Afars and Isaas = Djibouti			
Afghanistan	D		PAL
Albania	B	G	PAL
Algeria	B		PAL
Andorra	B		PAL
Angola	I		PAL
Antigua and Barbuda	M		NTSC
Argentina	N		PAL
Armenia	D	K	SECAM
Ascension Islands	I		
Australia	B	B	PAL
Austria	B	G	PAL
Azerbaijan	D	K	SECAM
Azores	B		PAL
Bahamas	M		NTSC
Bahrain	B	G	PAL
Bangladesh	B		PAL
Barbados	M		NTSC
Belarus	D	K	SECAM
Belgium	B	H	PAL
Benin	K′		SECAM
Bermuda	M		NTSC
Bolivia	M	M	NTSC
Bosnia and Herzegovina	B	G	PAL
Botswana	I		PAL
Brazil	M		PAL
Brunei Darussalam	B		PAL
Bulgaria	D	K	P & S
Burkina Faso	K′		SECAM
Burma = Myanmar			
Burundi		K′	SECAM
Cambodia	B		PAL
Cameroon	B		PAL
Canada	M	M	NTSC
Canary Islands	B	G	PAL
Cape Verde Islands	I		PAL
Cayman Islands	M		NTSC
Central African Republic	K′		SECAM
Ceylon = Sri Lanka			
Chad	K′		SECAM
Channel Islands		I	PAL
Chile	M		NTSC
China	D	D	PAL
Colombia	M		NTSC
Commonwealth of Independent States: see state			
Comores	K′		
Congo	K′		SECAM
Costa Rica	M		NTSC
Côte d'Ivoire	K′		SECAM
Croatia	B	G	PAL
Cuba	M		NTSC
Curacao	M	M	NTSC
Cyprus	B	G	P & S
Czech Republic	D	K	S P
Dahomey = Benin			
Denmark	B	G	PAL
Diego Garcia	M		NTSC
Djibouti	K′		SECAM
Dominican Republic	M		NTSC
Ecuador	M		NTSC
Equtorial Guinea = Fernando Po			
Egypt	B	G	P & S
El Salvador	M		NTSC

Table 4. (*Continued*)

Territory	VHF	UHF	Color
Eritrea	B		PAL
Estonia	D	K	SECAM
Ethiopia	B		PAL
Faeroe Islands	B	G	PAL
Falkland Islands	I		PAL
Fernando Po	B		PAL
Fiji	M		NTSC
Finland	B	G	PAL
France	L′	L	SECAM
French Guyana	K′		SECAM
French Polynesia	K′		SECAM
Gabon	K′		SECAM
Gambia	B		PAL
Georgia	D	K	SECAM
Germany	B	G	PAL
Ghana	B		PAL
Gibraltar	B	G	PAL
Greece	B	G	SECAM
Greenland	B		PAL
Grenada	M		NTSC
Guadeloupe	K′		SECAM
Guam	M	M	NTSC
Guatemala	M		NTSC
Guinea	K′		PAL
Guinea-Bissau	I		
Guyana, Republic of	M		NTSC
Haiti	M		NTSC
Honduras	M		NTSC
Hong Kong		I	PAL
Hungary	D	K	P & S
Iceland	B	G	PAL
India	B		PAL
Indonesia	B		PAL
Iran	B		SECAM
Iraq	B		SECAM
Ireland	I	I	PAL
Israel	B	G	PAL
Italy	B	G	PAL
Ivory Coast = Côte d'Ivoire			
Jamaica	M		NTSC
Japan	M	M	NTSC
Johnston Islands	M		NTSC
Jordan	B	G	PAL
Kampuchea = Cambodia			
Kazakhstan	D	K	SECAM
Kenya	B		PAL
Korea, Democracy of (N)	D		PAL
Korea, Republic of (S)	M	M	NTSC
Kuwait	B	G	PAL
Kyrgyzstan	D	K	SECAM
Laos	B		PAL
Latvia	D	K	SECAM
Lebanon	B	G	SECAM
Leeward Islands = Antigua			
Lesotho	I		PAL
Liberia	B		PAL
Libya	B		SECAM
Lichtenstein	B	G	PAL
Lithuania	D	K	SECAM
Luxembourg	B	G/L	P P/S
Macao		I	PAL
Macedonia	B	G	PAL
Madagascar	K′		SECAM

Table 4. (*Continued*)

Territory	VHF	UHF	Color
Madeira	B		PAL
Malawi	B		PAL
Malaysia	B		PAL
Maldives	B		PAL
Mali	K′		SECAM
Malta	B		PAL
Martinique	K′		SECAM
Mauritania	B		SECAM
Mauritius	B		SECAM
Mayotte	K′		SECAM
Mexico	M	M	NTSC
Micronesia	M		NTSC
Moldova	D	K	SECAM
Monaco	L	G/L	S P/S
Mongolia	D		SECAM
Montserrat	M		NTSC
Morocco	B		SECAM
Mozambique		B	PAL
Myanmar	M		NTSC
Namibia	I		PAL
Nepal	B		PAL
Netherlands	B	G	PAL
Netherlands Antilles	M		NTSC
New Caledonia	K′		SECAM
New Zealand	B	G	PAL
Nicaragua	M		NTSC
Niger	K′		SECAM
Nigeria	B		PAL
Norway	B	G	PAL
Oman	B	G	PAL
Pakistan	B		PAL
Palau	M		NTSC
Panama	M		NTSC
Papua New Guinea	B	G	PAL
Paraguay	N		PAL
Peru	M	M	NTSC
Philippines	M		NTSC
Poland	D	K	P & S
Portugal	B	G	PAL
Puerto Rico	M	M	NTSC
Qatar	B	G	PAL
Reunion	K′		SECAM
Romania	D	G/K	PAL
Russia	D	K	SECAM
Rwanda	K′		
St. Helena	I		
St. Pierre et Miquelon	K′	K′	SECAM
St. Kitts and Nevis	M		NTSC
Samoa (American)	M		NTSC
Samoa (Western)	B		PAL
São Tomé e Príncipe	B		PAL
San Andres Islands	M		NTSC
San Marino	B	G	PAL
Saudi Arabia	B	G	S/P S
Senegal	K′		SECAM
Serbia	B	G	PAL
Seychelles	B		PAL
Sierra Leone	B		PAL
Singapore	B		PAL
Slovakia	B	G/K	PAL
Slovenia	B	G	PAL
Society Islands = French Polynesia			
Somalia	B		PAL
South Africa	I	I	PAL
S. West Africa = Namibia			
Spain	B	G	PAL
Sri Lanka	B		PAL
Sudan	B		PAL
Suriname	M		NTSC
Swaziland	B	G	PAL
Sweden	B	G	PAL
Switzerland	B	G	PAL
Syria	B	G	P & S
Tahiti = French Polynesia			
Taiwan	M	M	NTSC
Tajikistan	D	K	SECAM
Tanzania	B	I	PAL
Thailand	B	M	P N
Togo	K′		SECAM
Trinidad and Tobago	M	M	NTSC
Tunisia	B	G	P S
Turks and Caicos	M		NTSC
Turkey	B	G	PAL
Turkmenistan	D	K	SECAM
Uganda	B		PAL
Ukraine	D	K	SECAM
USSR: see independent state			
United Arab Emirates	B	G	PAL
United Kingdom		I	PAL
United States	M	M	NTSC
Upper Volta = Burkina Faso			
Uruguay	N		PAL
Uzbekistan	D	K	SECAM
Venezuela	M		NTSC
Vietnam	D/M		S/N
Virgin Islands	M		NTSC
Yemen	B		PAL
Yugoslavia: see new state			
Zaire	K′		SECAM
Zambia	B		PAL
Zanzibar = Tanzania			
Zimbabwe	B		PAL

These specifications were standardized in ITU-R BT.601 — hence the abbreviated reference, "601 Video." The first component digital tape machine standard was designated "D1" by SMPTE. This term has become used in place of more correct designations. For wide-screen applications, a 360-Mbps standard scales up the number of sampling points for a 16 : 9 aspect ratio.

Interconnecting digital video equipment is vastly simplified by using a serial interface. Originally, an 8/9 block code was devised to facilitate clock recovery by preventing long strings of ones or zeros in the code. This would have resulted in a serial data rate of 243 Mbps. To permit serializing 10-bit data, scrambling is employed, and complementary descrambling is at the receiver. NRZI coding is used, so the fundamental frequency is half the bit rate of 270 MHz.

Composite Digital. Time-base correctors for composite 1 in. videotape recorders had been developed that had several lines of storage capability. Some early devices sampled at three times the color subcarrier frequency

Table 5. Component Analog Video Format Summary

Format	Color Bar Amplitudes (mV)			Peak Excursions (mV)		
	Channel 1	Channel 2	Channel 3	Synchronization Channels/Signals		Setup
	100% 75%	100% 75%	100% 75%			
R/G/B/S[a]	+1V/+750	+1V/+750	+1V/+750	S = −4V		No
G/B/R	+700/+525	+700/+525	+700/+525	G, B, R = −300		No
Y/I/Q (NTSC)	+714/+549	±393/ ± 295	±345/ ± 259	Y = −286		Yes
Y/Q/I (M_I)	+934/+714	±476/ ± 357	±476/ ± 357	Y = −286	I = −600	Yes
Y/R - Y/B - Y*	+700/+525	±491/ ± 368	±620/ ± 465	Y = −300		No
Y/U/V (PAL)	+700/+525	±306/ ± 229	±430/ ± 323	Y = −300		No
Betacam 525	+714/+549	±467/ ± 350	±467/ ± 350	Y = −286		Yes
2 CH Y/CTDM	+714/+549	±467/ ± 350		Y = ±286	C = −420	
Betacam 625	+700/+525	±467/ ± 350	±467/ ± 350	Y = −300		No
2 CH Y/CTDM	+700/+525	±467/ ± 350		Y = ±300	C = −420	
M_{II} 525	+700/+538	±324/ ± 243	±324/ ± 243	Y = −300		Yes
2 CH Y/CTCM	+714/+549	±350/ ± 263		Y = −286	C = −650	
M_{II} 625	+700/+525	±350/ ± 263	±350/ ± 263	Y = −300		No
2 CH Y/CTCM	+700/+525	±350/ ± 263		Y = −300	C = −650	
SMPTE/EBU ($Y/P_B/P_R$)	+700/+525	±350/ ± 263	±350/ ± 263	Y = −300		No

[a]Other levels possible with this generic designation.

Table 6. Sampling Structures for Component Systems[a]

	Sample/Pixel						Sample/Pixel				
	1	2	3	4	5		1	2	3	4	5
Line	YC_bC_r	YC_bC_r	YC_bC_r	YC_bC_r	YC_bC_r	Line	YC_bC_r	Y	YC_bC_r	Y	YC_bC_r
	YC_bC_r	YC_bC_r	YC_bC_r	YC_bC_r	YC_bC_r		YC_bC_r	Y	YC_bC_r	Y	YC_bC_r
	YC_bC_r	YC_bC_r	YC_bC_r	YC_bC_r	YC_bC_r		YC_bC_r	Y	YC_bC_r	Y	YC_bC_r
	YC_bC_r	YC_bC_r	YC_bC_r	YC_bC_r	YC_bC_r		YC_bC_r	Y	YC_bC_r	Y	YC_bC_r
			4 : 4 : 4						4 : 2 : 2		
	Sample/Pixel						Sample/Pixel				
	1	2	3	4	5		1	2	3	4	5
Line	YC_bC_r	Y	Y	Y	YC_bC_r	Line	Y	Y	Y	Y	Y
							C_bC_r		C_bC_r		C_bC_r
	YC_bC_r	Y	Y	Y	YC_bC_r		Y	Y	Y	Y	Y
	YC_bC_r	Y	Y	Y	YC_bC_r		Y	Y	Y	Y	Y
							C_bC_r		C_bC_r		C_bC_r
	YC_bC_r	Y	Y	Y	YC_bC_r		Y	Y	Y	Y	Y
			4 : 1 : 1						4 : 2 : 0		

[a]Y = luminance sample; C_bC_r = chrominance samples; YC_bC_r = pixels so shown are cosited. Boldfaced type indicates bottom field, if interlaced.

($3f_{SC}$); however, better filter response could be obtained by $4f_{SC}$ sampling. The sampling instants correspond to peak excursions of the I and Q subcarrier components in NTSC. The total number of samples per scanning line is 910 for NTSC and 1,135 for PAL. To accommodate the 25 Hz offset in PAL, lines 313 and 625 each have 1,137 samples. The active picture portion of a line consists of 768 samples in NTSC and 948 samples in PAL. These specifications are standardized as SMPTE 244M (NTSC) and EBU Tech. 3,280 (PAL).

Unlike component digital, nearly the entire horizontal and vertical blanking intervals are sampled and quantized, which degrades the amplitude resolution (Fig. 18). However, in PAL, no headroom is provided for sync level, and the sampling instants are specified at 45° from the peak excursions of the V and U components of the subcarrier (Fig. 19). This allows "negative headroom" in the positive direction. Thus, an improvement of about 0.5 dB in the signal-to-noise ratio is obtained.

Waveform location	Voltage level	8-Bit Decimal Value	8-Bit Hex Value	8-Bit Binary Value	10-Bit Decimal Value	10-Bit Hex Value	10-Bit Binary Value
Excluded	766 .3 mV				1023	3FF	11 1111 1111
Excluded	763 .9 mV	255	FF	1111 1111	1020	3FF	11 1111 1100
Peak	700 .0 mV	235	EB	1110 1011	940	3AC	11 1010 1100
Black	0 .0 mV	16	10	0001 0000	64	040	00 0100 0000
Excluded	−48.7 mV				3	003	00 0000 0011
Excluded	−51.1 mV	0	0	0000 0000	0	000	00 0000 0000

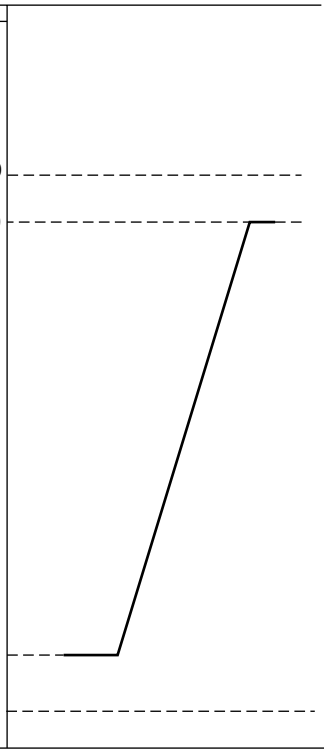


Figure 16. Quantizing levels for component digital luminance.

Waveform location	Voltage level	8-Bit Decimal Value	8-Bit Hex Value	8-Bit Binary Value	10-Bit Decimal Value	10-Bit Hex Value	10-Bit Binary Value
Excluded	399.2 mV				1023	3FF	11 1111 1111
Excluded	396.9 mV	255	FF	1111 1111	1020	3FC	11 1111 1100
Max positive	350.0 mV	240	F0	1111 0000	960	3C0	11 1100 0000
Black	0.0 mV	128	80	1000 0000	512	200	10 0000 0000
Max negative	−350.0 mV	16	10	0001 0000	64	040	00 0100 0000
Excluded	−397.7 mV				3	003	00 0000 0011
Excluded	−400.0 mV	0	00	0000 0000	0	000	00 0000 0000

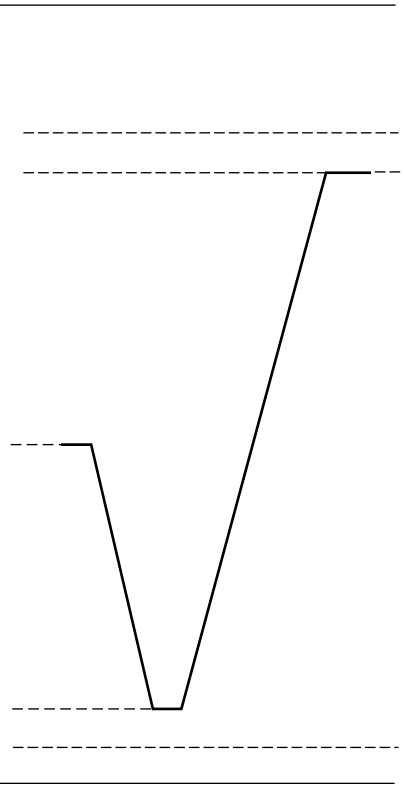


Figure 17. Quantizing levels for component digital color difference.

Waveform location	Voltage level	IRE units	8-Bit Decimal value	8-Bit Hex value	8-Bit Binary value	10-Bit Decimal value	10-Bit Hex value	10-Bit Binary value
Excluded	998.7 mV	139.8				1023	3FF	11 1111 1111
Excluded	994.9 mV	139.3	255	FF	1111 1111	1020	3FC	11 1111 1100
100% Chroma	907.7 mV	131.3	244	F4	1111 0100	975	3CF	11 1100 1111
Peak white	714.3 mV	100	200	C8	1100 1000	800	320	11 0010 0000
Blanking	0.0 mV	0	60	3C	0011 1100	240	0F0	00 1111 0000
Sync tip	−285.7 mV	−40	4	04	0000 0100	16	101	00 0001 0000
Excluded	−302.3 mV	−42.3				3	003	00 0000 0011
Excluded	−306.1 mV	−42.9	0	00	0000 0000	0	000	00 0000 0000

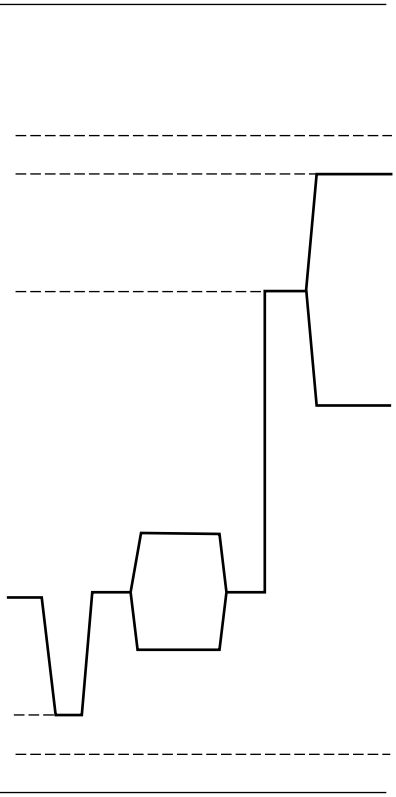


Figure 18. Quantizing levels for composite digital NTSC.

Rate conversion between component and composite digital television signals involves different sampling points and quantizing levels. Each conversion degrades the picture because exact levels cannot be reproduced in each pass. An important advantage of digital coding is thereby lost. In addition, decoding composite signals requires filtering to prevent cross-luminance and cross-color effects. This forever removes a part of the information; therefore, this process must be severely limited in its use.

Ancillary data may be added to digital component and composite video signals. AES/EBU-encoded digital audio can be multiplexed into the serial bit stream. Four channels are possible in the composite format, and 16 channels are possible in component digital video.

Component Video Standards

The video signals from a camera before encoding to NTSC, PAL, SECAM, or the ATSC Digital Standard are normally green (G), blue (B), and red (R). These are described as component signals because they are parts or components of the whole video signal. It has been found more efficient of bandwidth use for distribution and sometimes for processing to convert these signals into a luminance signal (Y) and two color-difference signals, blue minus luminance ($B - Y$) and red minus luminance ($R - Y$), where the color difference signals use 1/2 or 1/4 of the bandwidth of the luminance signal. The SMPTE/EBU Standard N10 adopted has a uniform signal specification for all 525/60 and 625/50 television systems. When the color-difference signals in this standard, are digitally formatted, they are termed C_b and C_r, respectively. At the same time, due to the lower sensitivity of the human eye to fine detail in color, it is possible to reduce the bandwidth of the color-difference signals compared to that of the luminance signal.

When these signals are digitized according to International Telecommunication Union, Radiocommunication Sector, (ITU-R) Recommendation 601, for both 525/60 and 625/50 systems, several modes of transmission may be used, all based on multiples of a 3.75 MHz sampling rate. For the ATSC standard, 4 : 2 : 0 is used (see below and the ATSC digital television standard). Either eight, or more frequently, 10 bits per sample are used.

4 : 4 : 4 Mode. The G, B, R or Y, C_b, C_r signal at an equal sampling rate of 13.5 MHz for each channel is termed the 4 : 4 : 4 mode of operation, and it yields 720 active samples per line for both 525/60 and 625/50 standards. This mode is frequently used for postproduction. If a (full-bandwidth) key signal must also be carried with the video, this combination is known as a 4 : 4 : 4 : 4 signal.

4 : 2 : 2 Mode. The 4 : 2 : 2 mode is more frequently used for distribution, where Y is sampled at 13.5 MHz, and the color-difference signals are sampled at a 6.25 MHz rate, corresponding to 360 active samples per line.

4 : 1 : 1 Mode. The 4 : 1 : 1 mode is used where bandwidth is at a premium, and the color-difference signals are each sampled at a 3.75 MHz rate, corresponding to 180 samples per line.

Waveform location	Voltage level	8-Bit Decimal value	8-Bit Hex value	8-Bit Binary value	10-Bit Decimal value	10-Bit Hex value	10-Bit Binary value
Excluded	913.1 mV				1023	3FF	11 1111 1111
Excluded	909.5 mV	255	FF	1111 1111	1020	3FC	11 1111 1100
100% Chroma Highest sample	886.2 mV	250	F4	1111 1010	3E8	3E8	11 1110 1000
Peak white	700.0 mV	211	D3	1101 0011	844	34C	11 0100 1100
Blanking	0.0 mV	64	40	0100 0000	256	100	01 0000 0000
Sync tip	–300.0 mV	1	01	0000 0001	4	004	00 0000 0010
Excluded	–301.2 mV				3	003	00 0000 0011
Excluded	–304.8 mV	0	00	0000 0000	0	000	00 0000 0000

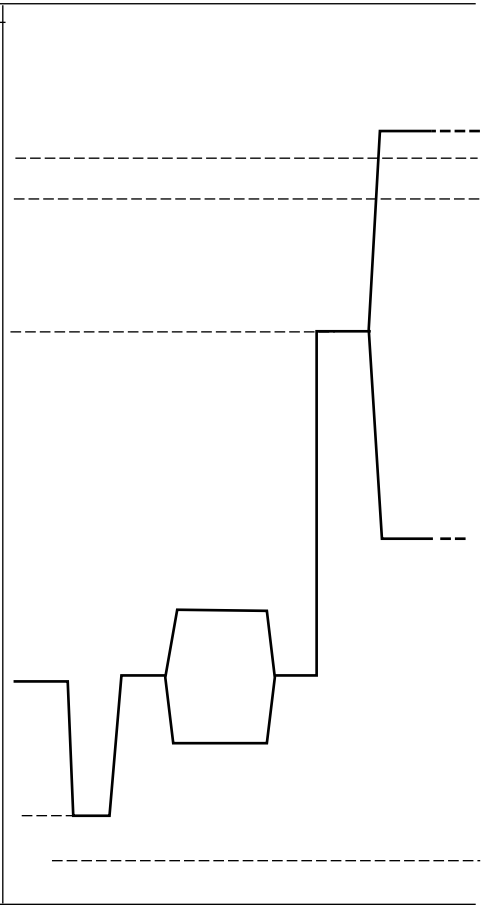


Figure 19. Quantizing levels for composite digital PAL.

4 : 2 : 0 Mode. A further alternative, 4 : 2 : 0 mode, whose structure is not self evident, is derived from a 4 : 2 : 2 sampling structure but reduces the vertical resolution of the color-difference information by 2 : 1 to match the reduced color-difference horizontal resolution. Four line (and field sequential if interlaced) cosited C_b, C_r samples are vertically interpolated weighted toward the closest samples, and the resultant sample is located in between two adjacent scanning lines. This mode is used in MPEG bit-rate reduced digital signal distribution formats, and hence in the ATSC digital television standard.

These four modes are illustrated in Table 6.

ADVANCED TELEVISION SYSTEMS, CURRENT AND FUTURE

ATSC Digital Television Standard

Overview. From 1987 to 1995, the Advisory Committee on Advanced Television Service (ACATS) to the Federal Communications Commission, with support from Canada and Mexico, developed a recommendation for an Advanced Television Service for North America. The ACATS enlisted the cooperation of the best minds in the television industry, manufacturing, broadcasting, cable industry, film industry, and federal regulators, in its organization to develop an advanced television system that would produce a substantial improvement in video images and in audio performance over the existing NTSC, 525-line system. The primary video goal was at least a doubling of horizontal and vertical resolution and a widening in picture aspect ratio from the current 4 (W) × 3 (H) to 16 (W) × 9 (H), and this was named "high-definition television." Also included was a digital audio system consisting of five channels plus a low-frequency channel (5.1).

Twenty-one proposals were made for terrestrial transmission systems for extended-definition television (EDTV) or high-definition television (HDTV), using varying amounts of the RF spectrum. Some systems augmented the existing NTSC system by an additional channel of 3 MHz or 6 MHz, some used a separate simulcast channel of 6 MHz or 9 MHz bandwidth, and all of the early systems used hybrid analog/digital technology in signal processing by an analog RF transmission system. Later proposals changed the RF transmission system to digital along with all-digital signal processing.

It was also decided that the signal would be transmitted in a 6 MHz RF channel, one for each current broadcaster of the (6 MHz channel) NTSC system, and that this new channel would eventually replace the NTSC channels. The additional channels were created within the existing UHF spectrum by improved design of TV receivers, so that the previously taboo channels, of which there were many, could now be used.

In parallel with this effort, the Advanced Television Systems Committee (ATSC) documented and developed the standard known as the ATSC Digital Television Standard, and it is subsequently developing related implementation standards.

In countries currently using 625-line, 4 : 3 aspect ratio television systems, plans are being developed to use a 1,250-line, 16 : 9 aspect ratio system eventually, and the ITU-R has worked successfully to harmonize and provide interoperability between the ATSC and 1,250-line systems.

Figure 20 shows the choices by which the signals of the various television standards will reach the consumer. Other articles detail satellite, cable TV, and asynchronous transfer mode (ATM) common carrier networks.

The ATSC and the ITU-R have agreed on a digital terrestrial broadcasting model, which is shown in Fig. 21. Video and audio sources are coded and compressed in separate video and audio subsystems. The compressed video and audio are then combined with ancillary data and control data in a service multiplex and transport, in which form the combined signals are distributed to the terrestrial transmitter. The signal is then channel-coded, modulated, and fed at appropriate power to the transmission antenna. The receiver reverses the process, demodulates the RF signal to the transport stream, then demultiplexes the audio, video, ancillary, and control data into their separate but compressed modes, and the individual subsystems then decompress the bit streams into video and audio signals that are fed to display screen and speakers, and the ancillary and control data are used if and as appropriate within the receiver.

Information Service Multiplex and Transport System. These subsystems provide the foundation for the digital communication system. The raw digital data are first formatted into *elementary bit streams*, representing image data, sound data, and ancillary data. The elementary bit streams are then formed into manageable packets of information (*packetized elementary stream*, PES), and a mechanism is provided to indicate the start of a packet (synchronization) and assign an appropriate identification code (*packet identifier*, PID) within a header to each packet. The packetized data are then multiplexed into a *program transport stream* that contains all of the information for a single (television) program. Multiple program transport streams may then be multiplexed to form a *system level multiplex transport stream*.

Figure 22 illustrates the functions of the multiplex and transport system and shows its location between the application (e.g., audio or video) encoding function and the transmission subsystem. The transport and demultiplex subsystem functions in the receiver in the reverse manner, being situated between the RF modem and the individual application decoders.

Fixed-Length Packets. The transport system employs the fixed-length transportation stream packetization approach defined by the Moving Picture Experts Group (MPEG), which is well suited to the needs of terrestrial broadcast and cable television transmission of digital television. The use of moderately long, fixed-length packets matches well with the needs for error protection, and it provides great flexibility for initial needs of the service to multiplex audio, video, and data, while providing backward compatibility for the future and maximum interoperability with other media (MPEG-based).

Packet Identifier. The use of a PID in each packet header to identify the bit stream makes it possible to have a mix

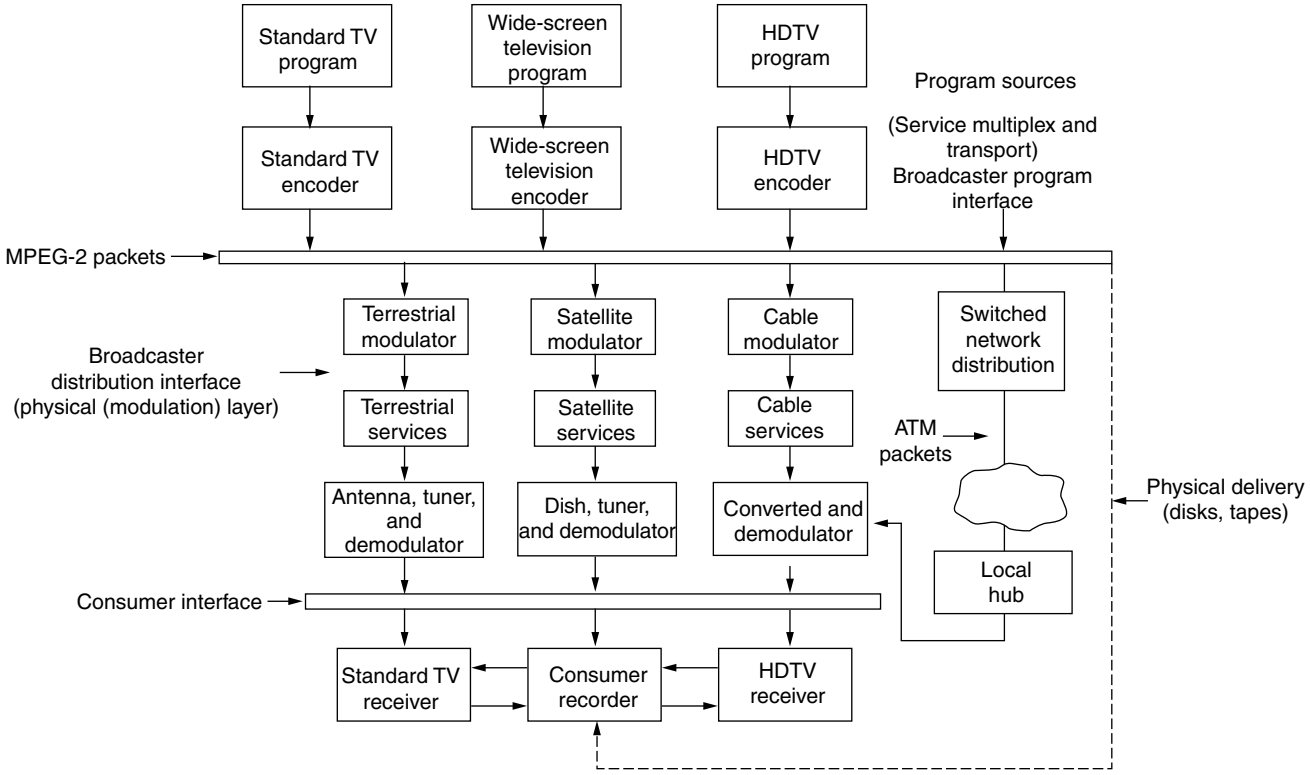


Figure 20. Television service model.

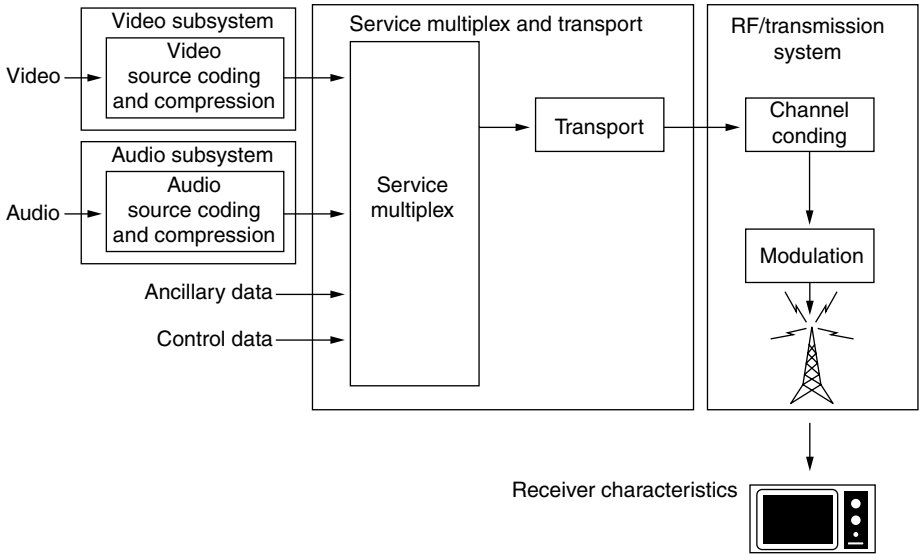


Figure 21. Block diagram showing ATSC and ITU-R terrestrial television broadcasting model.

of audio, video, and auxiliary data that is not specified in advance.

Scalability and Extensibility. The transport format is scalable in that more elementary bit streams may be added at the input of the multiplexer or at a second multiplexer. Extensibility for the future could be achieved without hardware modification by assigning new PIDs for additional elementary bit streams.

Robustness. After detecting errors during transmission, the data bit stream is recovered starting from the first good packet. This approach ensures that recovery is independent of the properties of each elementary bit stream.

Transport Packet Format. The data transport packet format, shown in Fig. 23, is based on fixed-length packets (188 bytes) identified by a variable-length header, including a sync byte and the PID. Each header identifies a

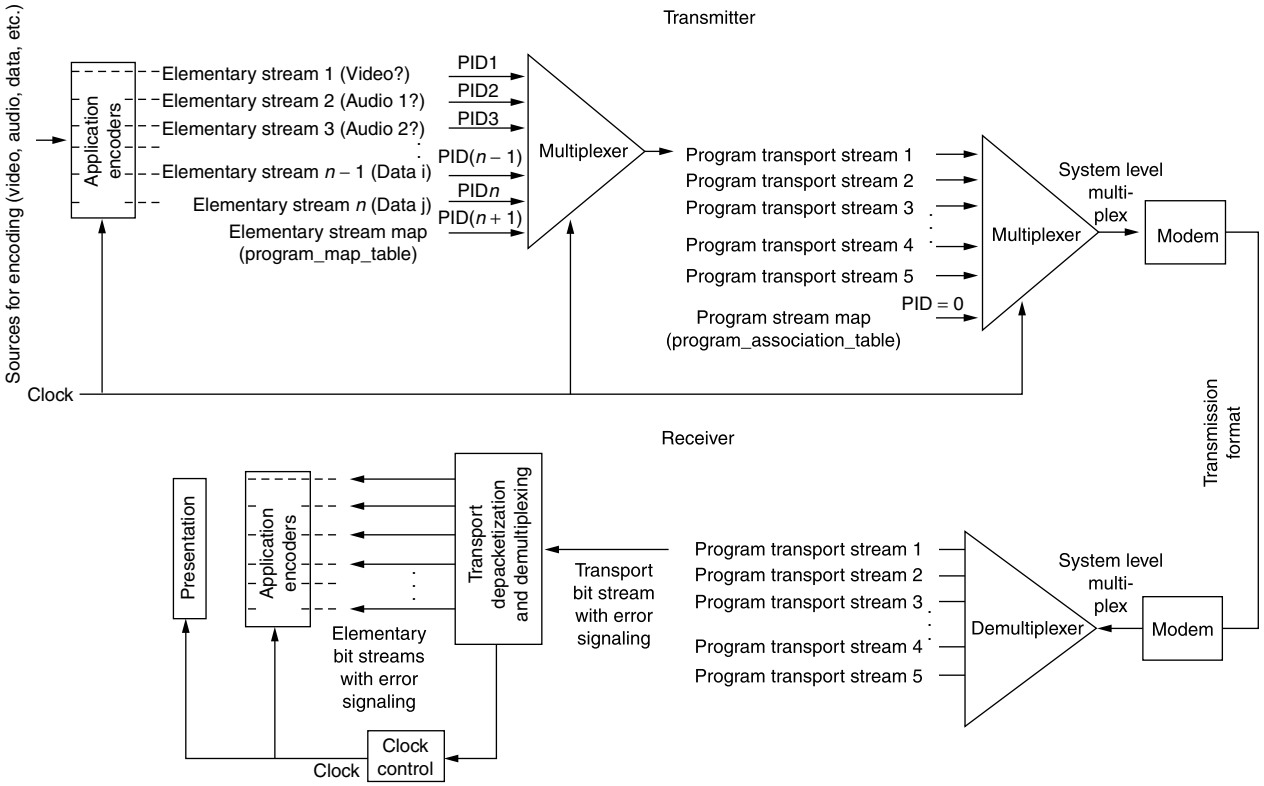


Figure 22. Organization of functionality within a transport system for digital TV programs.

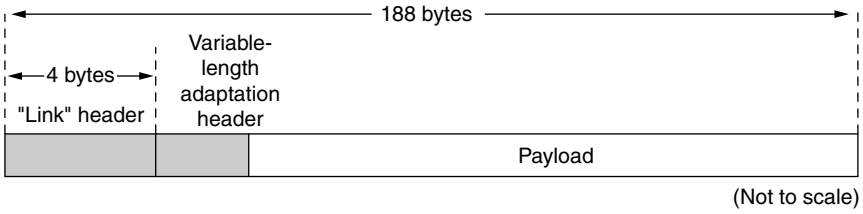


Figure 23. Transport packet format.

particular application bit stream (elementary bit stream) that forms the payload of the packet. Applications include audio, video, auxiliary data, program and system control information, and so on.

PES Packet Format. The elementary bit streams themselves are wrapped in a variable-length packet structure called the packetized elementary stream (PES) before transport processing. (Fig. 24). Each PES packet for a particular elementary bit stream then occupies a variable number of transport packets, and data from the various elementary bit streams are interleaved with each other at the (fixed-length) transport packet layer. New PES packets always start a new transport packet, and stuffing bytes (i.e., null bytes) are used to fill partially filled transport packets.

Channel Capacity Allocation. The entire channel capacity can be reallocated to meet immediate service needs. As an example, ancillary data can be assigned fixed amounts depending on a decision as to how much to allocate to video; or if the data transmission time is not critical, then it can be sent as *opportunistic data* during periods when the video channel is not fully loaded.

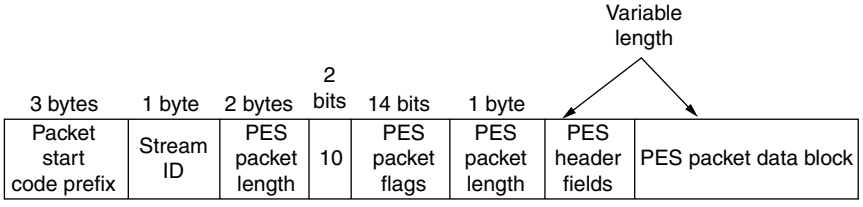


Figure 24. Structural overview of packetized elementary stream (PES) packet.

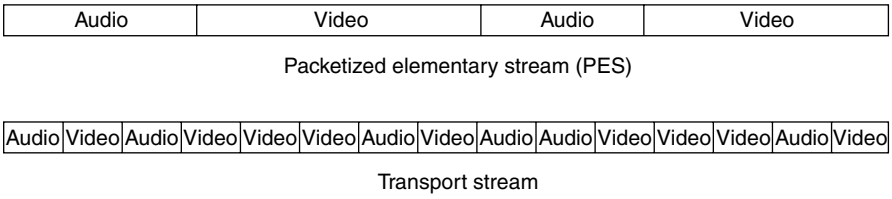


Figure 25. Variable-length PES packets and fixed-length transport packets.

Figure 25 illustrates how the variable-length PES packets relate to the fixed-length transport packets.

The transport system provides other features, including decoder synchronization, conditional access, and local program insertion. Issues relating to the storage and playback of programs are also addressed, and the appropriate hooks are provided to support the design of consumer digital products based on recording and playback of these bitstreams, including the use of "trick modes" such as slow motion and still frame, typical of current analog video cassette recorders (VCRs).

Local Program Insertion. This feature is extremely important to permit local broadcast stations to insert video, audio, or data unique to that station. As shown in Fig. 26 to splice local programs, it is necessary to extract (by demultiplexing) the transport packets, identified by the PIDs of the individual elementary bit streams, which make up the program that is to be replaced, including the *program map table*, which identifies the individual bit streams that make up the program. Program insertion can then take place on an individual PID basis, using the fixed-length transport packets.

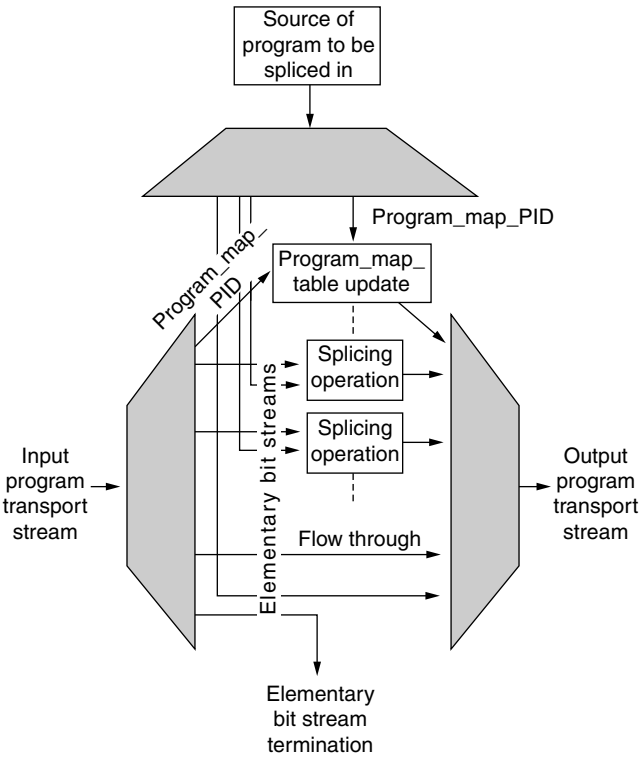


Figure 26. Example of program insertion architecture.

Presentation Time Stamp and Decoding Time Stamp. Both of these time stamps occur within the header of the PES packet, and they are used to determine when the data within the packet should be read out of the decoder. This process ensures the correct relative timing of the various elementary streams at the decoder relative to the timing at which they were encoded.

Interoperability with ATM. The MPEG-2 transport packet size (188 bytes) is such that it can easily be partitioned for transfer in a link layer that supports asynchronous transfer mode (ATM) transmission (53 bytes per cell). The MPEG-2 transport layer solves MPEG-2 presentation problems and performs the multimedia multiplexing function, and the ATM layer solves switching and network adaptation problems.

Video Systems

Compressed Video. Compression in a digital HDTV system is required because the bit rate required for an uncompressed HDTV signal approximates 1 Gbps (when the luminance/chrominance sampling is already compressed to the 4 : 2 : 2 mode). The total transmitted data rate over a 6 MHz channel in the ATSC digital television standard is approximately 19.4 Mbps. Therefore, a compression ratio of 50 : 1 or greater is required.

The ATSC Digital Television Standard specifies video compression using a combination of compression techniques which, for compatibility, conform to the algorithms of MPEG-2 Main Profile, High Level.

The goal of the compression and decompression process is to produce an approximate version of the original image sequence, such that the reconstructed approximation is imperceptibly different from the original for most viewers, for most images, and for most of the time.

Production Formats. A range of production format video inputs may be used. These include the current NTSC format of 483 active lines, 720 active samples/line, 60 fields, 2 : 1 interlaced scan (60I); the Standard Definition format of 480 active lines, 720 active samples/line, 60 frames progressive scan (60P); and high definition formats of 720 active lines, 1,280 active samples/line, 60P, or 1,080 active lines, 1,920 active samples/line, 60I.

Compression Formats. A large range of 18 compression formats is included to accommodate all of the those production formats. The 30P and 24P formats are included primarily to provide efficient transmission of film images associated with these production formats. The VGA Graphics format is also included at 480 lines and 640 pixels

Table 7. ATSC Compression Formats: A Hierarchy of Pixels and Bits[a]

Active Lines	Pixels per Line	Total Pixels per Frame	Uncompressed Payload Bit Rate in Mbps (8-bit 4 : 2 : 2 sampling) at Picture (Frame) Rate				Aspect Ratio and Notes
			60P	60I	30P	24P	
1,080	1,920	2,073,600	Future	995	995	796	16 : 9 only
720	1,280	921,600	885	—	442	334	16 : 9 only
480	704	337,920	324	162	162	130	16 : 9&4 : 3
480	640	307,200	295	148	148	118	4 : 3 only (VGA)
Vertical Resolution	Horizontal Resolution		Higher	← → Temporal Resolution		Lower	

[a]Data courtesy of Patrick Griffis, Panasonic, NAB, 1998.

(see later for pixel definition). Details of these compression formats are found in Table 7.

Colorimetry. The Digital Television Standard specifies SMPTE 274M colorimetry (same as ITU-R BT.709, 1990) as the default and preferred colorimetry. This defines the color primaries, transfer characteristics, and matrix coefficients.

Sample Precision. After preprocessing, the various luminance and chrominance samples will typically be represented using 8 bits per sample of each component.

Film Mode. In the case of 24 fps film which is sent at 60 Hz rate using a 3 : 2 pull-down operation, the processor may detect the sequences of three nearly identical pictures followed by two nearly identical pictures and may encode only the 24 unique pictures per second that existed in the original film sequence. This avoids sending redundant information and permits higher quality transmission. The processor may detect similar sequencing for 30 fps film and may encode only the 30 unique pictures per second.

Color Component Separation and Processing. The input video source to the video compression system is in the form of RGB components matrixed into luminance (Y) (intensity or black-and-white picture) and chrominance (C_b and C_r) color-difference components, using a linear transformation. The Y, C_b, and C_r signals have less correlation with each other than R, G, and B and are thus easier to code. The human visual system is less sensitive to high frequencies in the chrominance components than in the luminance components. The chrominance components are low-pass-filtered and subsampled by a factor of 2 in both horizontal and vertical dimensions (4 : 2 : 0 mode) (see section entitled Component Video Standards).

Representation of Picture Data. Digital television uses digital representation of the image data. The process of digitization involves sampling the analog signals and their components in a sequence corresponding to the scanning raster of the television format and representing each sample by a digital code.

Pixels. The individual samples of digital data are referred to as picture elements, "pixels" or "pels." When the ratio of active pixels per line to active pixels per frame is the same as the aspect ratio, the format is said to have "square pixels." The term refers to the spacing of samples, not the shape of the pixel.

Blocks, Macroblocks, and Slices. For further processing, pixels are organized into 8 × 8 blocks, representing either luminance or chrominance information. Macroblocks consist of four blocks of luminance (Y) and one each of C_b and C_r. Slices consist of one or more macroblocks in the same row, and they begin with a slice start code. The number of slices affects compression efficiency; a larger number of slices provides for better error recovery but uses bits that could otherwise be used to improve picture quality. The slice is the minimum unit for resynchronization after an error.

Removal of Temporal Information Redundancy: Motion Estimation and Compensation. A video sequence is a series of still pictures shown in rapid succession to give the impression of continuous motion. This usually results in much temporal redundancy (picture sameness) among adjacent pictures. Motion compensation attempts to delete this temporal redundancy from the information transmitted. In the standard, the current picture is predicted from the previously encoded picture by estimating the motion between the two adjacent pictures and compensating for the motion. This "motion-compensated residual" is encoded rather than the complete picture and eliminates repetition of the redundant information.

Pictures, Groups of Pictures, and Sequences. The primary coding unit of a video sequence is the individual video frame or picture, which consists of the collection of slices constituting the active picture area. A video sequence consists of one or more consecutive pictures, and it commences with a sequence header that can serve as an entry point.

One or more pictures or frames in sequence may be combined into a group of pictures (GOP), optional within MPEG-2 and the ATSC Standard, to provide boundaries for interpicture coding and registration of a time code.

Figure 27 illustrates a time sequence of video frames consisting of intracoded pictures (I-frames), predictive

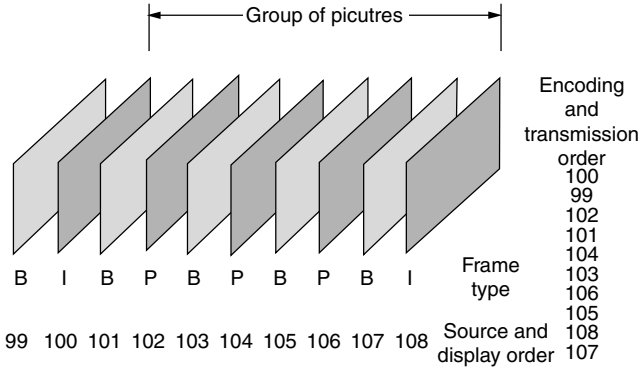


Figure 27. Video frame order, group of pictures, and typical I-frames, P-frames, and B-frames.

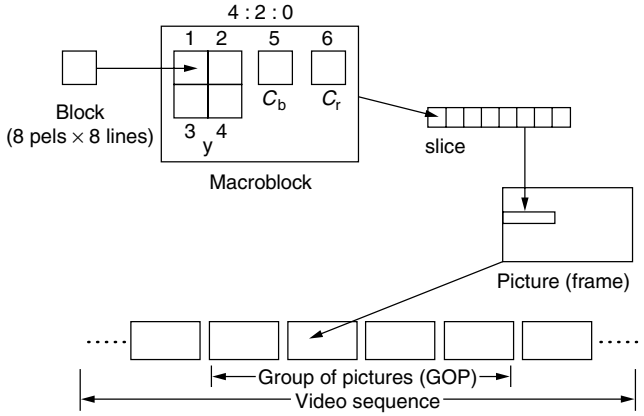


Figure 28. Video structure hierarchy.

coded pictures (P-frames), and bidirectionally predictive coded pictures (B-frames).

I-, P-, and B-Frames. Frames that do not use any interframe coding are referred to as I-frames (where I denotes intraframe coded). All of the information for a complete image is contained within an I-frame, and the image can be displayed without reference to any other frame. (The preceding frames may not be present or complete for initialization or acquisition, and the preceding or following frames may not be present or complete when noncorrectable channel errors occur.)

P-frames (where P denotes predicted) are frames where the temporal prediction is only in the forward direction (formed only from pixels in the most recently decoded I- or P-frame). Interframe coding techniques improve the overall compression efficiency and picture quality. P-frames may include portions that are only intraframe-coded.

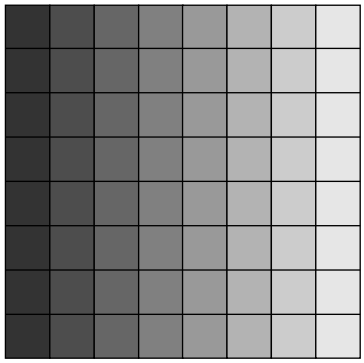
B-frames (where B denotes bidirectionally predicted) include prediction from a future frame as well as from a previous frame (always I- or P-frames). Some of the consequences of using future frames in the prediction are as follows: The transmission order of frames is different from the displayed order of frames, and the encoder and decoder must reorder the video frames, thus increasing the total latency. B-frames are used for increasing compression efficiency and perceived picture quality.

Figure 28 illustrates the components of pictures, as discussed before.

Removal of Spatial Information Redundancy: The Discrete Cosine Transform. As shown in Fig. 29, 8×8 blocks of spatial intensity that show variations of luminance and chrominance pel information are converted into 8×8 arrays of coefficients relating to the spatial frequency content of the original intensity information. The transformation method used is the discrete cosine transform (DCT).

As an example, in Fig. 29a, an 8×8 pel array representing a black-to-white transition is shown as increasing levels of a gray scale. In Fig. 29b, the gray-scale steps have been digitized and are represented by pel amplitude numerical values. In Fig. 29c, the gray-scale block is represented by its frequency transformation

(a)



(b)

0	12.5	25	37.5	50	62.5	75	87.5
0	12.5	25	37.5	50	62.5	75	87.5
0	12.5	25	37.5	50	62.5	75	87.5
0	12.5	25	37.5	50	62.5	75	87.5
0	12.5	25	37.5	50	62.5	75	87.5
0	12.5	25	37.5	50	62.5	75	87.5
0	12.5	25	37.5	50	62.5	75	87.5
0	12.5	25	37.5	50	62.5	75	87.5

(c) DCT-DC component

43.8	–40	0	–4.1	0	–1.1	0	0
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0

Figure 29. Discrete cosine transform.

coefficients, appropriately scaled. The DCT compacts most of the energy into only a small number of the transform coefficients. To achieve a higher decorrelation of the picture content, two-dimensional (along two axes) DCT coding is applied. The (0,0) array position (top left), represents the DC coefficient or average value of the array.

Quantizing the Coefficients. The goal of video compression is to maximize the video quality for a given bit rate. Quantization is a process of dividing the coefficients by a value of N which is greater than 1, and rounding the answer to the nearest integral value. This allows scaling the coefficient values according to their importance in the overall image. Thus high-resolution detail to which the human eye is less sensitive may be more heavily scaled (coarsely coded). The quantizer may also include a dead zone (enlarged interval around zero) to core to zero small noise-like perturbations of the element value. Quantization in the compression algorithm is a lossy step (information is discarded that cannot be recovered).

Variable Length Coding, Codeword Assignment. The quantized values could be represented using fixed-length codewords. However, greater efficiency can be achieved in bit rate by employing what is known as entropy coding. This attempts to exploit the statistical properties of the signal to be encoded. It is possible to assign a shorter code word to those values that occur more frequently and a longer code word to those that occur less frequently. The Morse code is an example of this method. One optimal code word design method, the Huffman code, is used in the Standard. Note that many zero value coefficients are produced, and these may be prioritized into long runs of zeros by zigzag scanning or a similar method.

Channel Buffer. Motion compensation, adaptive quantization, and variable-length coding produce highly variable amounts of compressed video data as a function of time. A buffer is used to regulate the variable-input bit rate into a fixed-output bit rate for transmission. The fullness of the buffer is controlled by adjusting the amount of quantization error in each image block (a rate controller driven by a buffer state sensor adjusts the quantization level). Buffer size is constrained by maximum tolerable delay through the system and by cost.

Audio

System Overview. The audio subsystem used in the ATSC Digital Television Standard is based on the AC-3 digital audio compression standard. The subsystem can encode from one to six channels of source audio from a pulse-code modulation (PCM) representation (requiring 5.184 Mbps for the 5.1 channel mode) into a serial bit stream at a normal data rate of 384 kbps. The 5.1 channels are left (front), center (front), right (front), left surround (rear), right surround (rear) (all 3 Hz to 20 kHz), and low-frequency subwoofer (normally placed centrally) (which represents the 0.1 channel, 3 Hz to 120 Hz). The system conveys digital audio sampled at a frequency of 48 kHz, locked to the 27 MHz system clock.

In addition to the 5.1 channel input, monophonic and stereophonic inputs and outputs can be handled. Monophonic and stereophonic outputs can also be derived from a 5.1 channel input, permitting backward compatibility.

The audio subsystem, illustrated in Fig. 30, comprises the audio encoding/decoding function and resides between the audio inputs/outputs and the transport system. The audio encoder(s) is (are) responsible for generating the *audio elementary stream(s)*, which are encoded representations of the baseband audio input signals.

The transport subsystem packetizes the audio data into PES packets which are then further packetized into (fixed-length) transport packets. The transmission subsystem converts the transport packets into a modulated RF signal for transmission to the receiver. Transport system flexibility allows transmitting multiple audio elementary streams. The encoding, packetization, and modulation process is reversed in the receiver to produce reconstructed audio.

Audio Compression. Two mechanisms are available for reducing the bit rate of sound signals. The first uses statistical correlation to remove redundancy from the bit stream. The second uses the psychoacoustical characteristics of the human hearing system such as spectral and temporal masking to reduce the number of bits required to recreate the original sounds.

The audio compression system consists of three basic operations, as shown in Fig. 31. In the first stage, the representation of the audio signal is changed from the time domain to the frequency domain, which is more efficient, to perform psychoacoustical audio compression. The frequency domain coefficients may be coarsely quantized because the resulting quantizing noise will be at the same frequency as the audio signal, and relatively low signal-to-noise ratios (SNRs) are acceptable due to the phenomena of psychoacoustic masking. The bit allocation operation determines the actual SNR acceptable for each individual frequency coefficient. Finally, the frequency coefficients are coarsely quantized to the necessary precision and formatted into the audio elementary stream.

The basic unit of encoded audio is the AC-3 sync frame, which represents six audio blocks of 256 frequency coefficient samples (derived from 512 time samples), a total of 1,536 samples. The AC-3 bit stream is a sequence of AC-3 sync frames.

Additional Audio Services. Additional features are provided by the AC-3 subsystem. These include loudness normalization, dynamic range compression that has an override for the listener, and several associated services; dialogue, commentary, emergency, voice-over, help for the visually impaired and hearing-impaired (captioning), and multiple languages. Some of these services are mutually exclusive, and multilanguage service requires up to an extra full 5.1 channel service for each language (up to an additional 384 kbps).

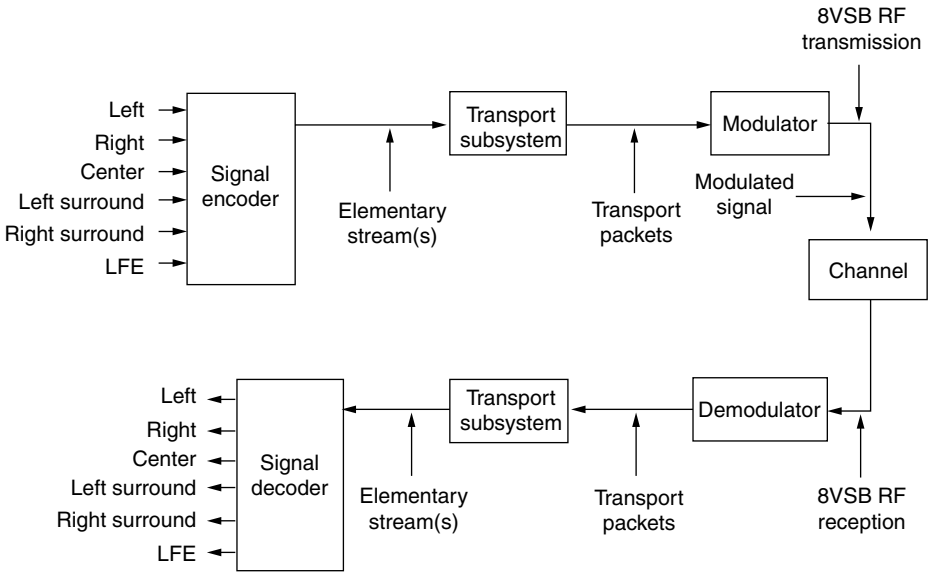


Figure 30. Audio subsystem within the digital television system.

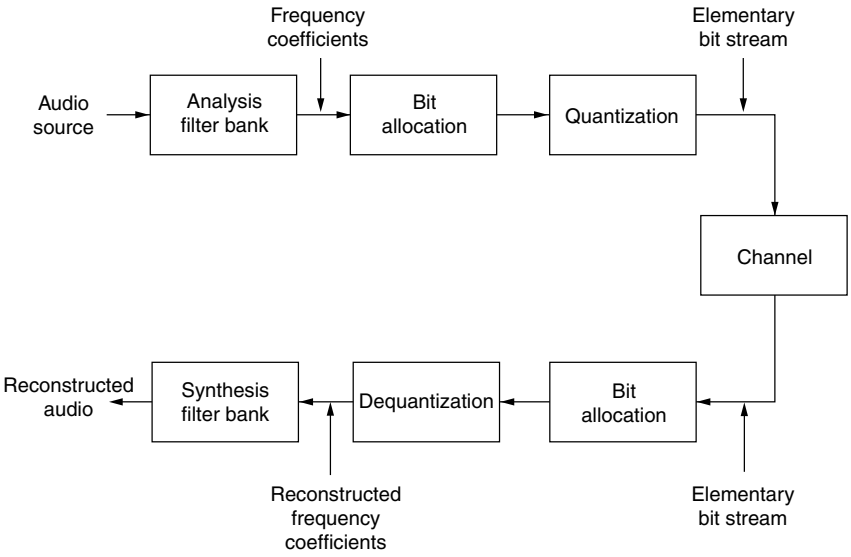


Figure 31. Overview of audio compression system.

Ancillary Data Services

Several data services have been included in the ATSC Standard. Other services can be added in the future. Currently, program subtitles (similar to closed captioning in NTSC), emergency messages (mixed into baseband video in NTSC), and program guide information are included.

Possible Future Data Services. Information data related to the following may be desired: conditional access, picture structure, colorimetry, scene changes, local program insertion, field/frame rate and film pull-down, pan/scan, multiprogram, and stereoscopic image.

Transmission Characteristics

The transmission subsystem uses a vestigial sideband (VSB) method: (1) 8-VSB for simulcast terrestrial broadcast mode and (2) a 16-VSB high data rate mode. VSB includes a small part of the lower sideband and the full upper sideband. Sloped filtering at the transmitter and/or the receiver attenuates the lower end of the band. The 8-VSB coding maps three bits into one of eight signal levels. The system uses a symbol rate of 10.76 Msymbols/s, capable of supporting a data stream payload of 19.39 MBits/s. See Fig. 32 VSB in a 6 MHz channel.

Modulation techniques for some other planned broadcast systems use orthogonal frequency division multiplexing (OFDM) or coded OFDM (COFDM), which is a form of multicarrier modulation where the carrier spacing is selected, so that each subcarrier within the channel is orthogonal to the other subcarriers; this mathematically ensures that during the sampling time for one carrier, all other carriers are at a zero point.

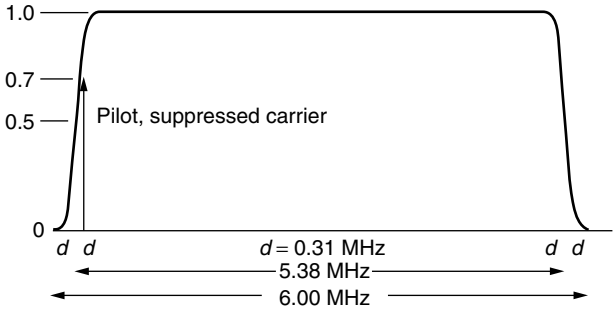


Figure 32. Vestigial sideband (VSB) in a 6 MHz channel for digital transmission.

The 8-VSB subsystem takes advantage of a pilot, segment sync, and a training sequence for robust acquisition and operation. To maximize service area, an NTSC rejection filter (in the receiver) and trellis coding are used. The system can operate in a signal-to-additive-white-Gaussian noise (S/N) environment of 14.9 dB. The transient peak power to average power ratio measured on a low-power transmitted signal that has no nonlinearities is no more than 6.3 dB for 99.9% of the time.

A block diagram of a generic transmitter subsystem is shown in Fig. 33. The incoming data (19.39 MB/s) are randomized and then processed for forward error correction (FEC) in the form of Reed–Solomon coding (20 RS parity bits are added to each packet, known as outer error correction). Data interleaving to reorganize the data stream so that it is less vulnerable to bursts of errors, then interleaves to a depth of about 1/6 data field (4 ms deep).

The second stage, called inner error correction, consists of a 2/3 rate trellis coding. This encodes one bit of a two-bit pair into two output bits, using a 1/2 convolutional code, whereas the other input bit is retained as precoded. Along with the trellis encoder, the data packets are precoded into data frames and mapped into a signaling waveform using an eight-level (3 bit), one-dimensional constellation (8 VSB). Data segment sync (4 symbols = 1 byte) at the beginning of a segment of 828 data plus parity symbols, and data field sync at the beginning of a data field of 313 segments (24.2 ms), are then added. data field sync includes the training signal used for setting the receiver equalizer.

A small in-phase pilot is then added to the data signal at a power of 11.3 dB below the average data signal power. The data are then modulated onto an IF carrier, which is the same frequency for all channels. The RF up-converter then translates the filtered, flat IF data signal spectrum to the desired RF channel. It is then amplified to the appropriate power for the transmitting antenna.

For the same approximate coverage as an NTSC transmitter (at the same frequency), the average power of the ATV signal is approximately 12 dB less than the NTSC peak sync power.

The frequency of the RF up-converter oscillator will typically be the same as that for NTSC (except for offsets). For extreme cochannel situations, precise RF carrier frequency offsets with respect to the NTSC cochannel carrier may be used to reduce interference into the ATV signal. The ATV signal is noise-like, and its interference into NTSC does not change with precise offset.

The ATV cochannel pilot should be offset in the RF up-converter from the dominant NTSC picture carrier by an odd multiple of half the data segment rate. An additional offset of 0, +10 kHz, or −10 kHz is required to track the principal NTSC interferer.

For ATV-into-ATV cochannel interference, precise carrier offset prevents the adaptive equalizer from misinterpreting the interference as a ghost.

The Japanese High-Definition Television Production System

This television production system was developed by the Japanese Broadcasting Corporation (NHK). It was standardized in 1987 by the Broadcast Technology Association (BTA), now renamed the Association of Radio Industries and Business (ARIB), in Japan and in the United States by SMPTE (240M and 260M Standards). It uses a total of 1,125 lines (1,035 active lines), is interlaced at a field rate of 60 Hz, and has an aspect ratio of 16 : 9. It requires a bandwidth of 30 MHz for the luminance signal (Y), and 15 MHz for each of the two color difference signals P_B and P_R When digitized at eight bits per sample, it uses 1,920 pixels per line, and it requires a total bit rate of 1.2 Gbps. Note that this production system is similar to the interlaced system used in the ATSC standard, except that the latter uses 1,080 active lines.

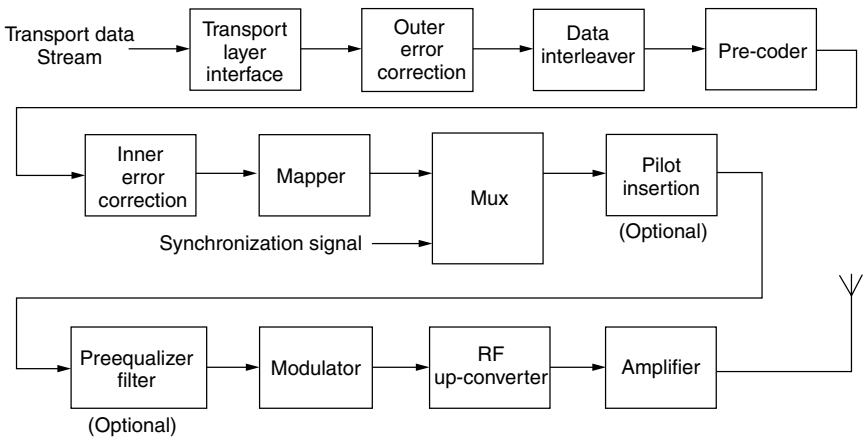


Figure 33. 8-VSB transmitter subsystem block diagram.

Japanese MUSE Transmission Systems

A range of transmission systems were developed by NHK based on the multiple sub-Nyquist encoding (MUSE) transmission scheme (see Table 8). MUSE (8.1 MHz bandwidth) was developed for DBS broadcasting and MUSE-T (16.2 MHz bandwidth) was developed for satellite transmission. MUSE-6 was designed to be compatible with a 6 MHz channel and NTSC receivers. MUSE-9 uses a 3 MHz augmentation channel in addition to the standard 6 MHz channel and is NTSC receiver-compatible.

Japanese Hi-Vision System

This system incorporates the $1{,}920 \times 1{,}035$ television production system and the MUSE-E transmission system. MUSE-E uses an 8.1 MHz bandwidth and is incompatible with standard NTSC receivers and channel allocations. Four audio channels are time-division-multiplexed with the video signals in the blanking intervals. The encoding and decoding processes are both very complex and require many very large scale integration (VLSI) chips. This system requires a MUSE-E receiver, or a set-top box equipped with a MUSE decoder that feeds either a 16 : 9 display or a 4 : 3 aspect ratio conventional receiver. In the near-term, NHK will use simultaneous Hi-Vision/NTSC program production.

The MUSE systems are not receiver-compatible with either the North-American ATSC system or the European DVB system (see later).

The Japanese Enhanced Definition Television System (EDTV-II)

EDTV-II is an NTSC-compatible letterbox analog transmission system standardized by the ARIB in Japan. The input signal is a 525-line, 60-frame progressive scan (525P) that has a 16 : 9 aspect ratio. A 525-line, 30-frame interlaced scan (525I) can be up-converted as an input signal. Note that the 525P signal is one of the SDTV signal formats defined in the ATSC Standard (720×480 at 60 P). It is also defined as a production format in SMPTE 293M and SMPTE 294M standards documents.

Compared with the current 525I standard, the frame rate has been doubled from 30 to 60. The sampling frequency in the format has been doubled to 27 MHz, compared to 525I, and the aspect ratio has been changed from 4 : 3 to 16 : 9. This increase of sampling frequency permits maintaining comparable resolution in H and V axes. The production system is effectively an 8 : 4 : 4 digital system that has production interfaces at 540 Mbps. A 4 : 2 : 0 system can also be used in production and would require interfacing at 360 Mbps. Horizontal blanking is shrunk to achieve this bit rate.

The EDTV-II analog transmission system is used for both terrestrial and satellite broadcasting. It requires the same bandwidth as the NTSC system, and no changes are needed in transmitter implementations. The image is displayed on an EDTV-II receiver, progressively by 480 lines and a 16 : 9 aspect ratio. It is compatible with existing NTSC receivers, except that the display image has a 16 : 9 aspect ratio and so appears in a letterbox format that has black bars at top and bottom. The 525P signal requires a video bandwidth of approximately 6.2 MHz. The EDTV-II system creates three enhancement signals in addition to an NTSC signal, with which they are then frequency-domain-multiplexed.

Main Picture (MP). The 525P 16 : 9 signal is reduced from 6.2 MHz to 4.2 MHz bandwidth, and the 480 lines are decimated to 360 lines to produce a letterbox display on the NTSC 4 : 3 receiver. Black bars at top and bottom are each 60 lines wide. Thus, horizontal and vertical resolution are reduced to conform to the NTSC format, but to maintain the 16 : 9 aspect ratio.

Horizontal High (HH 4.2 MHz to 6.2 MHz). A frequency enhancement signal is extracted from the original 525P image and is multiplexed into the MP signal to increase the horizontal bandwidth to 6.2 MHz in the EDTV-II receiver. For transmission, the HH signal is downshifted to 2 to 4 MHz and frequency-division-multiplexed into an unused vertical temporal frequency domain in the conventional NTSC system called the Fukinuki hole. The Fukinuki hole may be used only for correlated video information, which applies in this case.

In the EDTV-II receiver, a motion detector multiplexes the HH signal only onto the still parts of the picture where there is more need for high resolution to satisfy human vision characteristics. Two enhancement signals are frequency-division-multiplexed together into the top

Table 8. MUSE Transmission Systems

Transmission System	Type of Transmission	Bandwidth	Channel Compatible	Compatible with NTSC
MUSE	Direct broadcast by satellite (DBS)	8.1 MHz	NA	No
MUSE-T	Satellite	16.2 MHz	NA	No
MUSE-6	Terrestrial broadcast	6 MHz	Yes	Yes
MUSE-9	Terrestrial broadcast	6 + 3 MHz Augmentation	Yes, with 2nd 3 MHz channel	Yes
MUSE-E	Terrestrial broadcast	8.1 MHz	No	No

and bottom panels, which together occupy one-third as much area as the main picture. As these are generated in a 360-line format, they must be compressed by a 3 to 1 pixel downsampling decimation process to fit into the 120 lines of the top and bottom panels.

Vertical High Frequency (VH). The VH signal enhances the vertical still picture resolution back up to 480 lines. The signal is transmitted only for stationary areas of the image, and temporal averaging is applied.

Vertical Temporal Frequency (VT). The VT enhancement signal is derived from the progressive-to-interlace scan conversion at the encoder and improves the interlace-to-progressive scan (360/2 : 1 to 360/1 : 1) conversion in the receiver.

The EDTV-II receiver performs the reverse of the encoding process. The NTSC receiver uses the MP signal directly.

The European DVB System

The Digital Video Broadcast (DVB) system has been designed for MPEG-2-based digital delivery systems for satellite, cable, community cable, multichannel multipoint distribution (MMDS), and terrestrial broadcasting. Service information, conditional access, and teletext functions are also available. All DVB systems are compatible.

DVB-T, the terrestrial broadcasting standard, is similar in many respects to the ATSC standard. However, there are a number of significant differences. DVB-T uses coded orthogonal frequency division multiplexing (COFDM). This technique is already being used for direct audio broadcast (DAB). Individual carriers 1,704 (2 k) or 6,816 (8 k) may be used. The 8-k system is more robust, but increases receiver complexity and cost. Some broadcasters have already adopted the 2-k system, although it will not be compatible with the 8-k system. DVB-T uses the MPEG-2 Layer II Musicam audio standard, a 50 Hz frame rate, and aspect ratios of 4 : 3, 16 : 9, or 20 : 9.

The European PALplus System

This is an analog delivery system that uses a current TV channel to transmit an enhanced wide-screen version of the PAL signal. A conventional receiver displays the PALplus picture as a letterbox in a 4 : 3 aspect ratio. A wide-screen receiver shows the same transmitted picture in a 16 : 9 format at higher resolution. European broadcasters are divided on whether to use this format. The PALplus concept is similar to the Japanese EDTV-II format described before.

Acknowledgments

The authors sincerely thank the following for permission to use portions of their work in this article:

- The Advanced Television Systems Committee (ATSC) and its Executive Director, Craig Tanner, for text and figures from Standards A/52, A/53, and A/54.
- Mr. Stanley N. Baron for text and figures from his book *Digital Image and Audio Communications, Toward a Global Information Infrastructure*.
- Mr. Patrick Griffis for the data in a figure from his article Bits = Bucks, Panasonic, paper presented at NAB, 1998, unpublished.

BIBLOGRAPHY

1. M. Ashibe and H. Honma, A wide-aspect NTSC compatible EDTV system, *J. SMPTE*, Mar. 1992, p. 130.
2. ATSC Digital Television Standard, Advanced Television Systems Committee, Doc. A/53, 16 Sept., 1995.
3. S. N. Baron, ed., *Composite Digital Television: A Primer*, Soc. of Motion Picture and Television Eng., White Plains, NY, 1996.
4. S. N. Baron and M. I. Krivocheev, Digital Image and Audio Communications, *Toward a Global Information Infrastructure*, Van Nostrand Reinhold, New York, 1996.
5. K. B. Benson, ed., *Television Engineering Handbook*, revised by J. Whitaker, McGraw-Hill, New York, 1992.
6. Digital Audio Compression Standard (AC-3), Advanced Television Systems Committee, Doc. A/52, 20 Dec., 1995.
7. A. Dubec, *The SECAM Colour Television System*, Compagnie Française de Télévision, Paris, 1976.
8. P. Griffis, Bits = Bucks, Panasonic, paper at NAB, 1998.
9. Guide to the Use of the ATSC Digital Television Standard, Advanced Television Systems Committee, Doc. A/54, 4 Oct., 1995.
10. A. Itoh, 525 line progressive scan signal digital interface standard and system, *J. SMPTE*, Nov. 1997, p. 768.
11. R. W. G. Hunt, *The Reproduction of Colour*, 5th ed., Fountain Press, Kingston-upon-Thames, 1995.
12. G. Hutson, P. Shepherd, and J. Brice, *Colour Television*, McGraw-Hill, London, 1990.
13. A. F. Inglis and A. C. Luther, *Video Engineering*, 2nd ed., McGraw-Hill, New York, 1996.
14. ISO/IEC IS 13818-1, International Standard MPEG-2 Systems, 1994.
15. ISO/IEC IS 13818-2, International Standard MPEG-2 Video, 1994.
16. ISO/IEC IS 13818-2, Section 8.
17. ITU-R BT.470-4, Characteristics of television systems, International Telecommunications Union, Geneva, 1995.
18. ITU-R Document, 11-3/15, MPEG digital compression systems, 9 Aug., 1994.
19. K. Jack, *Video Demystified*, 2nd ed., HighText Interactive, San Diego, 1996.
20. K. Jackson and B. Townsend, eds., *TV & Video Engineer's Reference Book*, Butterworth-Heinemann, Oxford, England, 1991.
21. H. Y. Kim and S. Naimpally, Digital EDTV, compatible HDTV, *J. SMPTE*, Feb. 1993, p. 119.
22. B. Marti et al., Problems and perspectives of digital terrestrial TV in Europe, *J. SMPTE*, Aug. 1993, p. 703.
23. R. Mäusl, *Refresher Topics — Television Technology*, Rohde & Schwarz, Munich, 1992.
24. R. S. O'Brien, ed., *Color Television, Selections from the Journal of the SMPTE*, Society of Motion Picture and Television Engineers, New York, 1970.
25. G. Pensinger, ed., *4 : 2 : 2 Digital Video Background and Implementation*, Society of Motion Picture and Television Engineers, White Plains, NY, 1989.
26. D. H. Pritchard and J. J. Gibson, Worldwide color television standards — similarities and differences, *J. Soc. Motion Picture and Television Eng.* **89**, 111–120 (1980).

27. *Proc. IRE, Color Television Issue* **39**(10) (1951).
28. V. Reimer, Advanced TV systems, Germany and Central Europe, *J. SMPTE*, May 1993, p. 398.
29. M. Robin, Addendum to Worldwide color television standards — similarities and differences, *J. Soc. Motion Picture and Television Eng.* **89**, 948–949 (1980).
30. M. Robin and M. Poulin, *Digital Television Fundamentals*, McGraw-Hill, New York, 1997.
31. T. Rzeszewski, ed., *Color Television*, IEEE Press, New York, 1983.
32. T. S. Rzeszewski, ed., *Television Technology Today*, IEEE Press, New York, 1984.
33. H. V. Sims, *Principles of PAL Colour Television and Related Systems*, Newnes Technical Books, London, 1969.
34. SMPTE 274 M Standard for Television, $1,920 \times 1,080$ Scanning and Interface, 1995.
35. SMPTE S17.392 Proposed Standard for Television, $1,280 \times 720$ Scanning and Interface, 1995.
36. V. Steinberg, *Video Standards, Signals, Formats, and Interfaces*, Snell & Wilcox, Durford Mill, England, 1997.
37. N. Suzuki et al., Matrix conversion VT resolution in letterbox, *J. SMPTE*, Feb. 1991, p. 104.
38. N. Suzuki et al., Experiments on proposed multiplexing scheme for vertical-temporal and vertical high helper signals in EDTV-II, *J. SMPTE*, Nov. 1994, p. 728.
39. Television Operating and Interface Standards, Society of Motion Picture and Television Engineers, 595 W. Hartsdale Ave., New York, NY 10607-1824.
40. Television Measurements Standards, Institute of Electrical and Electronic Engineers, Inc., Broadcast Technology Society, c/o IEEE Service Center, 445 Hoes Lane, Box 1,331, Piscataway, NJ 08855.
41. J. Watkinson, *Television Fundamentals*, Focal Press, Oxford, England, 1996.

TERAHERTZ ELECTRIC FIELD IMAGING

X. -C. ZHANG
Rensselaer Polytechnic Institute
Troy, NY

INTRODUCTION TO THE TERAHERTZ WAVE

Various frequencies are spaced along the frequently used electromagnetic spectrum, including microwaves, infrared, visible light, and X rays. Terahertz radiation between microwave and infrared frequencies lies (Fig. 1). In the electromagnetic spectrum, radiation at 1 THz has a period of 1 ps, a wavelength of 300 μm, a wave number of 33 cm^{-1}, photon energy of 4.1 meV, and an equivalent temperature of 47.6 K. In the same way that visible light can create a photograph, radio waves can transmit sound, and X rays can see shapes within the human body, terahertz waves (Trays) can create pictures and transmit information.

Until recently, however, the very large terahertz portion of the spectrum has not been particularly useful because there were neither suitable emitters to send out controlled terahertz signals nor efficient sensors to collect them and record information. Recent developments in terahertz time-domain spectroscopy and related terahertz technologies now lead us to view the world in a new way. As a result of developing research, terahertz radiation now has widespread potential applications in medicine, microelectronics, agriculture, forensic science, and many other fields.

Three properties of THz wave radiation triggered research to develop this frequency band for applications:

- Terahertz waves have low photon energies (4 meV at 1 THz) and thus cannot lead to photoionization in biological tissues.
- Many molecules exhibit strong absorption and dispersion rates at terahertz frequencies, due to dipole allowed rotational and vibrational transitions. These transitions are specific to the molecule and therefore enable terahertz wave fingerprinting.
- Coherent terahertz wave signals can be detected in the time domain by mapping the transient of the electrical field in amplitude and phase. This gives access to absorption and dispersion spectroscopy.

Coherent terahertz time-domain spectroscopy that has an ultrawide bandwidth provides a new method for characterizing the electronic, vibronic, and compositional properties of solid, liquid, and gas phase materials, as well as flames and flows. In theory, many biological and chemical compounds have distinct signature responses to terahertz waves due to their unique molecular vibrations and rotational energy levels, this implies that their chemical compositions might be examined

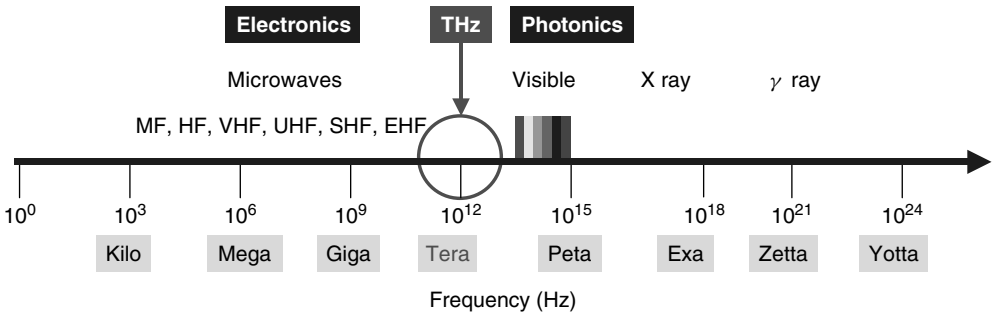


Figure 1. The terahertz gap: a scientifically rich but technologically limited frequency band between microwave and optical frequencies.

using a terahertz beam. Such capability could be used diagnosing disease, detecting pollutants, sensing biological and chemical agents, and quality control of food products. It is also quite possible that plastic explosives could be distinguished from suitcases, clothing, common household materials, and equipment based on molecular structure. Detecting the binding state of genetic materials (DNA and RNA) by directly using terahertz waves, without requiring markers, provides a label-free method for genetic analysis for future biochip technologies.

A T-ray imaging modality would produce images that have "component contrast" enabling analysis of the water content and composition of tissues in biological samples. Such capability presents tremendous potential to identify early changes in composition and function as a precursor to specific medical investigations and treatment. Moreover, in conventional optical transillumination techniques that use near-infrared pulses, large amounts of scattering can spatially smear out the objects to be imaged. T-ray imaging techniques, due to their longer wavelengths, can provide significantly enhanced contrast as a result of low scattering (Rayleigh scattering).

TECHNICAL BACKGROUND

The development of terahertz time-domain spectroscopy has recently stimulated applications of this unexplored frequency band. Hu and Nuss first applied THz pulses to imaging applications (1). Far-infrared images (T-ray imaging) of tree leaves, bacon, and semiconductor integrated chips have been demonstrated.

In an imaging system that has a single terahertz antenna, the image is obtained by pixel-scanning the sample in two dimensions (2–4). As a result, the time for acquiring an image is typically of the order of minutes or hours, depending on the total number of pixels and the lowest terahertz frequency components of interest. Although it is highly desirable to improve the data acquisition rate further for real-time imaging by fabricating a focal plane antenna array, technical issues such as high optical power consumption and limits on the antenna packaging density would hinder such a device (5).

Recently, a free-space electro-optic sampling system has been used to characterize the temporal and 2-D spatial distribution of pulsed electromagnetic radiation (6–8). A T ray can be reflected as a quasi-optical beam, collimated by metallic mirrors, and focused by a plastic or high-resistivity silicon lens. The typical powers of a T-ray sensing (single pixel) and an imaging (2-D array) system are microwatts and milliwatts, respectively. Thus a terahertz imaging system, based on the electro-optic sampling technique, shows promise for 2-D real-time frame imaging using terahertz beams.

GENERATION OF TERAHERTZ BEAMS

Currently, photoconduction and optical rectification are the two basic approaches for generating terahertz beams by using ultrafast laser pulses. The photoconductive approach uses high-speed photoconductors as transient current sources for radiating antennas (9). These antennas include elementary hertzian dipoles, resonant dipoles, tapered antennas, transmission lines, and large-aperture photoconducting antennas. The optical rectification approach uses electro-optic crystals as a rectification medium (10). The rectification can be a second-order (difference frequency generation) or higher order nonlinear optical process, depending on the optical fluency.

The physical mechanism for generating a terahertz beam by photoconductive antennas is the following: a laser pulse ($h\omega \geq E_g$) creates electron–hole pairs in the photoconductor, the free carriers then accelerate in the static field to form a transient photocurrent, and the fast time-varying current radiates electromagnetic waves. In the far field, the electrical component of terahertz radiation is proportional to the first time derivative of the photocurrent. The waveform is measured by a 100-μm photoconducting dipole that can resolve subpicosecond electrical transients. Because the radiating energy comes mainly from stored surface energy, the terahertz radiation energy can scale up with the bias and optical fluency (11).

Optical rectification is the inverse process of the electro-optic effect (12). In contrast to photoconducting elements where the optical beam functions as a trigger, the energy of terahertz radiation during transient optical rectification comes from the excitatory laser pulse. The conversion efficiency depends on the value of the nonlinear coefficient and the phase matching condition. In the optical rectification mode, the terahertz pulse duration is comparable to the optical pulse duration, and the frequency spectrum is limited mainly by the spectral broadening of the laser pulse, as determined by the uncertainty principle. Materials used for terahertz sources have been adapted from conventional electro-optic crystals and include semiconductor and organic crystals. Enhancement and sign change of second-order susceptibility by optically exciting electronic resonance states have been reported (13).

FREE-SPACE ELECTRO-OPTIC DETECTION

Fundamentally, the electro-optic effect is a coupling between a low-frequency electrical field (terahertz pulse) and a laser beam (optical pulse) in the sensor crystal. Free-space electro-optic sampling via the linear electro-optic effect (Pockels effect) offers a flat frequency response across an ultrawide bandwidth. Because field detection is purely an electro-optic process, the system bandwidth is limited mainly by either the pulse duration of the probe laser or the lowest transverse optical (TO) phonon frequency of the sensor crystal. Furthermore, because electro-optic sampling is purely an optical technique, it does not require electrode contact or wiring on the sensor crystal (14,15).

Figure 2 is a schematic of the experimental setup for using optical rectification and electro-optic effects. Nonlinear optics forms the basis of the terahertz system. A

mode-locked Ti:sapphire laser is used as the optical source. Several different gigahertz/terahertz emitters can be used, including photoconductive antennas (transient current source) and a ⟨111⟩ GaAs wafer at normal incidence (optical rectification source) (16–18). Generally, an optical rectification source emits terahertz pulses whose duration is comparable to that of the optical excitatory pulse, and a transient current source radiates longer terahertz pulses.

Figure 3 shows the details of the sampling setup. Simple tensor analysis indicates that using a ⟨110⟩ oriented zincblende crystal as a sensor gives the best sensitivity. The polarization of the terahertz beam and optical probe beam are parallel to the [1,−1,0] crystal direction. Modulating the birefringence of the sensor crystal via an applied electrical field (terahertz) modulates the polarization ellipticity of the optical probe beam that passes through the crystal. The ellipticity modulation of the optical beam can then be analyzed for polarization to provide information on both the amplitude and phase of the applied electrical field. The detection system will analyze a polarization change from the electro-optic crystal and correlate it with the amplitude and phase of the electrical test field. For weak field detection, the power of the laser beam $P_{\text{out}}(E)$ modulated by the electrical field of the terahertz pulse ($E = V/d$) is

$$P_{\text{out}}(E) = P_0\left[1 + \left(\frac{\pi E}{E_\pi}\right)\right], \tag{1}$$

where P_0 is the output optical probe power at zero applied field and E_π is the half-wave field of the sensor crystal of certain thickness. By measuring P_{out} from a calibrated voltage source as a function of time delay between the terahertz pulse and optical probe pulse, the time-resolved sign and amplitude of V can be obtained, and a numerical FFT provides frequency information. For a 3-mm thick ZnTe sensor crystal, the shot-noise limit gives a minimum detectable field of 100 $\text{nVcm}^{-1}\,\text{Hz}^{1/2}$ and a frequency range from near dc to 4 THz.

Figure 4 is a plot of the temporal electro-optic waveform of a terahertz pulse whose a half-cycle duration is 1 ps, as measured by a balanced detector using a ⟨110⟩ ZnTe sensor crystal. The time delay is provided by changing the relative length of the optical beam path between the terahertz pulses and the optical probe pulses. Detection sensitivity is significantly improved by increasing the interactive length of the pulsed field and the optical probe beam within the crystal. The dynamic range can exceed $10,000:1$ using unfocused beams, $100,000:1$ using unamplified focused beams, and $5,000,000:1$ using focused amplified beams and a ZnTe sensor crystal. Figure 5 is a plot of the signal and noise spectra, where the a SNR $> 50,000$ from 0.1 to 1.2 terahertz, corresponding to the waveform in Fig. 4.

A linear response in both generating and detecting of the terahertz pulses is crucial. Figure 6 is a plot of the electro-optic signal versus peak terahertz field strength. Excellent linearity is achieved. By increasing the optically illuminated area of the photoconductor on the terahertz emitter, the total emitted terahertz power scales linearly with the illumination area (assuming a nonsaturating optical fluence).

A shorter laser pulse (<15 fs) and a thinner sensor (<30 μm) allow us to enter the mid-IR range. Figure 7 shows a typical mid-IR waveform. A 30-μm and a 27-μm thick ZnTe crystal are used as an emitter and a sensor,

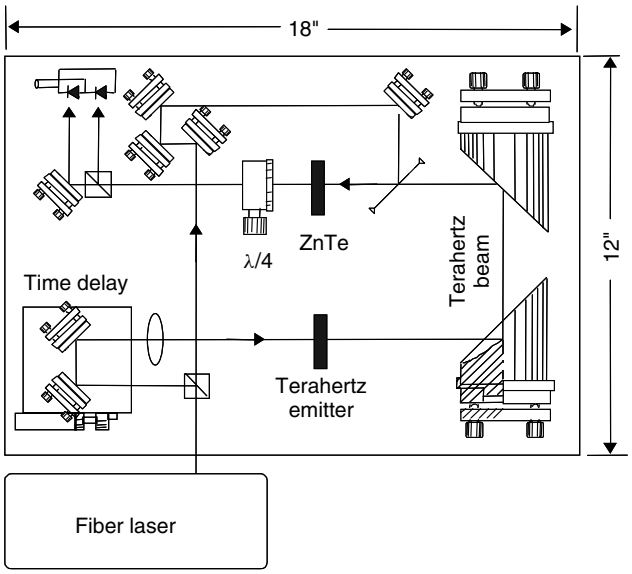


Figure 2. A simplified terahertz system.

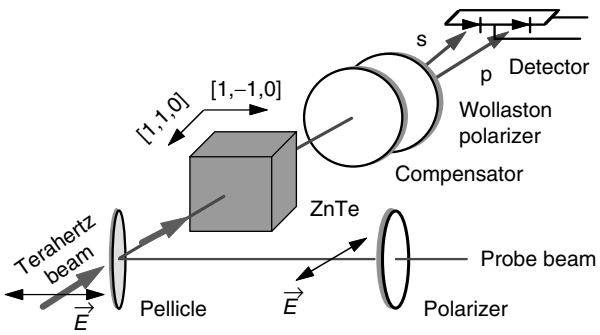


Figure 3. Details of a ZnTe sensor setup.

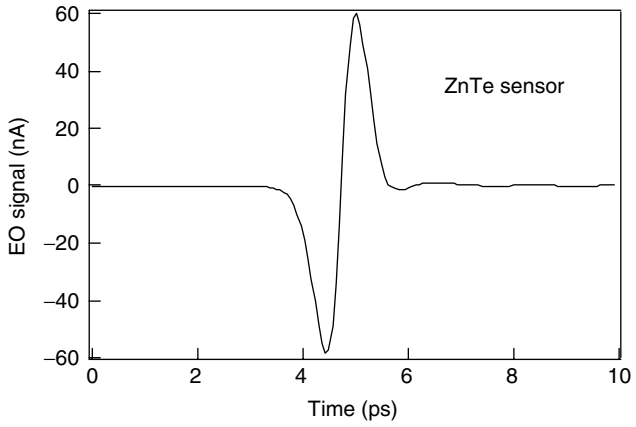


Figure 4. Temporal electro-optic signal of a 1-ps terahertz pulse measured by a ZnTe sensor.

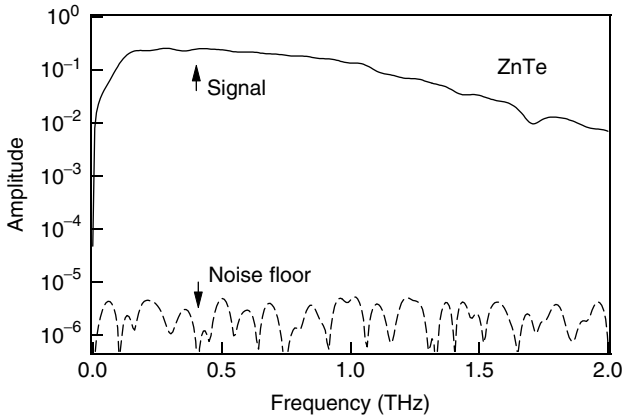


Figure 5. Amplitude spectrum of the waveform in Fig. 4 that shows a SNR >100,000 from 0.1 to 1.2 THz.

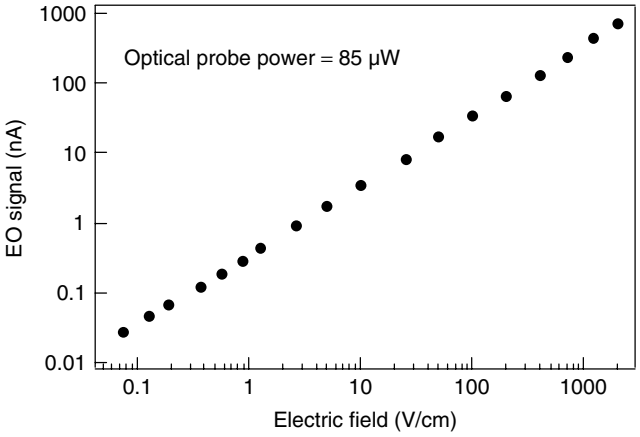


Figure 6. Electro-optic signal versus terahertz field.

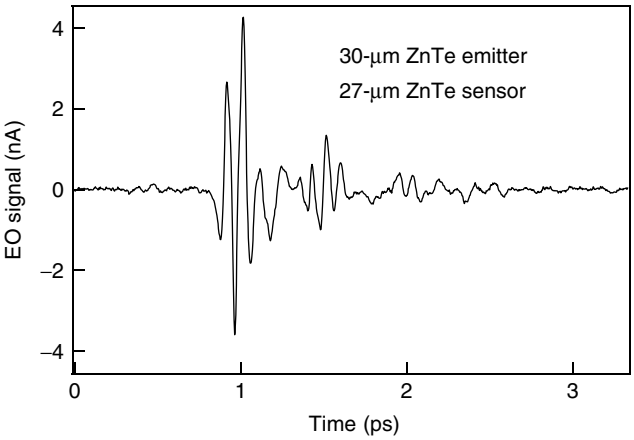


Figure 7. Temporal waveform of the terahertz radiation measured by a 30-μm ZnTe sensor. The shortest oscillation period is 31 fs (corresponding to 8 μm).

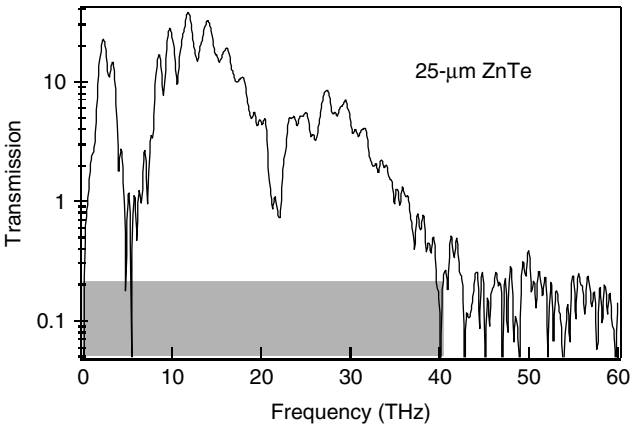


Figure 8. Frequency spectrum of the terahertz field in Fig. 7. The absorption dips at 5.3 THz are due to the phonon band. By using GaP, this dip can be moved to more than 10 THz.

respectively. The shortest period of the oscillation is 31 fs. The Fourier transform of the mid-IR pulse after it passes through a GaAs wafer is shown in Fig. 8, where the highest frequency response reaches more than 30 terahertz. The dip at 5.3 terahertz is due to phonon absorption by the ZnTe crystals. The preliminary result demonstrates the advantages of using the electro-optic sampling technique for measuring of ultrafast far-infrared to mid-infrared pulsed electromagnetic radiation. Current research has demonstrated that by using a thin ZnTe crystal (of the order of 10 μm), a detection bandwidth of more than 40 THz is feasible (19,20).

ELECTRO-OPTIC MEASUREMENT USING A CHIRPED PROBE PULSE

Using a linearly chirped optical probe pulse in electro-optic measurements, a temporal waveform of an electric field is linearly encoded onto the frequency spectrum of the probe pulse, and then decoded by dispersing the probe pulse from a grating to a detector array (21). Real-time and single-shot acquisition of picosecond terahertz field pulses is achieved without using a mechanical time-delay device (22).

Free-space terahertz electro-optic detection employing a chirped pulse technique may provide a solution for measuring freely propagating electromagnetic pulses (23) at an unprecedented acquisition time. A terahertz imaging system using a free-space electro-optic field sensor provides the capability of real-time and single-shot recording (24,25).

Figure 9 schematically illustrates the concepts of electro-optic measurement using the standard sampling method and a chirped optical probe beam. Figure 10 shows chirped pulse measurement. The geometry is similar to the conventional free-space electro-optic sampling setup, except for the use of a grating pair for chirping and stretching the optical probe beam and a grating–lens combination using a detector array for measuring the spectral distribution. Due to the negative chirp of the grating (pulse has decreasing frequency versus time), the blue component of the pulse leads the red component. The fixed delay line is used only to position the terahertz pulse within the duration of the synchronized probe pulse (acquisition window) and for temporal calibration.

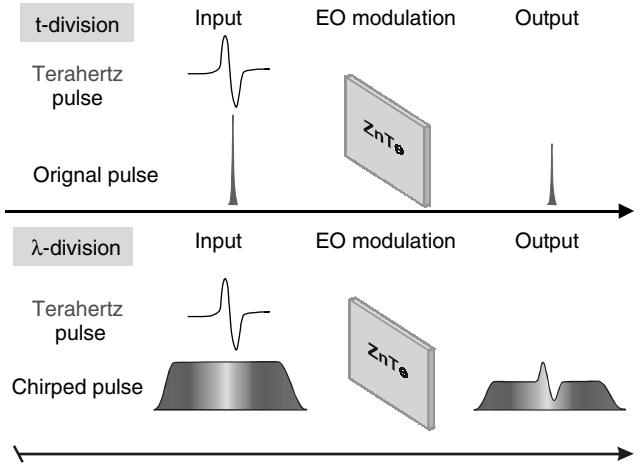


Figure 9. Schematics for (top) time and (bottom) wavelength division multiplexing and demultiplexing of the terahertz pulse. The wavelength division technique uses a chirped pulse.

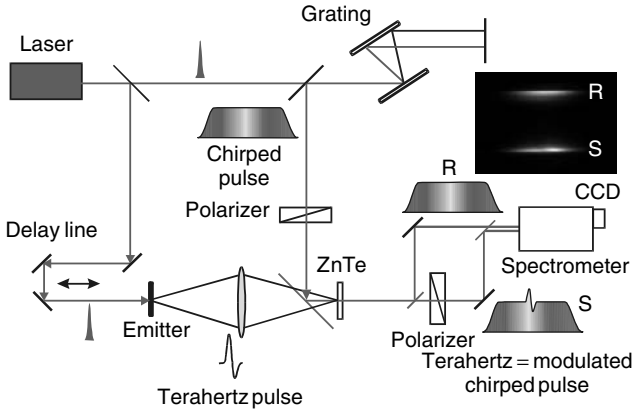


Figure 10. Schematic experimental setup of electro-optic measurement using a chirped optical probe beam. Images of the chirped probe pulses; S and R denote image traces of the signal and reference beam, respectively.

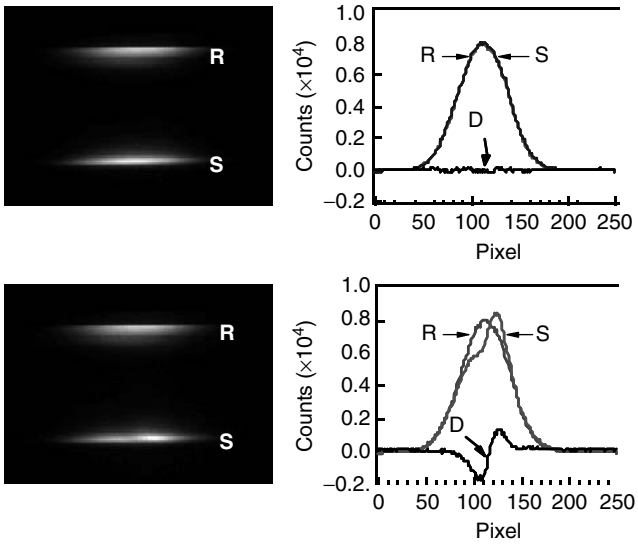


Figure 11. (Left) Imaging traces with/without a terahertz signal. (Right) 2-D plots of spectral distribution reconstructing the temporal profile of the terahertz signal.

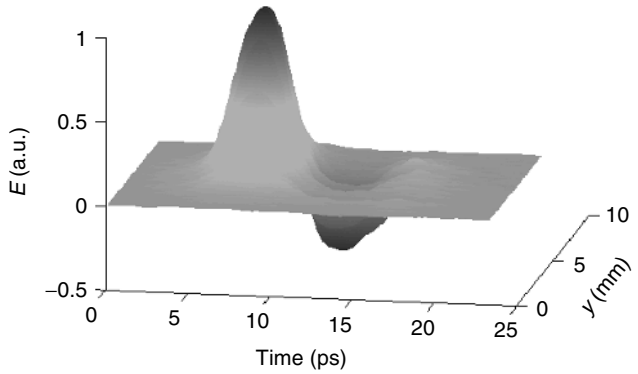


Figure 12. A *single-shot* spatiotemporal (10 mm–25 ps) image of a terahertz pulse measured by a chirped optical probe pulse. The emitter is a 7-mm dipole antenna.

The temporal terahertz signal can be extracted by measuring the difference between the spectral distributions of the probe pulse with and without terahertz pulse modulation. Figure 11 shows the spectral distributions of the chirped probe pulse measured by the CCD camera with and without a terahertz signal. The normalized distribution reconstructs both the amplitude and phase of the temporal waveform of the terahertz pulse.

Electro-optic measurement using a chirped optical probe pulse is capable of single-shot measurement of a full terahertz waveform. Figure 12 is a plot of a *single-shot* image from a GaAs photoconductive dipole antenna. The total time for wavelength division multiplexing and demultiplexing is a few picoseconds. The dipole length is 7 mm, and the bias voltage is 5 kV. One-dimensional spatial distribution across the dipole and its temporal terahertz waveform are obtained simultaneously in a single laser pulse. The size of the spatio temporal image is 10 mm by 25 ps. Typical oscillatory features and the symmetrical spatial distribution of the far-field pattern are obtained from a dipole photoconductive emitter.

TERAHERTZ IMAGING

One of the most advantageous applications of free-space electro-optic detection is two-dimensional terahertz wave imaging (26–30). Figure 13 schematically illustrates an experimental arrangement for free-space electro-optic terahertz imaging. Silicon lenses are used to focus the terahertz radiation on a ⟨110⟩ oriented ZnTe crystal. An optical readout beam whose diameter is larger than that of the terahertz beam probes the electric field distribution within the crystal via the Pockels effect. The 2-D terahertz field distribution in the sensor crystal is converted into a 2-D optical intensity distribution after the readout beam passes through a crossed polarizer (analyzer), and the optical image is then recorded by a linear diode array

or a digital CCD camera. Figure 14 shows one of the ZnTe crystals used for the terahertz imaging system and several electrical and optical parameters. By illuminating an electro-optic crystal using a terahertz beam and an optical readout beam, then detecting the optical beam using a linear diode array or CCD camera, time-resolved 1-D or 2-D images of pulsed far-infrared radiation can be achieved, respectively. This system is can noninvasively image of moving objects, turbulent flows, or explosions.

Furthermore, real-time 2-D terahertz imaging is feasible using the electro-optic technique. By chopping the terahertz beam, the change in optically recorded gray levels can be observed in real time by the human eye. In this system, the temporal resolution is about 50 fs, and the frame transfer rate is 38 images per second. Astonishingly, the transfer rate can be extended to as much as 190 images/second (digital out) and 2,000 images/second (video out)!

Figure 15 shows terahertz images of a mammographic phantom, which is designed to test the performance of a mammographic system by quantitatively evaluating of the system's ability to image small structures similar to those found clinically. Images of these fibers, masses, and specks, that are as thin as 0.24 mm have been resolved.

The diameter of the current image is 1 cm, using an intense laser system, an imaging size of 20 cm is possible. Figure 16 shows a standard photo and a terahertz image of compressed breast tissue obtained by the scanning image system. A 0.6 mm abnormal wire structure is clearly distinguished as a dark line in the breast tissue in the image range of 1×1 cm^2.

Figure 17 shows a standard photo, a transmitted terahertz image, and the terahertz transmission waveform through normal and cancerous tissues (breast cancer phantom). The dense tumor has a lower water concentration than the normal tissue, so the transmitted waveform has a temporal delay and provides less attenuation compared with normal tissue. Significant spectral difference between the normal and cancerous tissues can be seen between 30 GHz and 0.2 THz. The electro-optic imaging system can image fast moving species, such as real-time imaging of living insects. Figure 18 demonstrates real-time *in vivo* terahertz images of insects, such as a fly, worm, ant and ladybug. The antennae and legs of the ant and the organs in the ladybug can be resolved. Terahertz radiation has no ionizing effect, and the spectrum of a

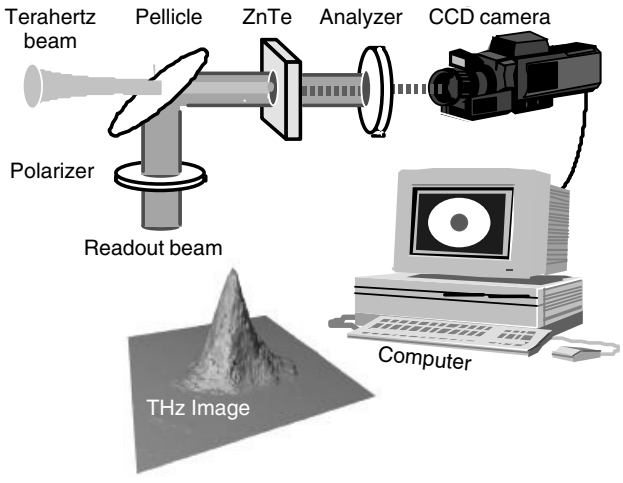


Figure 13. Setup for converting a terahertz image into an optical image. The 2-D field distribution in the sensor crystal is converted into a 2-D optical intensity distribution after the readout beam passes through a crossed analyzer.

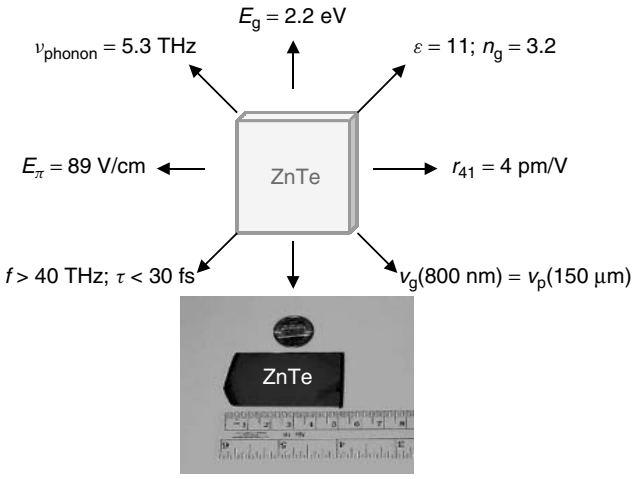


Figure 14. One of the ZnTe crystals used in terahertz imaging. The useful area is more than 3×3 cm^2.

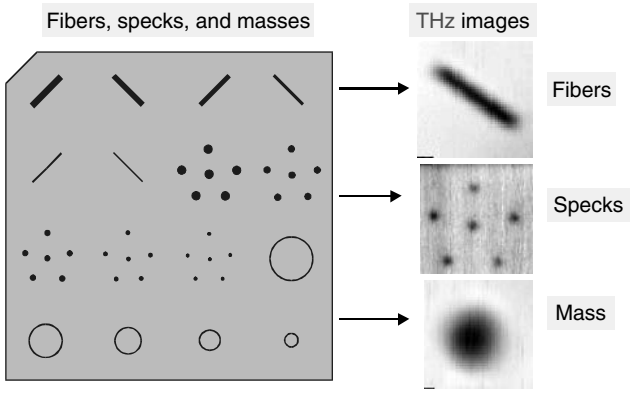


Figure 15. Terahertz images of fibers, mass, and specks. Small structures less than 0.5 mm thick and diameter less than 0.24 mm can be resolved.

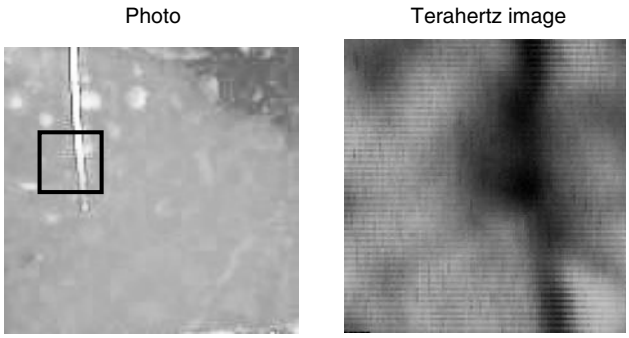


Figure 16. Photo of human breast tissue and a T-ray image of a 0.6 mm abnormal structure (shadow).

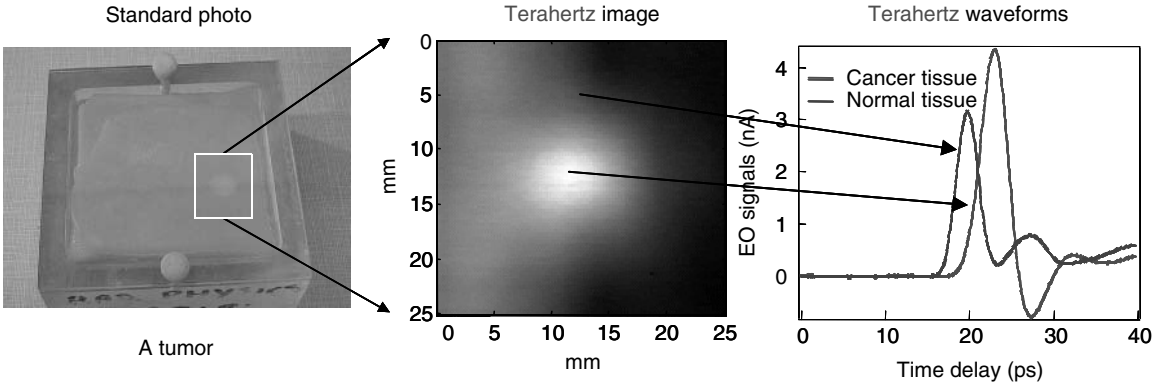


Figure 17. A breast tissue sample for terahertz measurement. The light spot near the center is a cancerous tumor. Transmitted terahertz waveform is from normal tissue and cancerous tissue.

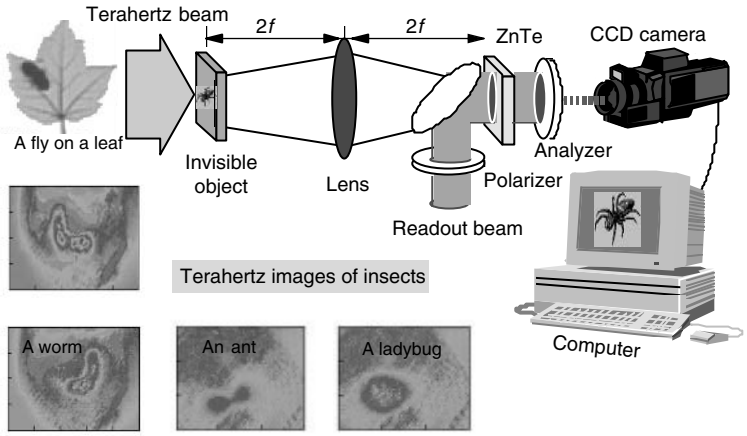


Figure 18. *In vivo* terahertz images of insects, such as a fly, worm, ant and ladybug. An image rate as fast as 30 frames/s is achieved.

terahertz wave falls within the heat range; therefore it is safe for medical applications.

Figure 19 shows the schematic for terahertz imaging of currency watermarks. Unlike intensity images viewed by a visible beam, the watermark images in Fig. 19 are obtained purely by the phase difference of the terahertz pulse transmitted through the watermarks. The maximum phase shift is less than 60 fs. The terahertz absorption is less than 1%. Clearly these terahertz watermark images in the terahertz spectrum show an alternative method for detecting-counterfeiting.

Electro-optic imaging makes it possible to see terahertz wave images of electrical fields, diseased tissue, the chemical composition of plants, and much more that is undetectable by other imaging systems. Real-time monitoring of a terahertz field supports real-time diagnostic techniques.

THZ WAVE TRANSCEIVER

In a conventional experimental setup of terahertz time-domain spectroscopy, a separate terahertz transmitter and terahertz receiver are used to generate and detect the terahertz signal. However, because electro-optic detection is the reverse of rectified generation, the transmitter and the receiver can be the same crystal (31). Therefore, a terahertz transceiver, which alternately transmits pulsed electromagnetic radiation (optical rectification) and receives the returned signal (electro-optic effect) is feasible. The use of a transceiver has its advantages for terahertz range remote sensing and tomographic imaging.

Theoretically and experimentally, it has also been demonstrated that the working efficiency of an electro-optic transceiver constructed from a (110) zincblende crystal is optimized when the pump beam polarization is 26° counterclockwise from the crystallographic Zaxis of the crystal. An experimental setup of a terahertz imaging system using an electro-optic transceiver is shown in Fig. 20. Compared to the traditional terahertz tomographic setup in reflective geometry, this imaging system using an electro-optic transceiver is simpler and easier to align. Besides, the normal incidence of the terahertz beam on the sample can be maintained. Greater than 50 meters of free-space terahertz generation, propagation and detection has been demonstrated by using this transceiver.

Terahertz tomographic imaging using an electro-optic transceiver is illustrated by using a razor pasted on a metal

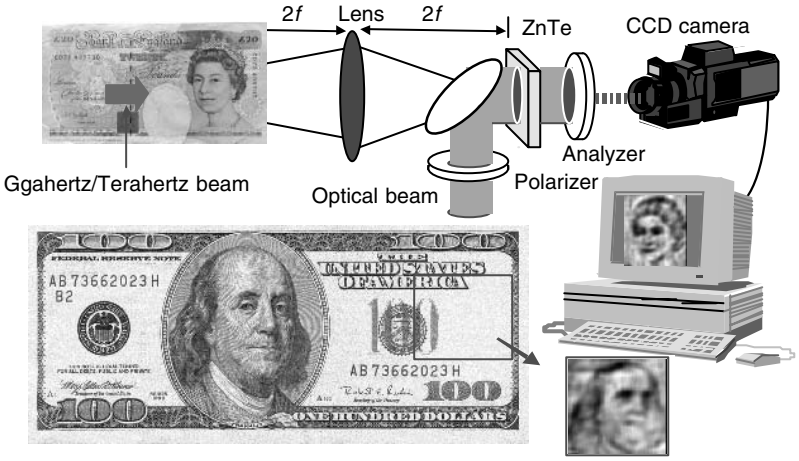


Figure 19. Terahertz image based on the phase difference in a currency watermark structure. A phase shift as small as a few femtoseconds can be resolved.

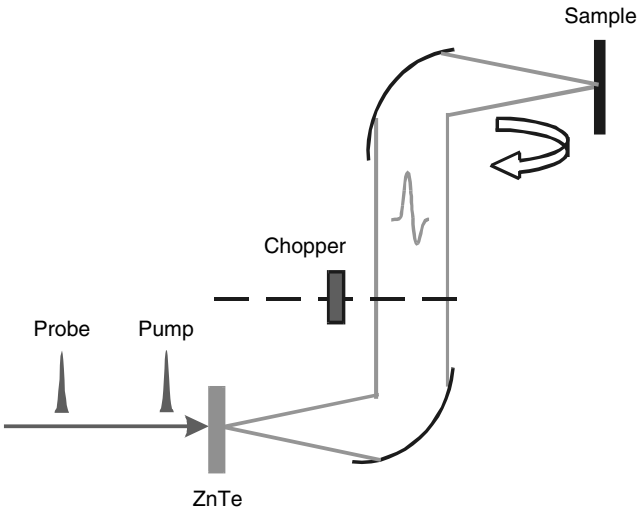


Figure 20. Schematic experimental setup of an electro-optic terahertz transceiver. The terahertz signal is generated and detected by the same ZnTe crystal.

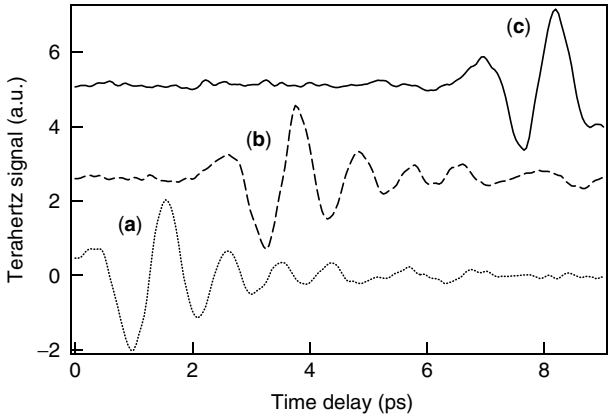


Figure 21. Terahertz waveforms reflected from (**a**) the metal handle of a razor, (**b**) the razor surface, and (**c**) the metal mirror.

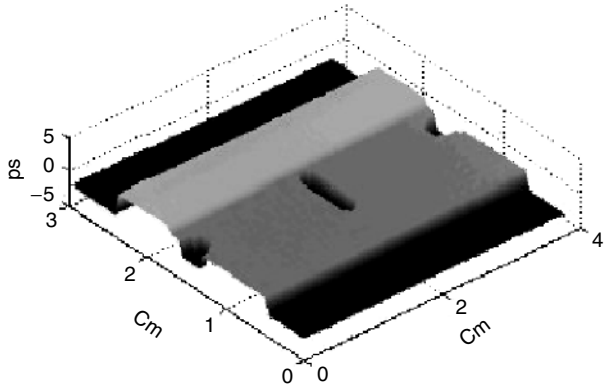


Figure 22. Terahertz tomographic image of a razor, the gray level represents the timing of the peak intensity.

mirror. There are three different reflective metal layers in this sample; the first is the metal handle of the razor, the second is the razor surface, and the third is the metal mirror. Figure 21 shows the terahertz waveforms reflected from these three different layers; the timing difference in the peak intensity spatially separate these layers, which can be used to construct a three-dimensional tomographic image of a razor, as shown in Fig. 22.

Using the same imaging system, terahertz tomographic images of a quarter dollar and a 50-pence piece are shown in Fig. 23. The image contrast is limited by the terahertz beam focal size and the flatness of the background metal surface. The width of the short timing window is determined by degree of "unflatness" of the target. If two images are from two different reflective layers and their spatial separation (depth) is large enough, the image can be displayed in this fashion at two different timing positions; the timing difference is proportional to the depth difference between two layers. Three-dimensional terahertz imaging can still be realized without displaying the image in terms of the timing of peak intensity.

TERAHERTZ WAVE NEAR-FIELD IMAGING

The terahertz wave near-field imaging technique can greatly improve the spatial resolution of a terahertz wave sensing and imaging system (32). Dr. Klass Wynne

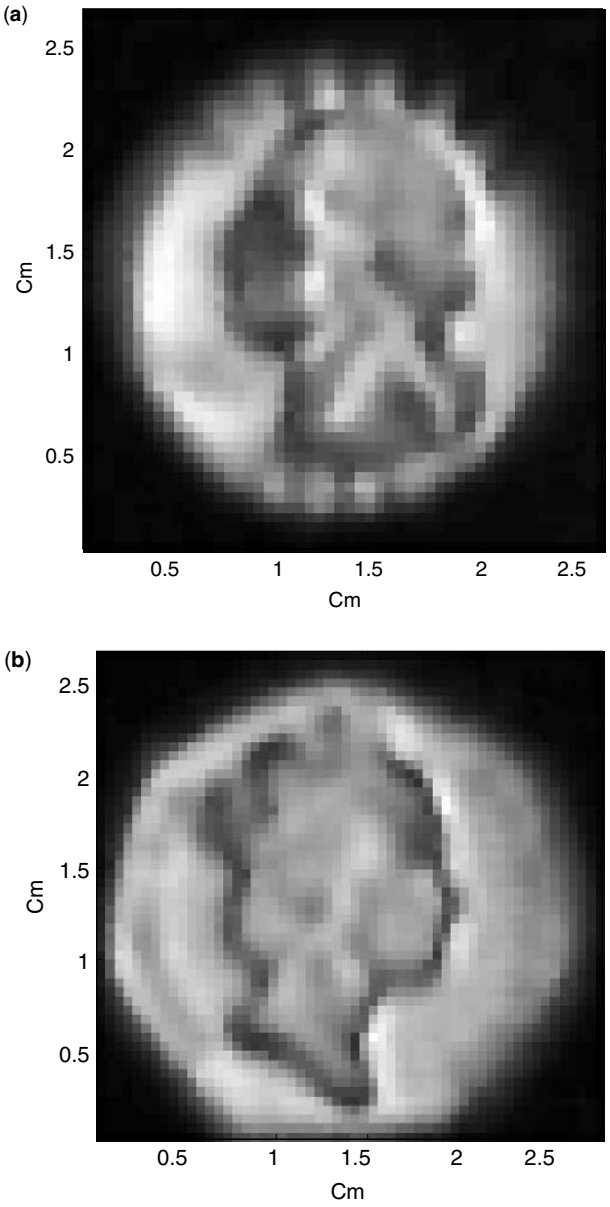


Figure 23. Terahertz image of (**a**) a quarter dollar; (**b**) a fifty-pence piece; the gray level represents the peak intensity within a certain temporal window.

in the United Kingdom has demonstrated 110-μm and 232-μm spatial resolution for $\lambda = 125$ μm and $\lambda = 1$ mm, respectively (33). The improvement factor is about 2 to 4. O. Mitrofanov and John Federici at the New Jersey Institute of Technology and Bell Laboratory reported the use of collection mode near-field imaging to improve spatial resolution (34–36). The best result reported is 7-μm imaging resolution using 0.5 terahertz pulses. This is about 1/100 of the wavelength. The limitation of such a system is the low throughput of the terahertz wave past the emitter tip, the throughput terahertz wave field is inversely proportional to the third power of the aperture size of the emitter tip.

A newly developed dynamic-aperture method that introduces a third gating beam can image objects at a subwavelength resolution ($\lambda/100$); however, the drawback of this method is the difficulty in coating a gating material on the surface of biomedical samples such as cells and tissues (37).

Dr. Wynne's method (the use of an electro-optic crystal as a near-field emitter) led to the development of the terahertz wave microscope. Biomedical samples are mounted directly on the surface of the crystal. Figure 24 shows the terahertz wave near-field microscope for 2-D microscopic imaging. In this case, terahertz waves are generated in the crystal by optical rectification and are detected by a terahertz wave detector crystal by the electro-optic effect.

The spatial resolution is limited only by the optical focal size of the laser on the crystal (less than 1 μm due to the large refractive index of 2.8 for ZnTe) under moderate optical power, and it is independent of the wavelength of the terahertz wave.

A coated thin ZnTe plate (antireflective coating for the bottom surface and highly-reflective coating for the top surface) is placed at the focal plane of the microscope as the terahertz wave emitter. The coating prevents optical loss in the crystal and leakage of the optical beam into the tissue sample. The tissue can be monitored by the optical microscope. A laser beam is guided from the bottom of the microscope into the terahertz emitter. Terahertz waves generated by the emitter can be detected in the transmitted mode (a terahertz wave sensor is mounted on top of the microscope) and/or reflection mode (transceiver). The emitter, the sample, or the terahertz beam can be scanned laterally to obtain a 2-D image. Submicron spatial resolution is expected, even though the imaging wavelength is about 300 μm at 1 THz.

In the transmitted mode shown in Fig. 25, a separate ZnTe sensor crystal (terahertz detector) is required, and a probe beam is required to sample the terahertz wave in the sensor crystal. This construction also applies the concept of the terahertz wave transceiver, which combines the emitter and receiver in one crystal in the near-field range, as shown in Fig. 26. Both transmitted and reflected terahertz wave microscopic images can therefore be obtained from the proposed system.

When a Ti : sapphire laser where $\lambda = 0.8$ μm is used as the optical source, the smallest optical focal spot

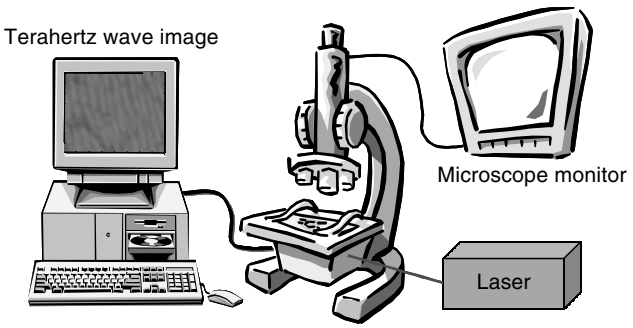


Figure 24. The concept of 2-D near-field inverted terahertz wave microscope imaging (left) and the schematic of a terahertz wave microscope system (right). A tissue sample is placed on top of a terahertz wave emitter.

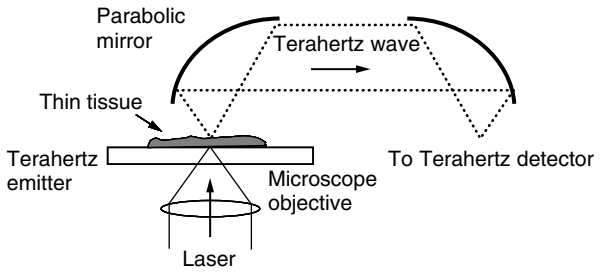


Figure 25. The Tray is generated and detected in one ZnTe crystal in the reflected geometry. The T-ray imaging spot on the tissue is comparable to the focal spot of the optical beam.

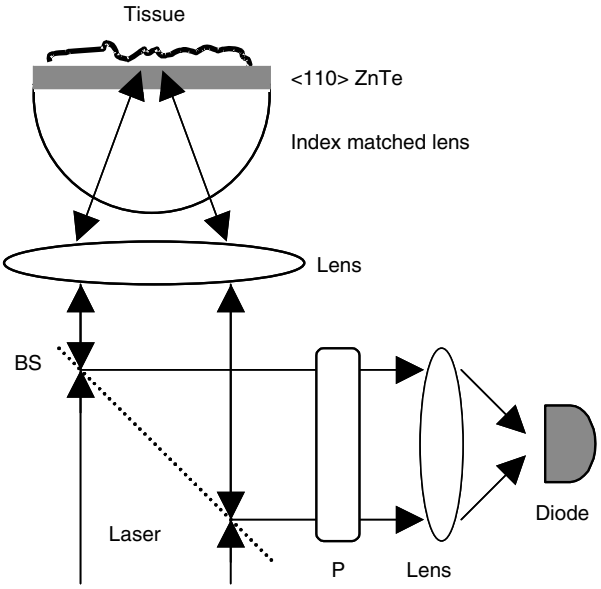


Figure 26. The T-ray is generated and detected in one ZnTe crystal in the reflected geometry.

a in the air is calculated from the standard equation $a = 1.22\lambda 2f/D$ the (1.22 factor comes from the diffraction limit under the Gaussian beam approximation) where f is the wavelength, D is the beam diameter, and $D/2f$ is the numerical aperture NA of the microscopes objective lens. Assuming the ideal case where NA = 1, then $a = 1$ μm.

A possible way of achieving submicron lateral resolution is to focus the optical beam into a high refractive index medium. The refractive index of ZnTe is greater than 1, therefore, the focal spot in a ZnTe must be smaller than that in air by the factor of the refractive index value. However, when directly focusing a laser beam from air into a ZnTe plate, as shown in Fig. 25, it is difficult to achieve a much smaller focal spot because of the change in the numerical aperture after optical refraction at the ZnTe interface by Snells law. This can be improved by using a high-index hemispherical lens, as shown in Fig. 26. The numerical apertures of the first focal lens and the hemispherical lens must be identical. A thin ZnTe plate is placed on the top of the hemispherical lens, which has the same refractive index as that of the ZnTe ($n = 2.8$).

The optical beam is focused in the ZnTe through the matching refractive index lens to a spot size comparable to $1.22\lambda/n$ (assume NA = 1). If $\lambda = 0.8$ μm and $n = 2.8$, in theory, a focal spot can be as small as 0.35 μm. By using a shorter optical wavelength, such as the second-harmonic wave from a Ti : sapphire laser, a smaller focal spot is expected.

An electro-optic terahertz transceiver can be used in the reflective mode of the near-field terahertz wave microscope. In this case, the terahertz wave is generated and detected at the same focal spot within a thin crystal (ZnTe). The target sample (biomedical tissue) is placed on top of the crystal (terahertz transceiver). The measured area of the tissue is comparable to the optical focal size.

Due to the intense power density at an optical focal spot (micron or submicron), higher order nonlinear phenomena other than optical rectification have to be considered. Some effects may limit T-ray generation and detection. For example, two-photon absorption (a third-order nonlinear optical effect) in ZnTe generates free carriers. In a tight focal spot, extremely high free-carrier density changes the ZnTe local conductivity, screens the Trays, and saturates the T-ray field. A possible solution is to reduce the optical peak power and increase the pulse repetitionrate. This method can maintain the same average power.

CONCLUSION

The terahertz band occupies an extremely broad spectral range between the infrared and microwave bands. However, due to the lack of efficient terahertz emitters and sensors, far less is known about spectra in the terahertz band than those in the rest of the electromagnetic spectrum. Recently developed photoconductive antennas and free-space electro-optic sampling provide measurement sensitivity by several orders better than conventional bolometer detection, but it is still far from the detection resolution achieved in other frequency bands. The development of instruments impacts physics and basic science; recent examples are the scanning tunneling microscope and near field optical microscope, which opened the following new fields to the physics community.

Terahertz System for Spectroscopy

The powerful capabilities of the time-domain terahertz spectroscopic technique results from the ability to provide 20-cm diameter real-time images at a variable frame rate up to 2,000 frames/second and the ability to image moving objects, turbulent flows, and explosions noninvasively. Furthermore, the imaging system will also have a subwavelength spatial resolution (1/1,000 λ), 50-femtosecond temporal resolution, sub-mV/cm field sensitivity and be capable of single-shot measurements.

New Terahertz Sources

New terahertz beam sources emphasize tunable narrowband terahertz lasers using novel semiconductor structures. For example, a terahertz laser was recently developed using a p-type Ge, which operates at the temperature of liquid nitrogen. In this laser, a novel unipolar-type

population inversion is realized by the streaming motion of carriers in the semiconductor. It may be possible to improve these solid-state terahertz laser sources further by using strained SiGe.

Terahertz BioChip and Spectrometer

The detection of nanolayers using terahertz techniques is a challenge, which cannot be solved by conventional transmission techniques. The fact that the thickness of the layer is orders of magnitude smaller than the wavelength of the terahertz radiation leads to layer specific signatures that are so small that they are beyond any reasonable detection limit. An alternative method for detecting nanolayers is grating couplers. Evanescent waves, that travel on the grating enlarge the interactive length between the nanolayers and terahertz radiation from nanometers to several tens of micrometers.

Quantum Terahertz Biocavity Spectroscopy

The concept is to design and fabricate photonic band-gap structures in the terahertz regime and place materials such as DNA, biological and chemical agents, or quantum dots in the active region of the cavity. This configuration will be useful for enhanced absorption as enhanced spontaneous emission spectroscopy. Previous research has indicated that DNA (and other cellular materials) possess large numbers of unique resonances due to localized phonon modes that arise from DNA base-pair interaction, that is absent from far-infrared data.

Terahertz Molecular Biology Spectroscopy

Understanding how genes are regulated is one of the grand challenges in molecular biology. Recent reports indicate that pulsed terahertz spectroscopy provides a new handle on the interactions of biopolymers. The use of pulsed terahertz spectroscopy in the study of the binding of transcription factors to these cognate DNA binding sites will be explored.

Near-Field Terahertz Imaging

The current near-field terahertz imaging system has a spatial resolution of 1/50 λ. The microscopic imaging system will achieve submicron spatial resolution ($<$1/1,000 λ resolution). If this can be done, it will provide direct spectroscopy and imaging of biological cells at terahertz frequency, which has not been possible in the past.

New Terahertz Source and Detector (Plasma Wave Generator)

Existing terahertz sources have limited power and typically deliver very broadband radiation. Existing terahertz detectors are also broadband. To achieve breakthroughs in the physics and applications of terahertz technology, we need to develop much more efficient and powerful sources of terahertz radiation and realize resonant and tunable detectors of terahertz radiation. To achieve these goals, we will need to develop entirely new physical approaches. The new physics will be based on using collective electronic motion plasma waves, rather than relying on transit and/or tunneling phenomena for individual electrons. Just as sound waves propagating in air achieve velocities that are orders of magnitude higher than the velocities of individual molecules propagating from the sound source to a sound detector, such as the human ear, in the same way electron plasma waves can propagate at much higher velocities and can generate and detect terahertz radiation. A preliminary theoretical foundation of this approach to terahertz generation and detection has already been established and the first experimental results have been obtained for detecting and generating terahertz radiation by a two-dimensional electron gas.

Acknowledgments
This work was supported by the U.S. Army Research Office and the U.S. National Science Foundation.

ABBREVIATIONS AND ACRONYMS

THz	terahertz
GaAs	gallium arsenate
ZnTe	zinc telluride
2-D	two-dimensional
IR	infrared
DNA	deoxyribonucleic acid
RNA	ribonucleic acid
CCD	charged couple device
ps	picosecond
fs	femtosecond

BIBLIOGRAPHY

1. B. B. Hu and M. C. Nuss, *Opt. Lett.* **20**, 1,716–1,718 (1995).
2. M. C. Nuss, *Circuits & Devices* **12**, 25–30 (1996).
3. D. M. Mittleman, R. H. Jacobsen, and M. C. Nuss, *IEEE-JSTQE* **3**, 678–692 (1996).
4. D. M. Mittleman, J. Cunningham, M. C. Nuss, and M. Geva, *Appl. Phys. Lett.* **71**, 16–18 (1997).
5. S. Mickan, D. Abbott, and X. -C. Zhang, *Microelectronics J.* **31**, 503–514 (2000).
6. Q. Wu and X. -C. Zhang, *Appl. Phys. Lett.* **67**, 3,523–3,525 (1995).
7. A. Nahata, D. H. Auston, T. F. Heinz, and C. Wu, *Appl. Phys. Lett.* **68**, 150–152 (1996).
8. P. Uhd Jepsen et al., *Phys. Rev. E* **53**, 3,052–3,054 (1996).
9. Ch. Fattinger and D. Grischkowsky, *Appl. Phys. Lett.* **53**, 1,480–1,482 (1988).
10. A. Rice et al., *Appl. Phys. Lett.* **64**, 1,324–1,326 (1994).
11. J. T. Darrow, X. -C. Zhang, D. H. Auston, and J. D. Morse, *IEEE J. Quantum Electron.* **28**, 1,607–1,616 (1992).
12. M. Bass, P. A. Franken, J. F. Ward, and G. Weireich, *Phys. Rev. Lett.* **9**, 446–448 (1962).
13. X. -C. Zhang, Y. Jin, K. Yang, and L. J. Schowalter, *Phys. Rev. Lett.* **69**, 2,303–2,306 (1992).
14. Q. Wu and X. -C. Zhang, *IEEE-JSTQE* **3**, 693 (1996); *Opt. Quantum Electron.* **28**, 945–951 (1996).
15. A. Nahata and T. F. Heinz, *IEEE-JSTQE* **3**, 701–708 (1996).
16. J. T. Darrow, B. B. Hu, X. -C. Zhang, and D. H. Auston, *Opt. Lett.* **15**, 323–325 (1990).
17. X. -C. Zhang and D. H. Auston, *J. Electromagnetic Waves Appl.* **6**, 85–106 (1992).

18. X. -C. Zhang and D. H. Auston, *J. Appl. Phys.* **71**, 326–338 (1992).
19. Q. Wu and X. -C. Zhang, *Appl. Phys. Lett.* **70**, 1,784–1,786 (1997).
20. Q. Wu and X. -C. Zhang, *Appl. Phys. Lett.* **71**, 1,285–1,286 (1997).
21. Z. Jiang and X. -C. Zhang, *Appl. Phys. Lett.* **72**, 1,945–1,947 (1998).
22. Z. Jiang and X. -C. Zhang, Engineering Laboratory Note, *Optic & Photonic News* **9**(1-2), (1998). Also in *Appl. Opt.* **37**, 8,145–8,146 (1998).
23. F. G. Sun, Z. Jiang, and X. -C. Zhang, *Appl. Phys. Lett.* **73**, 2,233–2,235 (1998).
24. Z. Jiang and X. -C Zhang, *Opt. Lett.* **23**, 1,114–1,116 (1998).
25. Z. Jiang and X. -C. Zhang, *Opt. Express* **5**(11), 243–248 (1999).
26. X. -C. Zhang and Q. Wu, *Opt. & Photonics News* **12**, 9 (1996).
27. X. -C. Zhang, Q. Wu, and T. D. Hewitt, in *Ultrafast Phenomena **X**, Springer Series in Chemical Physics*, **54**, 463–465 Berlin, Heidelberg, 1996.
28. Q. Wu, F. G. Sun, P. Campbell, and X. -C. Zhang, *Appl. Phys. Lett.* **68**, 3,224–3,226 (1996).
29. Q. Wu, T. D. Hewitt, and X. -C. Zhang, *Appl. Phys. Lett.* **69**, 1,026–1,028 (1996).
30. Z. G. Lu, P. Campbell, and X. -C. Zhang, *Appl. Phys. Lett.* **71**, 593–595 (1997).
31. Q. Chen, Z. Jiang, M. Tani, and X. -C. Zhang, *Electron. Lett.* **36**, 1,298–1,299 (2000).
32. S. Hunsche, M. Koch, I. Brener, and M. C. Nuss, *Opt. Commun.* **150**, 22–26 (1998).
33. K. Wynne and D. A. Jaroszynske, *Opt. Lett.* **24**, 25–27 (1998).
34. O. Mitrofanov et al., *Appl. Phys. Lett.* **78**, 252–254 (2001).
35. O. Mitrofanov et al., *Appl. Phys. Lett.* **77**, 3,496–3,498 (2000).
36. O. Mitrofanov et al., *Appl. Phys. Lett.* **77**, 591–593 (2000).
37. Q. Chen, Z. Jiang, G. X. Xu, and X. -C. Zhang, *Opt. Lett.* **25**, 1,122–1,124 (2000).

TOMOGRAPHIC IMAGE FORMATION TECHNIQUES

GABOR T. HERMAN
The Graduate Center, CUNY
New York, NY

THE RECONSTRUCTION PROBLEM

Computerized tomography is the process of reconstructing the interiors of objects from collected data based on transmitted or emitted radiation. This section explains the nature of the reconstruction problem, and the following sections present computer algorithms used for achieving the reconstructions and approaches for evaluating the efficacy of such algorithms.

In the reconstruction problem, there is a three-dimensional structure of unknown internal composition. The structure is subjected to some kind of radiation, either by transmitting the radiation through the structure or by introducing the emitter of the radiation into the structure. The radiation transmitted through or emitted from the structure is measured at a number of points. *Computerized tomography* (CT) is the process of obtaining from these measurements the distribution of the physical parameter(s) inside the structure that have an effect on the measurements. The problem occurs in a wide range of areas, such as X-ray CT, emission tomography, photon migration imaging, and electron microscopic reconstruction; see, for example (1,2). All of these are inverse problems of various sorts; see, for example (3). A snapshot of the recent state of the art regarding volumetric reconstruction of medical images is provided by the May 2000 Special Issue of the *IEEE Transactions on Medical Imaging*.

An important special case that provides useful insights into all variations of the reconstruction problem is estimating of a function of two variables (in other words, a planar object) from estimates of its line integrals. As is quite reasonable for any application, it is assumed that the domain of the function is contained in a finite region of the plane. Suppose that f is a function of the two polar variables r and ϕ. Let $[\mathcal{R}f](\ell, \theta)$ denote the line integral of f along the line that is at distance ℓ from the origin and makes an angle θ with the vertical axis. This operator $\mathcal{R}$ is usually referred to as the *Radon transform*. [An alternative name for the operator is the *X-ray transform*. In three dimensions (3D), the two operators differ from each other: the X-ray transform still maps the function to be reconstructed into its line integrals, but the Radon transform in 3D becomes an operator that maps a function f into a function that associates with each plane the integral of f across that plane. Such planar integrals are natural for certain applications, such as magnetic resonance imaging (4).]

The input data to a reconstruction algorithm are estimates (based on physical measurements) of the values of $[\mathcal{R}f](\ell, \theta)$ for a finite number of pairs (ℓ, θ); its output is an estimate, in some sense, of f. More precisely, suppose that the estimates of $[\mathcal{R}f](\ell, \theta)$ are known for I pairs: (ℓ_i, θ_i), $1 \leq i \leq I$. The notation y is used for the I-dimensional column vector (called the *measurement vector*) whose ith component, y_i, is the available estimate of $[\mathcal{R}f](\ell_i, \theta_i)$. The task of a reconstruction algorithm is

given the data y, **estimate** the function f.

Following (1), reconstruction algorithms are classified as either transform methods or series expansion methods. The following sections discuss the underlying ideas of these two approaches and give detailed descriptions of two algorithms from each category.

TRANSFORM METHODS

The Radon transform has an inverse, R^{-1}, defined as follows. For a function p of ℓ and θ,

$$[R^{-1}p](r, \phi) = \frac{1}{2\pi^2} \int_0^{\pi} \int_{-\infty}^{\infty} \frac{1}{r\cos(\theta - \phi) - \ell} p_1(\ell, \theta)\, d\ell\, d\theta, \tag{1}$$

where $p_1(\ell, \theta)$ denotes the partial derivative of p with respect to its first variable ℓ. [Note that it is intrinsically assumed in this definition that p is sufficiently smooth for the existence of the integral in Eq. (1).] It is known (1) for any function f that satisfies some physically reasonable conditions (such as continuity and boundedness), for all points (r, ϕ), that

$$[R^{-1}Rf](r, \phi) = f(r, \phi). \tag{2}$$

Transform methods are numerical procedures that estimate values of the double integral on the right-hand side of Eq. (1) from given values of $p(\ell_i, \theta_i)$, for $1 \le i \le I$. Examples of such methods are the widely adopted *filtered backprojection* (FBP) algorithm, called the *convolution* method in (1), and the more recently developed *linogram* method.

Filtered Backprojection (FBP)

In this algorithm, the right-hand side of Eq. (1) is approximated by a two-step process [for derivational details, see (1,5) or, in a more general context, (3)]. First, for fixed values of θ, convolutions defined by

$$[p *_Y q](\ell', \theta) = \int_{-\infty}^{\infty} p(\ell, \theta) q(\ell' - \ell, \theta)\, d\ell, \tag{3}$$

are carried out, using a *convolving function* q (of one variable) whose choice has important influence on the appearance of the final image. Second, the estimate f^* of f is obtained by *backprojection* as follows:

$$f^*(r, \phi) = \int_0^{\pi} [p *_Y q](r \cos(\theta - \phi), \theta)\, d\theta. \tag{4}$$

To make the implementation of this explicit for a given measurement vector, suppose that the data function p is known at points $(nd, m\Delta)$, $-N \le n \le N$, $0 \le m \le M - 1$, and $M\Delta = \pi$. Assume further that the function f is to be estimated at points (r_j, ϕ_j), $1 \le j \le J$. The computer algorithm operates as follows.

A sequence $f_0, \ldots, f_{M-1}, f_M$ of estimates is produced; the last of this is the output of the algorithm. First, let

$$f_0(r_j, \phi_j) = 0, \tag{5}$$

for $1 \le j \le J$. Then, for each value of m, $0 \le m \le M - 1$, let the $(m + 1)$st estimate be produced from the mth estimate by a two-step process:

1. For $-N \le n' \le N$, calculate

$$p_c(n'd, m\Delta) = d \sum_{n=-N}^{N} p(nd, m\Delta) q[(n' - n)d], \tag{6}$$

using the measured values of $p(nd, m\Delta)$ and precalculated values (same for all m) of $q[(n' - n)d]$. This is a discretization of Eq. (3). One possible choice is

$$q(nd) = \begin{cases} \dfrac{1}{4d^2}, & \text{if } n = 0, \\ -\dfrac{1}{(\pi dn)^2}, & \text{if } n \text{ is odd and nonzero,} \end{cases} \tag{7}$$

and $q(nd) = 0$, otherwise. [This is often referred to as the *Ram–Lak filter* (6); it is the sampled inverse Fourier transform of the ramp function, $Q(X) = |X|$, between $-1/2d$ and $1/2d$.] The actual choice of the convolving function q should depend on a number of things, especially on the noise level in the measurement data; see Section 8.6 of (1).

2. For $1 \le j \le J$, set

$$f_{m+1}(r_j, \phi_j) = f_m(r_j, \phi_j) + \Delta p_c[r_j \cos(m\Delta - \phi_j), m\Delta]. \tag{8}$$

This is a discretization of Eq. (4). To obtain the values needed in Eq. (8), the first variable of p_c has to be interpolated from the values calculated in Eq. (6). There are many ways to do this interpolation (7). One recommended methodology is to use *modified cubic splines* (8), that is to use a cubic to interpolate between any pair of neighboring sample points so that the slope of the interpolated function at the sample points is the same as the slope of the straight line that connects the values at the two sample points on either side of it. An alternative is to use *sinc interpolation*, which can be achieved as follows: instead of convolving, as in Eq. (6), calculate the Fourier transform of p, multiply it pointwise by the ramp function Q, and pad out the product with many zeros before taking the inverse Fourier transform. The result of this will be a finely sampled version of the sinc interpolation. In practice, once $f_{m+1}(r_j, \phi_j)$ has been calculated, $f_m(r_j, \phi_j)$ is no longer needed, and the computer can reuse the same memory location for $f_0(r_j, \phi_j), \ldots, f_{M-1}(r_j, \phi_j), f_M(r_j, \phi_j)$.

In a complete execution of the algorithm, the use of Eq. (6) requires $M(2N + 1)^2$ multiplications and additions. However, the computational demand can be reduced by taking the *discrete Fourier transforms* (DFTs) of both p and q, multiplying these transforms pointwise, and then taking the inverse DFT. The DFT can always be implemented (possibly after some padding by zeros) very efficiently (9) by using the *fast Fourier transform* (FFT) that reduces the order of the required operations to $M(2N + 1)\log(2N + 1)$. On the other hand, all of the uses of Eq. (8) require MJ interpolations and additions. Because J is typically of the order of N^2 and N itself in typical applications is between 100 and 1000, the computational cost of backprojection is likely to be much higher than the cost of convolution. In any case, reconstruction of a cross section consisting of $1{,}000 \times 1{,}000$ pixels (picture elements) from data collected by a typical X-ray CT device is not a challenge to state-of-the art computational capabilities; it is routinely done in time of the order of a second or so and can be done, using a pipeline architecture, in a fraction of a second (10).

The Linogram Method

This method was proposed in (11), and the reason for its name can be found there. The basic result that justifies the method is the well-known *projection theorem* which says that "taking the two-dimensional Fourier transform is the same as taking the Radon transform and then applying the Fourier transform with respect to the first variable" (1). (This theorem leads to great insight in many areas of tomography.) The main motivation for the method is its speed of execution. In the description that follows, the approach of (12) is used. That paper deals with the fully three-dimensional problem; here it is simplified to the two-dimensional case.

For the linogram approach, it is assumed that the data were collected at particular points whose locations are specified next; if collected otherwise, one needs to interpolate before reconstruction. If the function is to be estimated at an array of points that have rectangular coordinates $[(id, jd) | -N \le i \le N, -N \le j \le N]$ (it is assumed that this array covers the object to be reconstructed), then the data function p needs to be known at points

$$\{(nd_m, \theta_m) | -2N-1 \le n \le 2N+1, -2N-1 \le m \le 2N+1\} \tag{9}$$

and at points

$$\left\{\left(nd_m, \frac{\pi}{2} + \theta_m\right) \middle| -2N-1 \le n \le 2N+1, \right. \\ \left. -2N-1 \le m \le 2N+1\right\}, \tag{10}$$

where

$$\theta_m = \tan^{-1}\frac{2m}{4N+3}$$

and

$$d_m = d\cos\theta_m. \tag{11}$$

The linogram method produces estimates of the function values from such data at the desired points by using a multistage procedure. The most expensive computation that is used in any of the stages is the taking of DFTs, which can always be implemented very efficiently by using the FFT. The output of any stage produces estimates of function values at exactly those points where they are needed for the discrete computations of the next stage; there is never any need to interpolate between stages. These two properties indicate why the linogram method is both computationally efficient and accurate. The stages of the procedure are as follows.

1. *Fourier transforming of the data*. For each value of the second variable, take the DFT of the data with respect to the first variable in Eqs. (9) and (10). By the projection theorem, this provides the estimates of the two-dimensional Fourier transform F of the object at points (in a rectangular coordinate system)

$$\left\{\left(\frac{k}{(4N+3)d}, \frac{k}{(4N+3)d}\tan\theta_m\right) \middle| -2N-1 \le k \right. \\ \left. \le 2N+1, -2N-1 \le m \le 2N+1\right\} \tag{12}$$

and at points (also in a rectangular coordinate system)

$$\left\{\left(\frac{k}{(4N+3)d}\tan\left(\frac{\pi}{2}+\theta_m\right), \frac{k}{(4N+3)d}\right) \middle| -2N-1 \right. \\ \left. \le k \le 2N+1, -2N-1 \le m \le 2N+1\right\}. \tag{13}$$

2. *Windowing*. At this point, those frequencies that are considered noise-dominated may be suppressed by multiplication by a *window function* (corresponding to the convolving function in the FBP method).
3. *Separating into two functions*. The sampled Fourier transform F of the object to be reconstructed is written as the sum of two functions G and H. G has the same values as F at all of the points specified in Eq. (12) except at the origin and is zero-valued at all other points. H has the same values as F at all of the points specified in Eq. (13), except at the origin, and is zero-valued at all other points. Clearly, except at the origin, $F = G + H$. The idea is that by first taking the two-dimensional inverse Fourier transforms of G and H separately and then adding the results, an estimate of f is obtained [except for an additive constant, called the DC term, that has to be estimated separately, see (12)]. What needs to be done with G is described now; the treatment of H is analogous.
4. *Chirp z-transforming in the second variable*. Note that the way the θ_m were selected implies that for a fixed k, the sampling in the second variable of Equation (12) is uniform. Furthermore, it is known that G is zero outside the sampled region. Hence, for each fixed k, $0 < |k| \le 2N+1$, the chirp z-transform (13) can be used to estimate the inverse DFT in the second variable at points

$$\left\{\left(\frac{k}{(4N+3)d}, jd\right) \middle| -2N-1 \le k \le 2N+1, \right. \\ \left. -N \le j \le N\right\} \tag{14}$$

   The chirp z-transform can be implemented by using three FFTs; see (14).
5. *Inverse transforming in the first variable*. Now, the inverse Fourier transform of G can be estimated at the required points by taking, for every fixed j, the inverse DFT in the first variable of the values at the points of Eq. (14).

SERIES EXPANSION METHODS

This approach assumes that the function f to be reconstructed can be approximated by a linear combination of a finite set of known and fixed basis functions $b_j(r, \phi)$, that is,

$$f(r, \phi) \approx \sum_{j=1}^{J} x_j b_j(r, \phi), \tag{15}$$

and the task is to estimate the unknowns x_j. If it is assumed that the measurements depend linearly on the object to be reconstructed (certainly true in the special case of line integrals) and that one can calculate (at least approximately) what the measurements would be if the object to be reconstructed were one of the basis functions ($r_{i,j}$ is used to denote the value of the ith measurement of the jth basis function), then it can be concluded (1) that the ith of the measurements of f is approximately

$$\sum_{j=1}^{J} r_{i,j} x_j. \tag{16}$$

Then, the problem is to estimate the x_j from the measured approximations (for $1 \leq i \leq I$) of a Eq. (16). The estimate can often be selected as one that satisfies some *optimization criterion*.

To simplify notation, the image is represented by a J-dimensional *image vector* x (whose components are x_j), and the data form an I-dimensional measurement vector y. There is an assumed *projection matrix* R (that has entries $r_{i,j}$). Let r_i denote the transpose of the ith row of R ($1 \leq i \leq I$), and so the inner product $\langle r_i, x \rangle$ is the same as the expression in Eq. (16). Then, y is approximately Rx, and there may be further information that x belongs to a subset C of $\mathbf{R}^J$, the space of J-dimensional real vectors. In this formulation, R, C, and y are known, and x is to be estimated. Then, substituting the estimated values of x_j in Eq. (15) will give an estimate of the function f.

The simplest way of selecting the basis functions is by subdividing the plane into pixels (or 3-D space into voxels) and choosing basis functions whose value is one inside a specific pixel (or voxel) and is zero everywhere else. However, other choices may be preferable; for example, Ref. (15) uses spherically symmetrical basis functions that are spatially limited and also can be chosen to be very smooth. The smoothness of the basis functions then results in smoothness of the reconstructions, and the spherical symmetry allows easy calculation of the numbers $r_{i,j}$. It has been demonstrated (16) for fully three-dimensional *positron emission tomography* (PET) reconstruction that such basis functions lead to statistically significant improvements in the task-oriented performance of series expansion reconstruction methods.

In many situations, only a small proportion of the numbers $r_{i,j}$ are nonzero. (For example, if the basis functions are based on voxels in a $200 \times 200 \times 100$ array and the measurements are approximate line integrals, then the percent of nonzero $r_{i,j}$ is less than 0.01 because a typical line will intersect fewer than 400 voxels.) This makes certain types of iterative methods surprisingly efficient for estimating the x_j because one can use a subroutine that returns a list of those values of j, for any i, for which $r_{i,j}$ is not zero, together with the values of the $r_{i,j}$ (1,17). Two such iterative approaches are discussed now, the so-called *algebraic reconstruction techniques* (ART) and the use of *expectation maximization* (EM).

Algebraic Reconstruction Techniques (ART)

The basic version of ART operates as follows (1). The method cycles through the measurements repeatedly, considering only one measurement at a time. Only those x_j are updated for which the corresponding $r_{i,j}$ for the currently considered measurement i is nonzero, and the change made to x_j is proportional to $r_{i,j}$. The factor of proportionality is adjusted so that if Eq. (16) is evaluated for the resulting x_j, then it will match the ith measurement exactly. Other variants will use a block of measurements in one iterative step and will update the x_j in different ways to ensure that the iterative process converges according to a chosen optimization criterion.

Only one specific optimization criterion (and the associated algorithm) is discussed in detail here. [It is one of the several optimization criteria that are presented and justified in Section 6.4 of (1).] The task is to find the x in $\mathbf{R}^J$ that **minimizes**

$$r^2 \|y - Rx\|^2 + \|x - \mu_X\|^2, \tag{17}$$

($\|\cdot\|$ indicates the usual Euclidean norm) for a given constant scalar r (called the *regularization parameter*) and a given constant vector μ_X.

The algorithm, that uses an I-dimensional vector u of additional variables (one for each measurement), works as follows. First, define $u^{(0)}$ as the I-dimensional zero vector and $x^{(0)}$ as μ_X. Then, for $k \geq 0$, calculate

$$\begin{aligned} u^{(k+1)} &= u^{(k)} + c^{(k)} e_{i_k}, \\ x^{(k+1)} &= x^{(k)} + r c^{(k)} r_{i_k}, \end{aligned} \tag{18}$$

where e_i is the I-dimensional vector whose ith component is one, and all other components are 0, and

$$c^{(k)} = \lambda^{(k)} \frac{r(y_{i_k} - \langle r_{i_k}, x^{(k)} \rangle) - u_{i_k}^{(k)}}{1 + r^2 \|r_{i_k}\|^2}, \tag{19}$$

where $i_k = k (\mathrm{mod} I) + 1$.

Theorem. [see (1) for a proof]. Let y be any measurement vector, r be any real number, and μ_X be any vector in $\mathbf{R}^J$. Then, for any real numbers $\lambda^{(k)}$ satisfying

$$0 < \epsilon_1 \leq \lambda^{(k)} \leq \epsilon_2 < 2, \tag{20}$$

for arbitrary fixed ϵ_1 and ϵ_2, the sequence $x^{(0)}, x^{(1)}, x^{(2)}, \ldots$, determined by the algorithm given before, converges to the unique vector x that minimizes the expression in Eq. (17).

The implementation of this algorithm is as simple as that of basic ART which is described at the beginning of this subsection. An additional sequence of I-dimensional vectors $u^{(k)}$ is needed, but only one component of $u^{(k)}$ is used or altered in the kth iterative step. Because the indices i_k are defined in a cyclic order, the components of the vector $u^{(k)}$ (just as the components of the measurement vector y) can be sequentially accessed. (The exact choice of this — often referred to as *data access ordering* — is very

important for fast initial convergence; one such efficient ordering is described in (18). The underlying principle there is that the individual actions in any subsequence of steps should be as independent as possible.) Also, for every integer $k \geq 0$, a positive real number $\lambda^{(k)}$ needs to be selected. [These are the so-called *relaxation parameters* that are free parameters of the algorithm and in practice need to be optimized (18).] The vectors r_i are usually not stored at all, but the location and size of their nonzero components are calculated, as and when needed. Hence, the algorithm described by Eqs. (18) and (19) shares the storage-efficient nature of basic ART, and its computational requirements are essentially the same. Assuming, as is reasonable, that the number of nonzero $r_{i,j}$ is of the same order as J, the cost of cycling once through the data using ART is of the order IJ, which is approximately the same as the cost of a reconstruction using the FBP method. [That this is so is confirmed by the timings reported in (19).]

An important thing to note about the Theorem is that there are no restrictions of consistency of the system $y = Rx$ in its statement. Hence, the algorithm of Eqs. (18) and (19) will converge to the minimizer of Eq. (17) — the so-called *regularized least squares solution* — using the real data collected in any application.

ART can also be used as a *maximum entropy* technique. The *entropy* of a J-dimensional vector x is defined as

$$-\sum_{j=1}^{J} \frac{x_j}{J\overline{x}} \ln \frac{x_j}{J\overline{x}}, \tag{21}$$

where $\overline{x}$ is the average value of the x_j for $1 \leq j \leq J$. The use of this is usually justified by arguments (20) aimed at showing that of all x that satisfy some model equations, the one that maximizes entropy has the least information content, and so the resulting function f, defined by the right-hand side of Eq. (15), is the least likely to mislead the user by the presence of spurious features. There is a *multiplicative* version of ART [as opposed to the *additive* one, which is called so because of the additions on the right-hand side of Eq. (18)], which finds the maximum entropy solution of $Rx = y$, provided that there is a solution at all (21). A typical iterative step of such an algorithm (without a relaxation parameter, which may also be introduced here if desired) is

$$x_j^{(k+1)} = x_j^{(k)} \left(\frac{y_{i_k}}{\langle r_{i_k}, x^{(k)} \rangle} \right)^{r_{i_k j}}, \text{ for } 1 \leq j \leq J. \tag{22}$$

There are versions of this algorithm (22) that do not require exact satisfaction of a system of equations for convergence; such versions allow a theory that is applicable to real data.

Expectation Maximization (EM)

A reasonable alternative optimization task is to seek the x that maximizes the likelihood of observing the actual measurements, based on the assumption that the ith measurement comes from a Poisson distribution whose mean is given by Eq. (16). An iterative method to do exactly that, based on the so-called EM (*expectation maximization*) approach, was proposed in (23). The following variant of this approach was designed for a somewhat more complicated optimization criterion (24) that enforces smoothness of the results where the original maximum likelihood criterion may result in noisy images.

Let $\mathbf{R}_+^J$ denote those vectors of $\mathbf{R}^J$ in which all components are nonnegative. The task is to find the x in $\mathbf{R}_+^J$ that **minimizes**

$$\sum_{i=1}^{I} \left(\langle r_i, x \rangle - y_i \ln \langle r_i, x \rangle \right) + \frac{\gamma}{2} \langle x, Sx \rangle, \tag{23}$$

where ln is the natural logarithmic function, γ is a real regularization parameter, and the $J \times J$ matrix S (where entries are denoted by $s_{j,u}$) is a *modified smoothing matrix* (1) that has the following property. (This definition is applicable only if pixels are used to define the basis functions.) Let N denote the set of indexes corresponding to pixels that are not on the border of the digitization. Each such pixel has eight neighbors; let N_j denote the indexes of the pixels associated with the neighbors of the pixel indexed by j. Then,

$$\langle x, Sx \rangle = \sum_{j \in N} \left(x_j - \frac{1}{8} \sum_{k \in N_j} x_k \right)^2. \tag{24}$$

Consider the following rules for obtaining $x^{(k+1)}$ from $x^{(k)}$.

$$p_j^{(k)} = \frac{\sum_{i=1}^{I} r_{i,j}}{9\gamma s_{j,j}} - x_j^{(k)} + \frac{1}{9 s_{j,j}} \sum_{u=1}^{J} s_{j,u} x_u^{(k)}, \tag{25}$$

$$q_j^{(k)} = \frac{x_j^{(k)}}{9\gamma s_{j,j}} \sum_{i=1}^{I} \frac{r_{i,j} y_i}{\langle r_i, x^{(k)} \rangle}, \tag{26}$$

$$x_j^{(k+1)} = \frac{1}{2} \left[-p_j^{(k)} + \sqrt{(p_j^{(k)})^2 + 4 q_j^{(k)}} \right]. \tag{27}$$

Because the first term of Eq. (25) can be precalculated, executing Eq. (25) requires essentially no more effort than multiplying $x^{(k)}$ by the modified smoothing matrix. As explained in (1), there is a very efficient way of doing this. Executing Eq. (26) requires approximately the same effort as cycling once through the data set using ART; cf. Eq. (19). Algorithmic details of efficient computations of Eq. (26) appear in (25). Clearly, executing Eq. (27) requires a trivial amount of computing. Thus, one iterative step of the EM algorithm of Eq. (25) to Eq. (27) requires approximately the same total computing effort as cycling once through the data set using ART, which costs about the same as a complete reconstruction by the FBP method. A basic difference between the ART method of the previous subsection and the EM method of the current section is that the former updates its estimate based on one measurement at a time, whereas the latter deals with all measurements simultaneously.

Theorem. [see (24) for a proof]. For any $x^{(0)}$ that has only positive components, the sequence $x^{(0)}, x^{(1)}, x^{(2)}, \ldots,$

generated by the algorithm of Eq. (25) to Eq. (27), converges to the nonnegative minimizer of the expression in Eq. (23).

COMPARISON OF THE PERFORMANCE OF ALGORITHMS

Four very different-looking algorithms have been discussed previously, and the scientific literature is full of many others, only some of which are surveyed in books such as (1,5,26). Many of the algorithms are available in general purpose image reconstruction software packages, such as SNARK93 (17). The novice faced with a problem of reconstruction is justified in being puzzled as to which algorithm to use. Unfortunately, there is not a generally valid answer: the right choice may very well depend on the area of application and on the instrument used for gathering the data. This section contains only some general comments regarding the four approaches discussed before, followed by some discussion of the methodologies that are available for comparative evaluation of reconstruction algorithms for a particular application.

Concerning the two transform methods discussed, the linogram method (essentially an $N^2 \log N$ method) is faster than the FBP method (an N^3 method) and, when the data are collected according to the geometry expressed by Eqs. (9) and (10), then the linogram method is likely to be more accurate because it requires no interpolations. However, data are not normally collected in this way, and the need for an initial interpolation together with the more complicated looking expressions which need to be implemented for the linogram method, may steer some users toward the FBP method, despite its extra computational requirements.

The advantages of series expansion methods over transform methods include their flexibility (no special relationship needs to be assumed between the object to be reconstructed and the measurements taken, such as that the latter are samples of the Radon transform of the former) and the ability to control the desired type of solution by specifying the exact sense in which the image vector is to be estimated from the measurement vector; see Eqs. (17) and (23) . The major disadvantage is that it is computationally much more intensive to find these precise estimators than to perform a numerical evaluation of Eq. (1). If the model (the basis functions, the projection matrix, and the optimization criterion) is not well chosen, then the resulting estimate may also be inferior to that provided by a transform method. The recent literature has demonstrated that usually there are models that make the efficacy of a reconstruction provided by a series expansion method at least as good as that provided by a transform method. To avoid the problem of computational expense, one usually stops the iterative process involved in the optimization long before the method has converged to the mathematically specified optimizer. Practical experience indicates that this can be done very efficaciously. For example, as reported in (19), in the area of fully three-dimensional PET, the reconstruction times for FBP are slightly longer than for cycling just once through the data using a version of ART and spherically symmetric basis functions, and the accuracy of FBP is significantly worse than that obtained by this very early iterate produced by ART. Similar experience has been reported in the area of fully three-dimensional electron microscopy (27).

Because the iterative process in practice is stopped early, in evaluating the efficacy of the result of a series expansion method, one should look at the actual outputs rather than the ideal mathematical optimizer. Reported experiences comparing an optimized version of ART with an optimized version of EM (16,18) indicate that the former can obtain as good or better reconstructions as the latter, but at a fraction of the computational cost. This computational advantage results from not trying to use of all the measurements in each iterative step.

The proliferation of image reconstruction algorithms imposes a need to evaluate the relative performance of these algorithms and to understand the relationship between their attributes (free parameters) and their performance. In a specific application of an algorithm, choices have to be made regarding its parameters (such as the basis functions, the optimization criterion, constraints, relaxation parameters, etc.). Such choices affect the performance of the algorithm, and there is a need for an efficient and objective evaluative procedure for selecting the best variant of an algorithm for a particular task and for comparing the efficacy of different algorithms for that task.

An approach to evaluating an algorithm is to start first with a specification of the task for which the reconstructed image is to be used and then define a *figure of merit* (FOM) that determines quantitatively how helpful the image is and hence, the reconstruction algorithm, for performing that task. In the numerical observer approach (18,28,29), a task-specific FOM is *computed* for each image. Based on the FOMs for all of the images produced by two different techniques, one can calculate the statistical significance at which one can reject the null hypothesis that the methods are equally helpful for solving a particular task in favor of the alternative hypothesis that the method that has the higher average FOM is more helpful for solving that task. Different imaging techniques can then be rank-ordered on the basis of their average FOMs. It is strongly advised that a reconstruction algorithm should not be selected on the basis of the appearance of a few sample reconstructions, but rather that a study be carried out along the lines indicated before.

In addition to the efficacy of the images produced by the various algorithms, one should also be aware of the computational time and costs (possibly using special purpose hardware) needed for executing them. A survey from this point of view can be found in (2).

FIGURES OF MERIT

In this section, some examples of the FOMs mentioned in the last section are reviewed. Because this is a general discussion, only FOMs of general usefulness are presented; in a particular application, one may often wish to define FOMs that are much more task-specific (e.g., an FOM that measures how well a stenosis of a blood vessel can be seen in the reconstruction).

Resolution

In the reconstruction, it is generally desirable to be able to distinguish objects that are distinct in reality. To measure this ability, one can, for example, consider the following. If there are two identical small objects of uniform density (say, in PET, two small positron emitting spheres) near each other, then to be able to recognize them as separate, they need to be spaced at more than the full width at half maximum (FWHM) along the line connecting their centers. The FWHM is defined as the distance between the points on the two sides of the center of a single object at which the reconstructed density is half of the reconstructed density at the center of the object. If there are two objects, the reconstructed densities of each add up in the reconstruction, and if they are placed at less than the FWHM from each other, then the reconstructed value at the halfway point between their centers is likely to be actually larger than the reconstructed values at the centers themselves; this makes the two objects indistinguishable from a single elongated object. Because the FWHM in many situations depends on the direction of the line along which it is measured, it is advisable to report on its value for multiple directions, such as axial and tangential; see (30). The FWHM may also be location dependent, especially for fully three-dimensional modes of data collection.

Signal-to-Noise Ratio

Practical image reconstruction is inevitably a noisy process. Noise (defined in the very general sense, including both systematic effects and stochastic processes, as the difference between the actual reconstruction and what we would like to achieve in an absolutely ideal situation) is due to the counting statistics of the emitted or transmitted radiation, inaccuracies of the instrument, imperfections of the reconstruction algorithm, etc. It is clearly desirable to be able to distinguish actual signals in the reconstructions (such as radiation-emitting hot spots in the brain) from noise that may give rise to something of a similar appearance. The signal-to-noise ratio (SNR) is an FOM that measures this capability.

One way of defining SNR (30) is the following. Suppose that the values in the signal are higher than those in the background. Let V be the average reconstructed value in the region of the signal, and let B and σ_B be the mean and standard deviation, respectively, of the average reconstructed value in a number of regions (similar in size to the signal region) that are in the background. Then SNR is defined as $(V - B)/\sigma_B$. Such a measure is also not a single property of the reconstruction algorithm; for example, it is reported in (30) that the SNR is considerable larger when the signal regions are spheres 25 mm in diameter than when they are spheres 18 mm in diameter.

Artifact Limitations

Based on the general definition of noise given in the last subsection, "artifacts" are those components of the noise that are not due to statistical fluctuations. For example, a specific choice of the convolving function q in FBP may result in ripples coming off sharp edges of the image; see Fig. 8.8(a) in (1). Another choice of q may not suffer from such an artifact to the same extent, but usually there is a price to pay, such as reduced SNR; see Fig. 8.8(c) in (1), but compare also Fig. 8.9(c) with Fig. 8.9(a).

One kind of artifact that occurs frequently in practice is due to missing angles in the data collection. In the terminology of the previously introduced notation, this means that if the θ_is are arranged in increasing order, than there is an i for which $\theta_{i+1} - \theta_i$ is considerably larger than its typical value for $1 \leq i \leq I$ (using the definition that $\theta_{I+1} = \theta_1 + 2\pi$). Then, the reconstructions will typically appear to be blurred in the direction of the missing angles. One example of an application where this happens is the electron microscopic reconstruction of certain kinds of particles from data collected (due to the nature of these particles) using so-called conical tilting geometry. This is a three-dimensional mode of data collection, where projections are not collected for any of the directions within a cone of directions around an axis. This results in blurring along this axis, and the resolution in the direction of this axis can be measured by an appropriately defined FOM (27). Using such an FOM, it can be shown that some appropriate version of ART is better than some version of FBP from the point of view of blurring in the direction of missing angles. However, there are techniques to correct the output of FBP-like algorithms for data that have missing angles. For example, one can use prior knowledge, such as the region within which the object to be reconstructed must lie and the range within which the reconstructed values should be; see, Chapter 4 of Part I of (3).

SUMMARY

This article has covered some of the basic principles behind tomographic imaging, including such topics as the Radon transform, interpolation for backprojection (both modified cubic spline and sinc), the Ram–Lak filter, missing angle techniques, maximum entropy techniques, resolution, the signal-to-noise ratio, and artifact limitations. Further details may be found in the references.

BIBLIOGRAPHY

1. G. T. Herman, *Image Reconstruction from Projections: The Fundamentals of Computerized Tomography*, Academic Press, N.Y., 1980.
2. G. T. Herman, *J. Real-Time Imaging* **1**, 3–18 (1995).
3. G. T. Herman, H. K. Tuy, K. J. Langenberg, and P. C. Sabatier, *Basic Methods of Tomography and Inverse Problems*, Adam Hilger, Bristol, England, 1987.
4. P. Lauterbur, *Nature* **242**, 190–191 (1973).
5. A. C. Kak and M. Slaney, *Principles of Computerized Tomographic Imaging*, SIAM, Philadelphia, 2001.
6. G. N. Ramachandran and A. V. Lakshminarayanan, *Proc. Natl. Acad. Sci. U.S.A.* **68**, 2,236–2,240 (1971).
7. E. H. W. Meijering, W. J. Niessen, and M. A. Viergever, *Med. Imaging Anal.* **5**, 111–126 (2001).
8. G. T. Herman, S. W. Rowland, and M. -M. Yau, *IEEE Trans. Nucl. Sci.* **26**, 2,879–2,894 (1979).
9. E. A. Brigham, *The Fast Fourier Transform*, Prentice-Hall, Englewood Cliffs, N.J., 1974.

10. J. L. C. Sanz, E. B. Hinkle, and A. K. Jain, *Radon and Projection Transform-Based Computer Vision*, Springer-Verlag, Berlin, 1988.
11. P. Edholm and G. T. Herman, *IEEE Trans. Med. Imaging* **6**, 301–307 (1987).
12. G. T. Herman, D. Roberts, and L. Axel, *Phys. Med. Biol.* **37**, 673–687 (1992).
13. L. R. Rabiner, R. W. Schafer, and C. M. Rader, *Bell Syst. Tech. J.* **48**, 1,249–1,292 (1969).
14. P. Edholm, G. T. Herman, and D. A. Roberts, *IEEE Trans. Med. Imaging* **7**, 239–246 (1988).
15. R. M. Lewitt, *Phys. Med. Biol.* **37**, 705–716 (1992).
16. S. Matej et al., *Phys. Med. Biol.* **39**, 355–367 (1994).
17. J. A. Browne, G. T. Herman, and D. Odhner, *Technical Report MIPG198*, Dept. of Radiology, University of Pennsylvania, Philadelphia, 1993.
18. G. T. Herman and L. B. Meyer, *IEEE Trans. Med. Imaging* **12**, 600–609 (1993).
19. S. Matej and R. M. Lewitt, *IEEE Trans. Nucl. Sci.* **42**, 1,361–1,370 (1995).
20. R. Gordon, R. Bender, and G. T. Herman, *J. Theor. Biol.* **29**, 471–482 (1970).
21. A. Lent, in R. Shaw, ed. *Image Analysis and Evaluation*, Society of Photographic Scientists and Engineers, Washington, DC, 1977, pp. 249–257.
22. Y. Censor et al., in A. Wakulicz, ed., *Numerical Analysis and Mathematical Modelling*, Banach Center P, PWN — Polish Scientific, Warsaw, Poland, 1990, pp. 145–163.
23. L. A. Shepp and Y. Vardi, *IEEE Trans. Med. Imaging* **1**, 113–122 (1982).
24. G. T. Herman, A. R. De Pierro, and N. Gai, *J. Visual Commun. Image Representation* **3**, 316–324 (1992).
25. G. T. Herman, D. Odhner, K. D. Toennies, and S. A. Zenios, in T. F. Coleman and Y. Li, ed., *Large-Scale Numerical Optimization*, SIAM, Philadelphia, 1990, pp. 3–21.
26. F. Natterer, *The Mathematics of Computerized Tomography*, Wiley, Chichester, England, 1986.
27. R. Marabini, G. T. Herman, and J. M. Carazo, *Ultramicroscopy* **72**, 53–65 (1998).
28. K. M. Hanson, *J. Opt. Soc. Am. A* **7**, 1,294–1,304 (1990).
29. S. S. Furuie et al., *Phys. Med. Biol.* **39**, 341–354 (1994).
30. P. E. Kinahan et al., *IEEE Trans. Nucl. Sci.* **42**, 2,281–2,287 (1995).

U

ULTRASOUND IMAGING

NAVALGUND A. H. K. RAO
Rochester Institute of Technology
Rochester, NY

Ultrasound is defined as sound waves whose frequency is above 20 kHz, the upper limit of the human audible range. The past century has witnessed the growth of an imaging technology that uses ultrasound as the information carrier. A physical medium is required to sustain longitudinal and transverse ultrasound waves. The imaging can be passive, when the object to be detected is the source of sound, or active, when an external deliberately controlled source of sound is used and the echoes from objects in the medium are used to generate an image. State-of-the-art active systems can launch waves into a medium and nondestructively collect echo signals and display information pertaining to the inside of an optically opaque three-dimensional object. In medical applications, the propagation medium is soft tissue. Present day technology permits real-time (20 or more image frames per second) visualization of the gross anatomy and pathology. Medical ultrasound has found applications throughout the breadth of diagnostic investigations that includes obstetrics, gynecology, abdomen, ophthalmology, cardiovascular system, breast, and superficial structures. The frequency used can range from 1–10 MHz for imaging various large body parts and can go as high as 20–50 MHz for specialized imaging tasks where the required depth penetration is small such as in endoscopy, dermatology, or ophthalmology. Besides medical applications, ultrasound is also used in nondestructive evaluation (NDE) and sound navigation and ranging (SONAR). NDE is commonly used in the aerospace and nuclear power industries to inspect safety-critical parts for flaws during in-service use. The medium here is generally solid material. SONAR pertains to imaging under the sea. At the turn of the millennium, the medical application of ultrasound has emerged as the technology leader even though its use to detect internal defects in metal structures in 1940 preceded its first use in medicine. The imaging techniques and technologies used in these different application areas are quite similar, at least for medical and NDE. Therefore, the main body of the discussion will center on medical imaging.

An ultrasound imaging system usually conveys information about echo generating objects. The image can be thought of as a display of the multidimensional spatial distribution of some object-dependent entity. Visualization of this information is one of the objectives, and perhaps the major objective in medical ultrasound. The image value is generally related to the mechanical or elastic property of the object. A secondary objective therefore is to understand and quantify the relationship between the image value and the object property, also known as quantitative imaging or tissue characterization in medical ultrasound. This area still remains a subject of active research. This article deals only with the first objective.

Pulse echo imaging has become preeminent in medical ultrasound for reasons that will become clear later. The concept is illustrated by Fig. 1. Note that the figure depicts wave propagation only in two dimensions, whereas reality is three-dimensional. A finite size transducer located on the left side generates sound waves in response to electrical excitation and couples them to the anatomy under examination. The generated waves are of finite duration, and hence the disturbance in the tissue medium has a finite spatial extent in the propagative direction. This can be seen in a snap-shot of the disturbance in Fig. 1a, taken soon after transduction process is completed. Now, this disturbance propagates away from the transducer as time progresses. If the medium responds linearly, the propagative effects are governed by solving the space–time wave equation in free space (lossy or lossless) and at boundaries. The macroscopic effects are embodied in the familiar concepts of reflection, scattering, and refraction at the boundaries and diffraction in the homogeneous

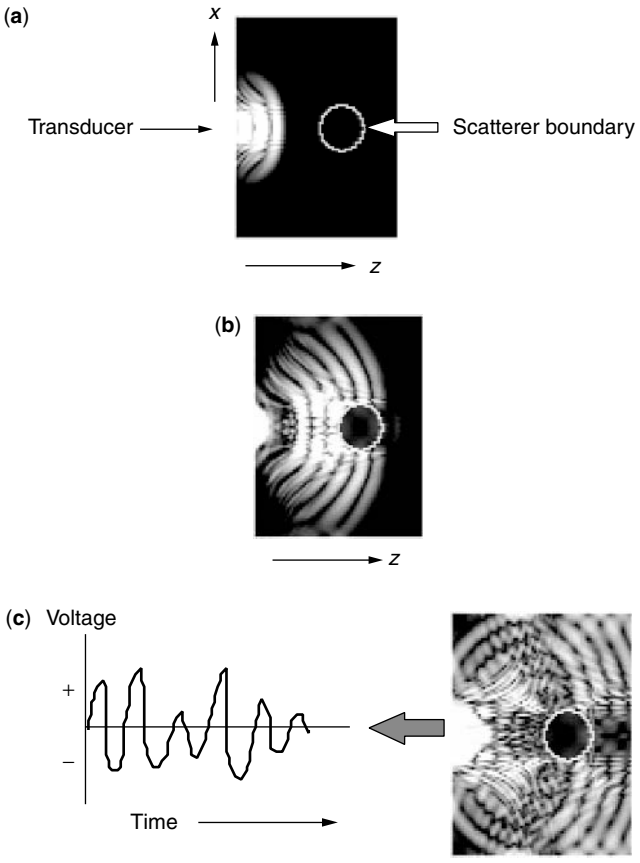


Figure 1. Physical processes involved in pulse echo imaging. Image data was generated using software WAVE 2000.

free space. Figure 1b shows a snapshot of the wave field after free space propagation to some depth z just in front of a scatterer boundary. Note that the spatiotemporal description of the wave field at any depth z is governed by diffraction, a phenomenon that plays an important role in determining the image quality in pulse echo imaging. Figure 1c depicts a later situation where the spatiotemporal echo field is seen propagating back toward the transducer. The transmitted primary field can also be seen propagating down along the z axis. This can generate additional echoes that arrive at the transducer later. Time encodes the range or depth of the scatterer. The echo pressure wave field incident on the same transducer generates an electrically time-varying voltage signal. The wave field incident on the face of the transducer is complex; it has a magnitude and phase at each point. If the voltage signal at any given instant, is considered proportional to the spatial integration of the complex wave field over the face of the transducer, we refer to it as phase-sensitive detection. The majority of transducers in use today are of this type. The detected voltage–time signal is referred to as the A-line signal. The three steps involved in the imaging process, namely, transduction and reception by the transducer, two-way propagation in the medium to some depth, and the scattering/reflection phenomenon will be considered in detail. But first we digress to consider different modes of scanning and information display employed in medical diagnosis and NDE.

SCANNING MODES AND IMAGE DISPLAY FORMATS

The A scan is simply the display of the transducer voltage–time signal generated by echoes. It is readily observed by connecting the received output to an oscilloscope triggered at the instant of initial excitation. If the ultrasound propagative speed c in the object is constant, time $t = 2z/c$ represents the range distance z. A large amount of information is contained in the radio-frequency (Rf) signal displayed in the A scan. The disadvantage of the A-mode is that it provides information about the object only along the line of sight and within the beam width of the transducer (Fig. 2). It is tedious

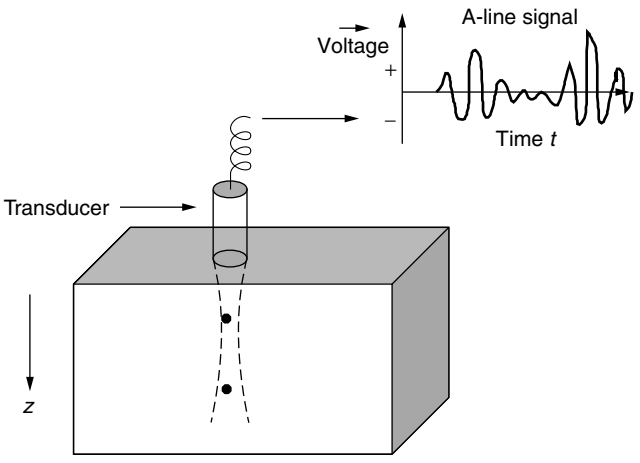


Figure 2. A-scan process.

and time-consuming to move the transducer around. Furthermore, a large amount of spatial information has to be interpreted quickly and kept in memory by the observer.

The B scan refers to a brightness scan. It is essentially a display of adjacent A-scan signals that are generated when the transducer is moved laterally along a line, say the x axis, as part of a linear scanning protocol (Fig. 3). However, only the envelope or amplitude of the oscillating voltage signal is used in the display. The signal values are mapped into gray-scale values that can be displayed as a range of brightness on a display device. The B-scan image represents a two-dimensional planar slice of the three-dimensional medium. The plane is determined by the transducer beam axis and the direction in which the scanning takes place. For example, the image in Fig. 3 is a x–z slice at some fixed y location. The scanning movement of the transducer was performed mechanically in the early 1970s. This is now largely replaced by electronic scanning using one-dimensional phased array transducers. The B scan is particularly suited for developing fast scanning techniques that can produce real-time images. For our present purpose, we may define real-time imaging technique as one that can acquire and display 20 or more image frames per second. For example, such high speeds are essential for adequate cardiovascular, abdominal, and fetal imaging. Qualitatively, the brightness at any given point in the image represents the scattering strength of the scatterers at that point in the chosen plane. Due to the finite size of the ultrasound beam, the plane has a finite thickness. We may define a point-spread function (PSF) here as the image of an infinitesimally small point scatterer located at some point in the plane that is being imaged. The PSF generally represents an image quality metric of any imaging system or modality. The spatial size of the PSF determines the resolution of that imaging system. The PSF for this modality varies spatially due to the divergence of the transducer beam. This will be considered in more detail later. B-scan imaging forms the basis for most of the medical ultrasound imaging techniques and their advancement (1).

The C scan refers to constant depth scanning. This technique also uses pulse echo signals, as illustrated in Fig. 4. Here, an electronically time-gated portion of the A-scan signal is used for a brightness-modulated display. A signal amplitude within a time gate centered at $t_0 = 2z_0/c$ only is used. The transducer has to be scanned laterally in the x, y plane for this modality. The mapping of the brightness extracted from the echo signal constitutes a two-dimensional image that provides a picture of the scatterers located in a thin slice in the x, y plane of the object at depth z_0. The thickness of the slice depends on the gate length. The PSF is spatially invariant. The advantage is that the transducer beam can be focused to a particular depth and optimal uniform resolution can be maintained everywhere in the image. A major disadvantage is that two-dimensional scanning requires considerable time for acquiring data. Large area coverage for real-time imaging is difficult now, even for one-dimensional phased array transducers. For this reason, it is rarely used for medical imaging but has found some applications in NDE. The

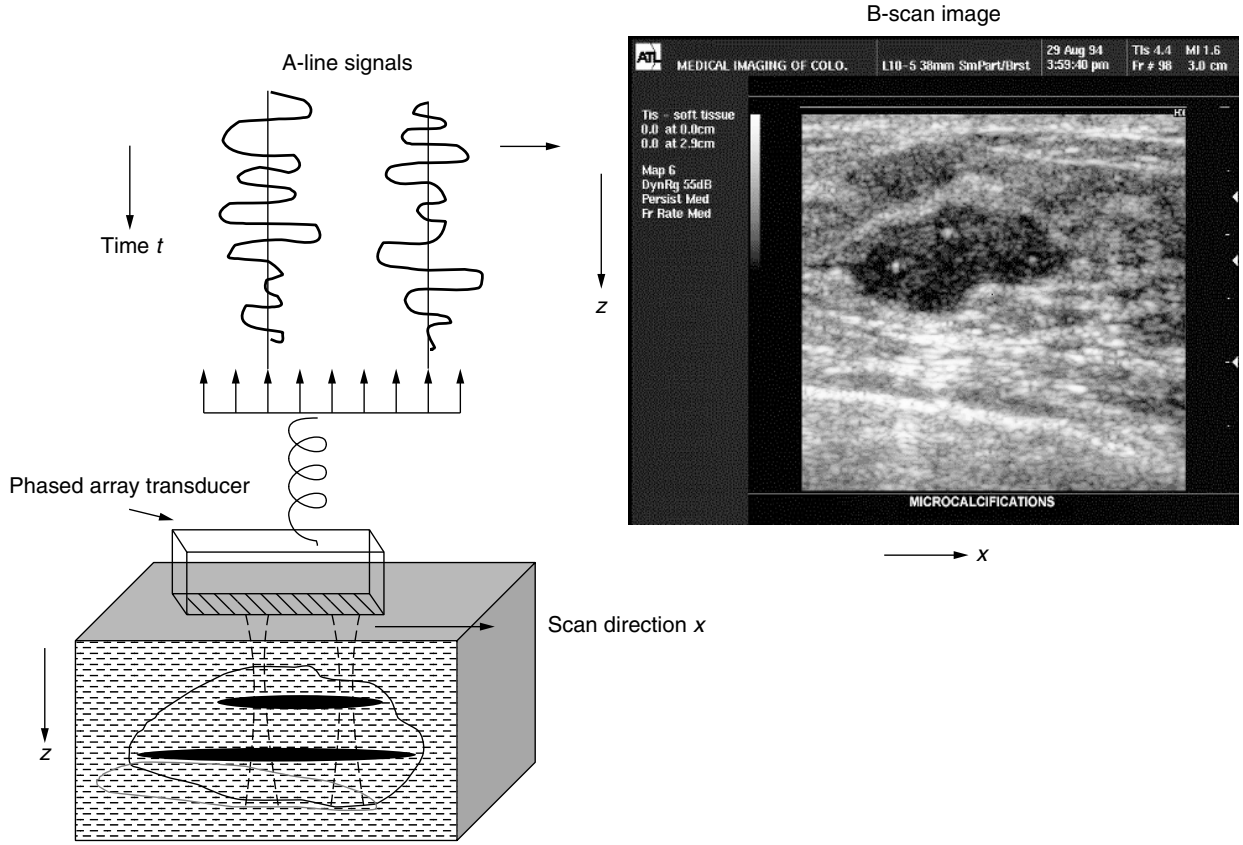


Figure 3. B-scan process.

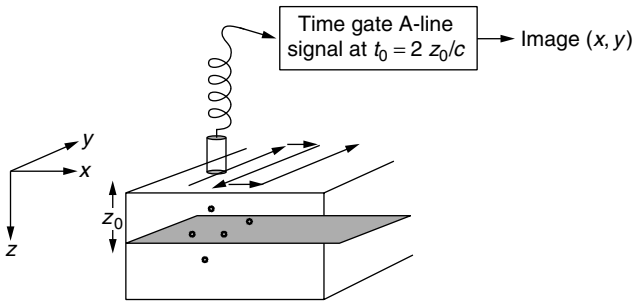


Figure 4. C-scan process.

future design of two-dimensional phased array scanners could possibly render real-time imaging feasible by this modality.

The M scan is used to monitor slow changes in a given A scan as a function of time. The transducer is held stationary at a fixed location, and the A-scans signals from the same line of sight are recorded at regular time intervals (T) (Fig. 5). The time-modulated brightness is displayed as a function of depth z (vertical axis) and T (horizontal axis). It is used to detect a pattern of slow movement in time of a reflector or scatterers along a fixed line of sight. The reflectors A and C in Fig. 5 are

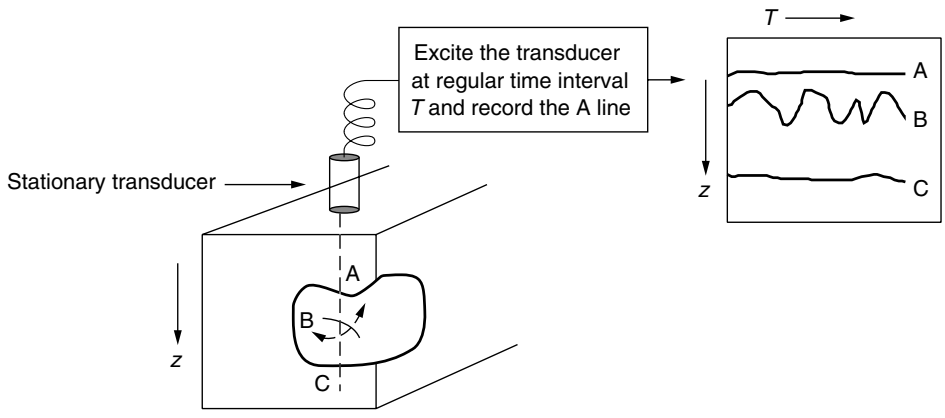


Figure 5. M-scan process.

stationary in time, and B is moving slowly. The motion of B can be studied by an M scan. A typical example of an application of M-mode scanning is examination of the pattern of movement of heart-valve leaflets (2).

ULTRASONIC PROPERTIES OF A BIOLOGICAL MEDIUM AND CONTRAST AGENTS

We will consider here those medium properties that are of major importance for medical ultrasound. The medium here therefore refers mainly to different types of soft tissue. This is essential for understanding the technology development, its limitations, and image interpretation. Contrast agents are now used clinically. Because these agents are injected intravenously into the body, a brief study of their properties is warranted here. The reader is referred to (3,4) for medium properties that are important in NDE.

Propagative Speed *c*

Soft tissue behaves as a fluid and therefore cannot sustain shear or transverse waves. We need consider only longitudinal or compressional waves. Conventional pulse echo image reconstruction is based on assuming a constant compressional wave propagative speed c in the human body. Fortunately, for soft tissue, 1,540 m/s is considered a good average value for c, and the range of variations for different tissues is generally less than 5% (see Table 1). Bone and air are exceptions, and consequently they do become a problem and a source of artifacts in medical ultrasound. Dispersion, that is, variation of c with frequency is negligible in soft tissue (less than 1% for frequencies in the 1–15 MHz range). Typical wavelengths can range from 1.5 mm at 1 MHz to 0.1 mm at 15 MHz. The finite speed also implies that all echoes from a body 15 cm thick can be received in less than 200 microseconds.

Characteristic Acoustic Impedance *Z*

This medium property is formally represented by $Z = \rho_o c = (\rho_o K)^{1/2}$ for the plane wave condition where ρ_o stands for equilibrium medium density and K is the compressibility. Table 1 provides a listing of values that are of interest in medical ultrasound.

Table 1. Ultrasonic Properties of Biological Tissues

Material	Speed, c $m\,s^{-1}$	Acoustic Impedance, Z $kg\,m^{-2}\,s^{-1} \times 10^{-6}$	Attenuation $db\,cm^{-1}\,MHz^{-1}$
Water	1,497	1.49	0.001
Air	330	0.0004	10 $dB\,cm^{-1}$ at 1 MHz
Bone	2,900	5.37	4.06
Liver, pig	1,570	1.67	0.11
Lung			40 $dB\,cm^{-1}$ at 1 MHz
Muscle	1,585		1.65–1.75 $dB\,cm^{-1}$ at 1 MHz
Fat, pig	1,479	1.39	0.30
Blood	1,580	1.67	0.11
Aluminum	6,419	17.33	0.006
Glass, Pyrex	5,640	12.63	0.001

Reflection/Scattering

This phenomenon is responsible for generating the echo signal from any point in the medium. We will start with the plane wave where it is assumed that the wave field is uniform in amplitude and phase across a plane that is perpendicular to the direction of propagation and this plane is of infinite extent (Fig. 6a). An interface is defined here as a planar boundary across which there is a *change* in the acoustic impedance from Z_1 to Z_2. When a wave meets an interface and its direction of propagation makes an angle θ_i with respect to the normal to the interface, in general, it will be reflected at angle θ_i and refracted at an angle θ_r. As in optics, reflection and transmission coefficients can be defined as

$$R = \frac{\text{reflected} - \text{amplitude}}{\text{incident} - \text{amplitude}} = \frac{Z_2 \cos\theta_i - Z_1 \cos\theta_t}{Z_2 \cos\theta_i + Z_1 \cos\theta_t}, \quad (1)$$

$$T = \frac{\text{transmitted} - \text{amplitude}}{\text{incident} - \text{amplitude}} = \frac{2Z_2 \cos\theta_i}{Z_2 \cos\theta_i + Z_1 \cos\theta_t}, \quad (2)$$

along with $\theta_i = \theta_r$ and $\sin\theta_i / \sin\theta_t = c_1/c_2$.

Even though the ideal situation described before is rarely manifested in soft tissue imaging, these equations are a starting point for deeper understanding. There are two different ways of looking at how and where the echoes are generated in tissue. An echo generating boundary can be defined as a surface across which a difference in Z exists. This thought process leads to a discrete model for the tissue where such boundaries, it is assumed, are distributed in three-dimensional space. The term scattering may be used when the incident wave energy is redirected in many directions instead of being specularly reflected or refracted. This can happen if the boundary is small compared to the wavelength or if its surface is rough. To understand scattering from tissue fully, it is essential to know what the scattering properties are for all possible boundaries of varying shape and sizes. Alternatively, one can assume that the acoustic impedance Z varies continuously in the medium. Its spatial derivative is the source of echoes. This can lead to a continuum

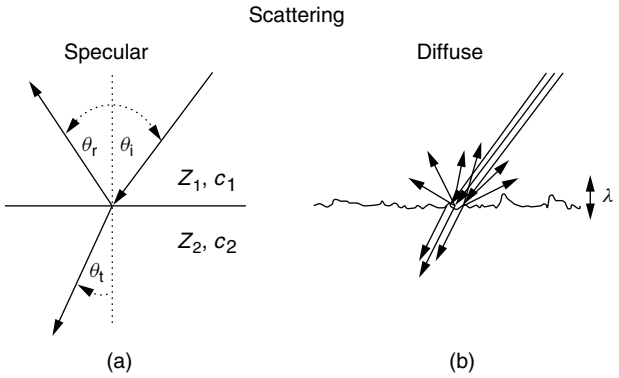


Figure 6. Specular reflection and diffuse scattering.

model where the inhomogeneous wave equation needs to be solved to determine the scattered signal (5,6). The difficult aspect of tissue as an echo generating medium is that our knowledge of the spatial variation of impedance in soft tissue is still in its infancy. It was surmised early that for frequencies in the range of 1–10 MHz, the collagenous connective tissues constitute the dominant boundaries (7). But much weaker echoes could be generated from blood and tissue cells as well. Theoretically, the scattering problem for boundaries that have deterministic shape and size, such as a point that is infinitesimally small compared to wave length, a sphere or cylindrical rod etc., has been solved (8). But for arbitrary shapes that are encountered in soft tissue imaging, the task is difficult. Quite frequently, a simplified stochastic model is assumed for theoretical understanding and for designing phantoms that mimic scattering from soft tissue. For example, the tissue could be modeled as a random distribution of point scatterers. The spacing, shape, and/or scattering coefficients, it may be assumed, are governed by some predetermined probability distribution functions (9–12). The tissue mimicking phantoms designed for imaging system quality control studies use small glass/plastic beads that are distributed randomly in the phantom, so that a B-scan image texture has an appearance similar to that seen in a clinical image. In what follows, we will proceed to interpret tissue scattering based on our current general understanding of scattering on different length scales.

Different kinds of scattering phenomena occur on different length scales "d" of the scattering structure relative to the wavelength λ (13). Structures within tissue that may scatter ultrasound can range from cells that are 10 microns in diameter to diaphragm, tissue–bone interface, or organ boundaries whose sizes are of the order of many centimeters. For the sake of discussion, we may classify them in three categories.

When $d \gg \lambda$ and the surface is very smooth, that is, the surface "roughness" is very small compared to λ, the beam is specularly reflected, as shown in Fig. 6a, and follows the law of reflection and refraction reasonably well. There are no such absolutely perfect reflectors in the body, though some organs such as bone surface or diaphragm may come close. It is more appropriate to interpret the reflection from such large-scale structures as diffuse, as shown in Fig. 6b. This can happen when the large-scale boundary is not smooth and surface irregularities are comparable in magnitude to λ. Diffuse scattering redirects the sound wave in many directions, including the backward direction, which we refer to here as backscatter. This fact is primarily responsible for visualization of such large-scale structures even, when they are not positioned perpendicularly to the ultrasound beam. Reflection or diffuse scattering is strong, and its frequency dependence is generally weak, $\approx f^0$. Except for the bone or air interface, the change in speed across the interface is usually small, so that refraction effects (beam bending) are negligible and the existence of a straight line of sight can be assumed for B-scan imaging.

The second category is defined by scatterers that possess the characteristic dimension $d \ll \lambda$. Blood cells and cells in tissue are some examples. Microbubbles that are used in contrast agents injected into the body for imaging also fit into this category. The scattering strength is generally very weak, microbubbles may be an exception because of a large acoustic impedance difference between tissue and air. Scattering in this regime is termed Rayleigh scattering, and it generally has f^4 frequency dependence (6). The angular dependence is fairly uniform, but backscatter may be predominant in some cases.

The third category corresponds to situations where the scatterer spacing is comparable to the wavelength, $d \approx \lambda$. This probably represents many tissue microstructures. It is also the most difficult to analyze because not much is known regarding what constitutes a scattering boundary in tissue parenchyma. It is a generally held hypothesis that the interfaces of collagenous connective tissue are dominant scatterers below 10 MHz (7). The microstructure formed by connective tissue appears in a variety of scales and shapes in the body. For example, evidence shows that the connective tissue matrix surrounding the lobular boundaries in liver tissue histology is the dominant scattering entity (11,12). The size or spacing is of the order of millimeters. Similarly, it has been argued that the collagenous septa surrounding cylindrically shaped muscle fibers and fascicles, known as the endomysium and perimysium, respectively, are the main contributors to scattering in muscle tissue (14). The diameters here can vary from 50 to 5,000 microns but the length can be in centimeters. In general, the scattering strength in this category is moderate but significantly weaker (10–30 dB lower) compared to the first category of nearly specular reflectors. In the 1–10 MHz range, the frequency dependence is $\approx f^m$ where $m \approx 1$–1.3; in other words, it is mostly linear. The scattering can be highly anisotropic, as in muscle fibers. Because the system resolution is on the order of $\lambda \approx$ a few millimeters, the echoes from adjacent scattering boundaries are not resolved, and an interference pattern of the echoes ensues. Due to the statistical variability inherent in this phenomenon, a noise-like texture can be seen in B-scan images. This is generally referred to as speckle (15).

Scattering characteristics are most conveniently expressed in terms of the scattering cross section. The reader is referred to the references cited for a more detailed model-based treatment of this topic (1,5,6). A model-based spectrum analysis of the radio-frequency (Rf) echo signal can provide some information about the tissue scattering microstructure. Clinical applications still remain a potentially useful research endeavor (16). Because of the nondeterministic nature of the tissue scattering microstructure, several researchers have tried to model it as a discrete random process and attempted to extract tissue specific information from the statistical analysis of the echo signal (9–12, 17–19). Though a great potential for characterizing tissue structure lies here, the understanding still remains in its infancy.

Attenuation

As a wave propagates in a medium, its amplitude or intensity can decrease with distance due to two types of losses: absorption of energy in the medium and loss of energy due

to scattering. The absorption mechanisms in soft tissue are quite complex and they show a frequency dependence (1). The loss due to both mechanisms collectively is termed attenuation. The wave intensity I_0 decreases exponentially to a value I_z as it propagates a distance z: $I_z = I_0 e^{-\mu \cdot z}$. An attenuation coefficient μ for soft tissue can be defined from measurement as

$$\mu(\text{dB cm}^{-1}) = -\left(\frac{1}{z}\right) 10 \log_{10}\left(\frac{I_z}{I_0}\right). \tag{3}$$

For many soft tissues, $\mu \approx \alpha f$, that is, the frequency dependence is nearly linear. Table 1 lists the values of α for several different tissues. A good average value of α for soft tissue turns out to be around 0.5 dB/cm/MHz. Note that the value for lung and skull bone is much higher and for water it is much lower. Attenuation poses a challenge for ultrasound imaging system designers. The total cumulative attenuation suffered by a wave that has to travel to a depth of 15 cm and back in tissue can be as high as 75 dB for a 5-MHz signal. The demand on the dynamic range to display all of this information is too high. The signal processing to meet this demand will be examined later. Attenuation and its frequency dependence are tissue parameters. Methods have been explored to extract this information from the echo signal for tissue characterization (16,18). The methods require the raw A-line radio frequency (Rf) data before it has undergone any envelope detection processing. The technical difficulties for *in vivo* data stem from (1) the need to correct for transducer diffraction effects and (2) the need for ensemble averaging across a finite volume of tissue to reduce the speckle interference artifact.

Nonlinearity Parameter B/A

Derivation of the basic differential equations that embody harmonic wave distortion for plane wave propagation in one dimension can be found in many texts (20). The following wave equation describes the nonlinear response of a medium to the wave amplitude or equivalently, the particle displacement $\xi(x, t)$in space and time:

$$\frac{\partial^2 \xi}{\partial t^2} = \frac{c_0^2}{\left(1 + \dfrac{\partial \xi}{\partial x}\right)^{(B/A)+2}} \frac{\partial^2 \xi}{\partial x^2} = \frac{c_0^2}{K\left(\xi, \dfrac{B}{A}\right)} \frac{\partial^2 \xi}{\partial^2 x}. \tag{4}$$

Note that c_o is the wave velocity when the nonlinear response is absent or can be ignored. Clearly, when the amplitude ξ is not infinitesimally small, the propagative speed begins to depend on the wave amplitude through its gradient $\partial \xi / \partial x$. The ratio B/A is the nonlinearity parameter of the medium. For nonlinear propagation, $K(\xi, B/A)$ deviates from its limiting value of one. The implication is that for a single frequency wave of finite amplitude, different parts of the wave propagate at a speed slightly different from c_o. The higher amplitude portion of the wave front will travel at a higher propagative speed than those portions of the wave front of lower amplitude. Consequently the single-frequency waveform is distorted. Through Fourier analysis, it can be shown that a distorted albeit periodic wave possesses harmonic components.

The Fubini solution (21) for a harmonic source signal that has a first-order approximation provides a harmonic description of the distorted wave as

$$\frac{u}{u_o} = 2 \sum_{n=1}^{\infty} \frac{J_n\left(\dfrac{nz}{l}\right)}{\left(\dfrac{nz}{l}\right)} \sin[n(\omega t - kz)] \tag{5}$$

where

$$l = \frac{c_o^2}{\left(\dfrac{B}{2A} + 1\right) \omega u_o},$$

where u_o is the initial wave amplitude, B/A is the nonlinearity parameter of the medium, c_o is the small amplitude velocity, $k = \omega / c_0$ is the wave number, and l is the propagative distance at which a shock wave is formed, that is, when the distorted wave assumes a sawtooth form. The solution is valid for $z < l$. The index n stands for the harmonics. The amplitudes of the harmonics depend on several system and medium-dependent parameters. The B/A ratio for water is 5, for blood it is 6, for fat it is 11, and for most soft tissue it is around 8 (6). Recent developments of a technique, widely known as harmonic imaging, has its basis in this phenomenon. Although the B/A ratio may be a tissue characterization parameter, this has not been investigated extensively.

As far as B-scan imaging is concerned, the properties of the biological media considered before pose several major implications and challenges for system designers. First, the echo signal strength can vary from 20–40 dB due to the different classes of scatterers in the tissue. Second, a speckle-like texture in the image is an accepted fact and is often considered desirable because it carries qualitatively useful information. Third, the predominantly diffuse nature of scattering in tissue ensures that, in general, there will be some backscattered signal from almost all types of boundaries. This has made the simple B-scan pulse echo mode of imaging with a single transmit-receive device a useful technique in medical ultrasound. Fourth, refraction at many soft tissue boundaries, is negligible. Therefore, no major correction is needed in most cases for proper registration of echoes in the B-scan image. In some situations, however, refractive artifacts can be observed (2). Fifth, due to the generally weak backscattering scenario, multiple reflection/scattering can be considered negligibly small. Again in some special cases artifacts can be observed when this is not the case (2). Sixth, a constant speed of propagation, it can be safely assumed, converts arrival times into depth. Again, only in some special, cases artifacts due to misregistration can occur. Seventh, the attenuation in tissue increases linearly with frequency. A typical value of 0.5 $\text{dB cm}^{-1}.\text{MHz}^{-1}$ has many implications. For body cross-sectional imaging, assuming a required depth range of 20 cm and transducer frequency of 4 MHz, the echoes from the deepest reflector can be 80 dB below that from the nearest one. Display of information over such a large dynamic range is a challenge. At still higher frequencies, the signal-to-noise ratio (SNR) can become very poor at large depths. Therefore, the attenuation in tissue sets the

upper limit on the usable frequency and the maximum depth that can be imaged with reasonably good SNR. Typically, imaging at frequencies higher than 5–7 MHz is possible only for superficial tasks that do not require large depth penetration. Eighth, harmonic signal generation due to nonlinear propagation in tissue is predominant at the pressure amplitudes that are commonly employed in medical ultrasound. Its clinical usefulness has been identified only recently. The details are considered in a later section under the heading of harmonic imaging.

Doppler Effect

The Doppler effect is a phenomenon in which an apparent change in the frequency (and wavelength) of sound is observed if there is relative motion between the source of sound and the receiver or scatterer. If the source is moving toward the receiver, the receiver is moving toward the source, or the scatterer is moving toward the source and the receiver (as in pulse-echo imaging), the received wave has higher frequency than would be experienced without the motion. For the pulse echo, the shift in the frequency f_D, commonly referred to as the Doppler shift, is given by

$$f_D = \frac{2vf\cos\phi}{c}, \tag{6}$$

where v is the velocity of the scatterer or the interface, f the frequency of the transducer, c the speed of sound in the medium, and ϕ is the angle between the direction of v and the sound propagative direction (see Fig. 24). If the motion is away (receding), the received wave has a lower frequency, and f_D is negative.

For medical applications, it was realized early that blood cells represent a source of moving scatterers and therefore can produce a measurable Doppler shift (1,2,13). For physiological blood velocities ranging from 20 to 200 cm/s and f from 2 to 7.5 MHz, the Doppler shift typically falls in the audible range (200 kHz to 10 kHz). Doppler ultrasound imaging technology is based on this phenomenon and continues to play an important role in medical ultrasound. The technology development to this date is summarized in a later section under the heading of Doppler ultrasound.

Contrast Agents

Ultrasound contrast agents (UCA) are now a clinical reality. They generally consist of air-filled micron-size gas bubbles that can be injected into the blood stream. The interaction between ultrasound and the contrast agent is still a subject of vigorous research. What is presented here are some general statements about the current understanding of bubble response when the bubble diameter is very small compared to the ultrasound wavelength. The response of a gas bubble to an ultrasound wave depends on the acoustic pressure amplitude and can be divided into three regimes. At small amplitudes, the relative compression and expansion of a bubble is linearly related to the instantaneous pressure. This is the linear regime that has moderate scattering at the insonifying frequency. At higher amplitudes, we enter a regime of nonlinear response. Bubble compression generally retards relative to expansion. Consequently, the bubble size is not linearly related to the applied acoustic pressure, and the bubble vibration contains second and higher multiples of the insonifying frequency (22). Thus, the backscatter signal contains the fundamental (transmitted) frequency, and also harmonic frequencies. If the amplitude is increased further, the scattering levels of most contrast agents increase abruptly for a short time. Now, it is believed that this is associated with bubble rupture and release of free gas bubbles (23). This irreversible effect is often termed transient cavitation and the scattering in this regime becomes highly nonlinear. Typical values for the range of pressure p for the three regimes are p less than 50 kPa for linear scattering, p between 50 kPa and 200 kPa for nonlinear harmonic regimen, and p between 0.2 MPa and 2 Mpa for transient scattering. Recent research efforts have explored all the three regimes for possible imaging applications. A summary of these efforts is given in a later section under the heading of contrast imaging.

INSTRUMENTATION AND PROCESSING

The basic engineering principles of B-scan imaging are described here starting with signal generation/reception followed by major elements of the signal processing in an ultrasound scanner. This is followed by a description of a number of technological improvements that have taken place in recent years.

Generation and Detection Transducers

The central function of the transducer is to convert an electrical signal to pressure waves that couple to the medium and vice versa. However, from the standpoint of image quality, the transducer is the single most important component. Its design parameters influence the resolution and performance of an imaging system. Figure 7 illustrates the main features of a single-element transducer. The active element, which is a piezoelectric material, is a thin plate that is responsible for the conversion. Its operation is based on the well-known piezoelectric effect. If a voltage of some polarity is applied across the two faces of a plate via gold-plated electrodes, the plate thickness increases in response, due to reorientation of the electric dipoles in the element.

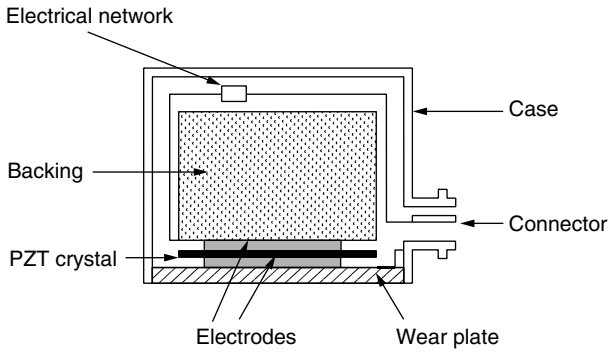


Figure 7. Components of a single element transducer.

If the polarity is reversed, the plate thickness can be made to decrease from its equilibrium thickness. Thus, the plate can be made to vibrate at some frequency by applying a sinusoidal voltage. Conversely, the application of a stress due to the incoming echo wave produces a sinusoidal voltage across the plate surface that can be picked up by sensitive electronics. The most commonly used material is lead zirconate titanate (PZT) which is commercially available. For details about other usable materials and their properties such as quartz, polyvinylidine difluoride (PVDF), and other newly developed PZT-polymer composites, the reader is referred to references cited in (1).

The thickness of the PZT crystal determines its resonant frequency f_o, the first among many others. By considering the two surfaces of the piezoelectric crystal as two independent vibrators, it is possible to see that the resonant frequencies are those where two waves add in phase. The thickness selected is one-half of the sound wavelength in PZT, considering that another 180° phase change takes place at the front surface. For example, for PZT, the thickness (in mm) $= 2/f_o$ (in MHz). The frequency response $T(f)$ of the transducer drops off above and below its resonant frequency, as shown in Fig. 8. Given the $T(f)$ of a transducer, its inverse Fourier transform, $F^{-1}[T(f)] = p(t)$, defines a time domain pulse $p(t)$. Considering the transducer as a linear system, it is possible to interpret $p(t)$ as the system impulse response. In other words $p(t)$ is the short pulse emitted by the transducer when the input electrical drive signal is a very short rectangular voltage pulse (mathematically called an impulse or a delta function) (24). Because a pulse of short duration is desirable for pulse-echo imaging, the transducers are normally excited by such impulse signals. The transducer quality factor, defined as $Q = f_o/\Delta f$, is an important parameter, where Δf stands for the -6 dB bandwidth of $T(f)$. As we will see later, the shorter the duration of a pulse, the better the resolution along the transducer axis. The pulse duration and the bandwidth Δf are inversely related (24). For superior resolution, the bandwidth has to be increased. This is achieved by a combination of two approaches: attaching a highly attenuating backing material to the back of the piezoelectric element to damp the vibrations and by electrically tuning the network, as shown in Fig. 7. This, of course, comes at a cost, a decrease in energy transfer efficiency. Typical Q values for medical transducers are around 2–4. PZT has an acoustic impedance that is 14 times that of water or soft tissue. Therefore, a matching layer is added to the front surface that also serves as a protective wear plate. As we shall see later, the shape and size of the element (circular, rectangular, etc.) combined with the dominant wavelength determines the lateral resolution that is limited by the diffractive effect. Focusing can be achieved by introducing curvatures into the element (1,2,13).

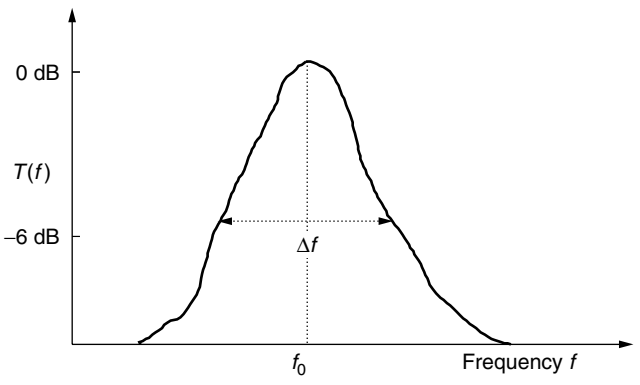


Figure 8. Typical frequency response of a transducer.

Processing the Echo Signal

Figure 9 illustrates the essential components of a B-scan pulse-echo imaging system. The transducer is driven by a very short duration, high voltage (100–400 volts) signal. To ensure synchronization with the lateral scanning motion, the timing is controlled by a trigger from a clock. The pulsed excitation has to be repeated for every A-line data acquisition. This pulse repetition frequency (PRF) must be chosen so that sufficient time is available to receive echoes from the farthest reflector of interest. For example, if we wish to image up to a depth of 15 cm, then the PRF^{-1} must be larger than 195 μs. Typical PRFs are of the order of 1 kHz. The A-line radio-frequency (Rf) echo signals are generally weak, so they require low-noise, low-voltage Rf preamplification. A transmit-receive circuit is used to protect the preamplifier from high voltage transmit signals (25). As illustrated in Fig. 9, a typical signal at this point has a dynamic range (ratio of the weakest to strongest echoes) of around 70–80 dB. Attenuation accounts for a 30–50 dB loss. These losses can be seen on an oscilloscope as an exponentially decreasing echo signal amplitude with time. All scanners have some sort of time-gain correction (TGC) circuit that attempts to correct approximately for this attenuative loss by applying time (depth)-dependent amplification (25).

Even after this correction, the remaining dynamic range is 40–50 dB. This is now mostly due to the differences in backscatter within the tissue microstructure. This is too large for a display monitor to handle and for the human visual system to comprehend. Therefore, the signal further undergoes compression amplification. This is usually done by a non linear amplifier that amplifies weak echo signals more than strong ones. The dynamic range can be brought to within 20–30 dB. At this point, the entire Rf echo signal time series is considered a slowly varying amplitude modulated carrier wave at frequency f_o (transducer center frequency). Only the slowly varying amplitude information is displayed in a B-scan image. Therefore, the signal is demodulated by using any of a number of techniques borrowed from communication engineering theory (26). For example, full or half-wave rectification followed by low-pass smoothing (filtering) is often performed by dedicated analog circuits. This envelope signal constitutes the A-line video signal that is digitized and stored for display. Once scanning is completed by, say, the lateral movement of the transducer in the x direction (see Fig. 3) and all of the A-line signals have been acquired, a static B-scan image $I(x, z)$ is produced by simply displaying the adjacent A-line values side by side in gray scale. Time encodes the depth $z = t/2c_0$. Analog and digital scan

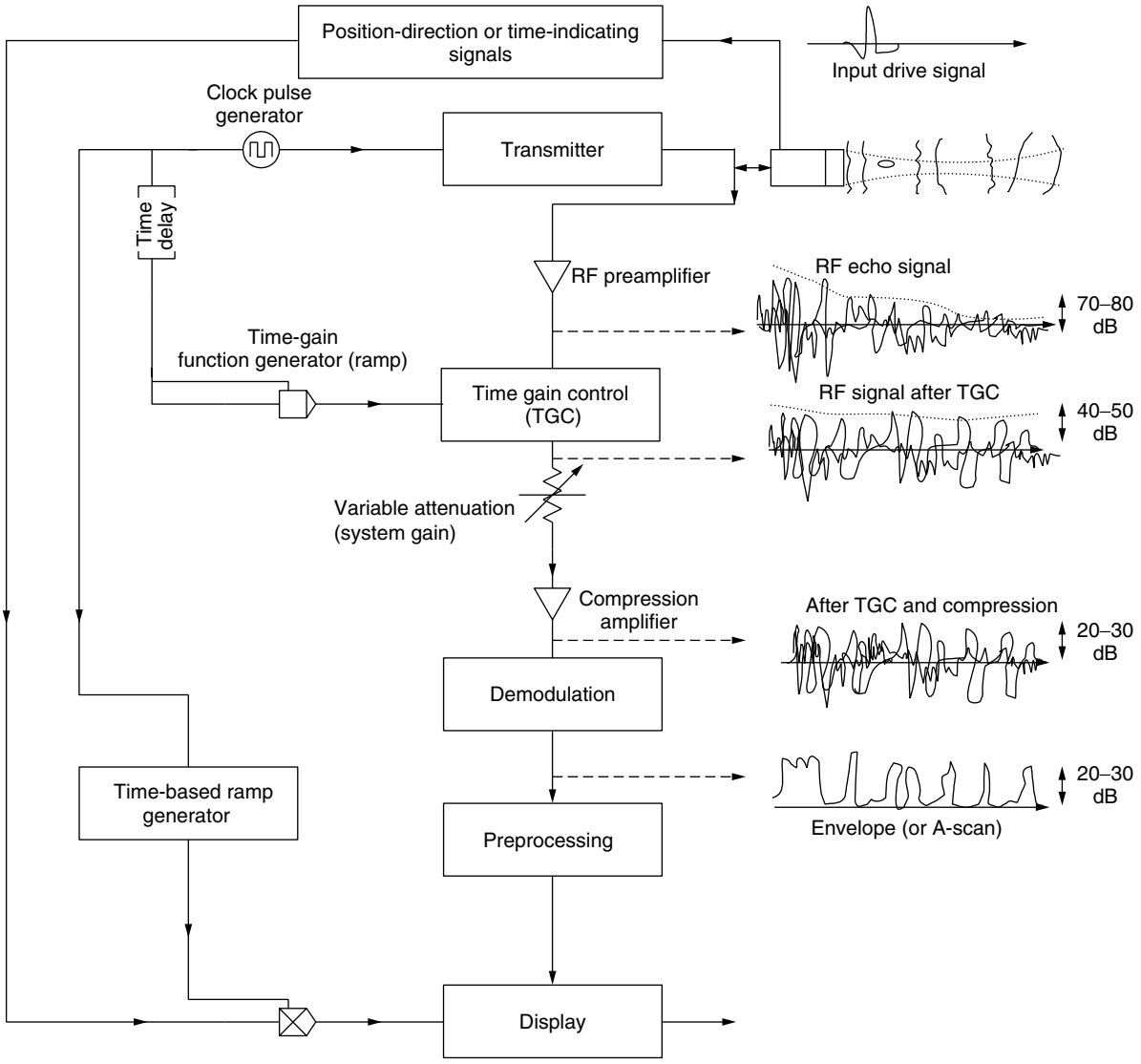


Figure 9. Block diagram depicting the essential echo signal processing in an ultrasound scanner. The modified echo signal is schematically shown to the right at various points in processing. In the schematic, time increases to the right. (From Bamber et al. (13)).

converters have been used in the past as a mapping and registration interface to the display unit (25). Further nonlinear gray value transformation, generally referred to as preprocessing, may be used under operator control before the display stage. The reader should bear in mind that many variations in the signal processing chain are possible and are commonly adopted in scanner design.

IMAGE QUALITY METRICS

Speckle

A typical ultrasound image from a tissue-mimicking phantom is shown in Fig. 10. The mottled or granular gray-scale structure seen in the image is called speckle. Speckle results from the random scattering microstructure of the tissue and the coherent nature of the transducer detection process (15). For a basic understanding of the process, consider the simple example in Fig. 11. Echoes from four closely spaced scatterers that are positioned randomly in the ultrasound beam can superimpose and produce an A-line envelope signal that shows seemingly random variations in amplitude with echo arrival time. Upon lateral scanning, these scatterers are imaged as randomly fluctuating gray values in the B-scan image. Note that because of the underlying interference phenomenon, there is no one-to-one correspondence between a high brightness value seen in the B scan and any particular scatterer location. In other words, the B-scan image does not turn out to be simply four distinct bright spots, one corresponding to each scatterer. This leads to an implication that for a speckle pattern to exist, the interscatterer spacing must be much smaller than some resolution dimensions of the B-scan imaging system. These dimensions are defined by the imaging system's point-spread function (PSF)

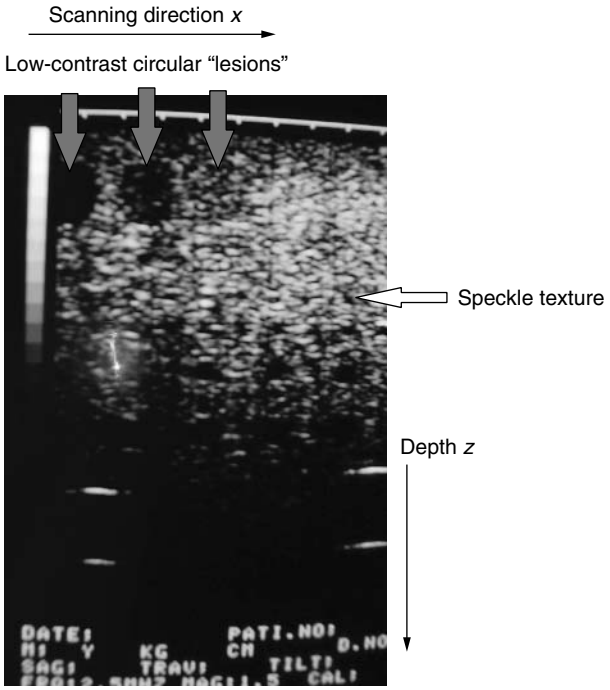


Figure 10. B-scan image of a tissue-mimicking phantom.

which we will consider shortly. It is generally believed, therefore, that speckle results from the random scattering microstructure when the interscatterer spacing is much smaller than the resolution dimensions of the imaging system. From our previous discussion on scattering, we understand that it is difficult to know this microstructure deterministically. Even more difficult, if not impossible, is the inverse problem: to determine the microstructure uniquely from the B-scan image. Nevertheless, given that speckle does carry some information about the tissue microstructure, several researchers have attempted to extract statistical information from the speckle for tissue characterization (9–12,17–19). These methods are not deterministic, and therefore we have to assume some statistical model for the scattering microstructure.

To a radiologist who is used to looking at images from other medical imaging modalities, speckle appears as fine randomly fluctuating gray values superimposed on an otherwise uniform background. Bear in mind that speckle is often a very low-level signal that is amplified nonlinearly in the processing stage. But for the speckle, it would be difficult to visualize low-contrast lesions in tissue. In Fig. 10, some circularly shaped low-contrast "lesions" can be visualized in the presence of speckle. Here, we will consider speckle from the standpoint of its role in low-contrast detection.

It is easier to recognize the quality (subjective evaluation) of an image than to define it. But subjective evaluation is not always convenient and is task-dependent. On the other hand, it is easy to define some objective image quality metrics that depend mainly on the design parameters of the imaging system (27,28). The relationship between objective and subjective evaluation is not well understood at present. We will consider only the objective metrics here. Some have been defined and have been in use for many years. Some others, such as contrast resolution and uniformity, have come into existence more recently as the use of computer technology continues to bring subtle improvements in image quality.

Point-Spread Function: Lateral and Axial Resolution

The concept of a two-dimensional point-spread function, $\mathrm{PSF}(x, y)$, in the theory of imaging systems analysis is well established (24,29). Extending it to three-dimensional

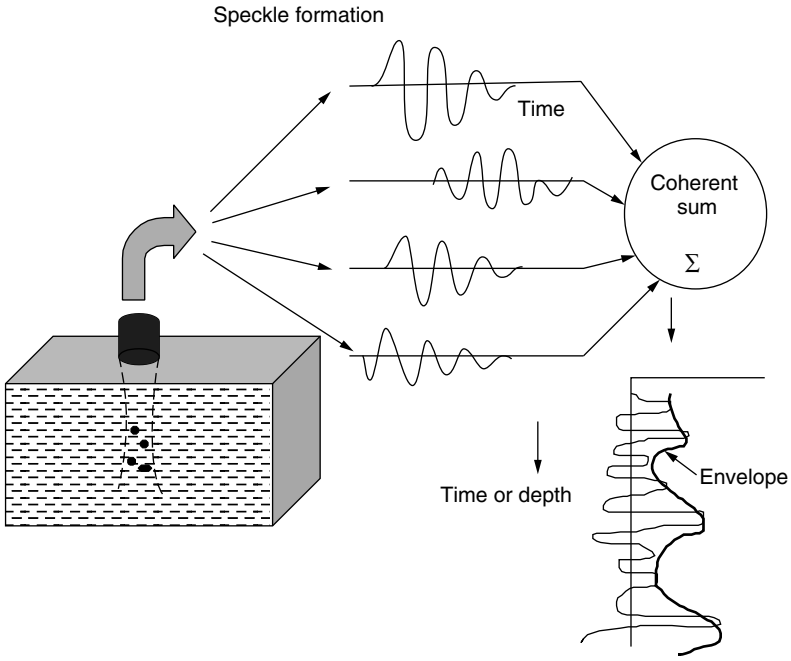


Figure 11. Coherent summation as a basis for speckle formation.

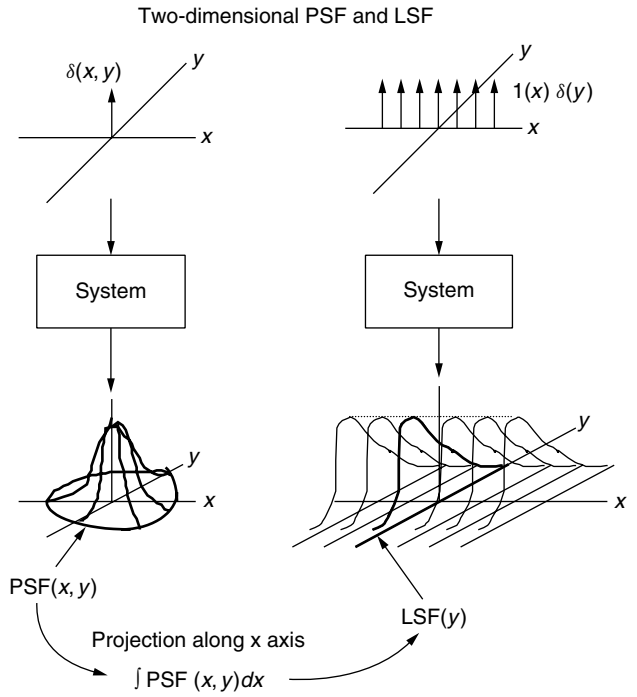


Figure 12. Two-dimensional point-spread function and its relationship to line-spread function.

ultrasonic imaging system analysis has remained a subtle distinction (4,29). Consider a two-dimensional imaging system as a black box with an access to its input (object) and the output (the final B-scanned image of the object). Figure 12 illustrates the input–output relationship. If the input is a point object [mathematically a two-dimensional delta function $\delta(x, y)$], then by definition, the output image is called the system point-spread function, $\mathrm{PSF}(x, y)$. Similarly if a line of delta functions (a thin line object) aligned along the x axis is the input, the resulting image is called a one-dimensional line-spread function, $\mathrm{LSF}(y)$. Mathematically, $\mathrm{LSF}(y) = \int \mathrm{PSF}(x, y)\, dx$. In general, it is an angular projection of the $\mathrm{PSF}(x, y)$ in a direction that is given by the line object (24).

This concept can be extended to a B-scan imaging system in a lossless medium. Figure 13 illustrates a rectangular element transducer that scans a point target located at some distance z_0. A hypothetical beam profile $h(x, y)$, chosen as a two-dimensional rectangular function for the sake of illustration, is also depicted. The exact form of $h(x, y)$ is determined by solving the diffraction problem (more on this later). Scanning can be performed by moving the transducer in the x, y plane. The resulting image, as shown, can be defined as the true three-dimensional $\mathrm{PSF}(x, y, z)$. We assume for simplicity that the PSF is separable into an x, y component that depends on the beam profile and a z component that depends on the envelope of the received temporal pulse $|p(t)|$ or equivalently, $|p(z)|$:

$$\mathrm{PSF}(x, y, z) = [h^2(x, y)]|p(z)| \tag{7}$$

The beam profile term appears as a square because transmission and reception are by the same transducer and the reciprocity principle holds for a beam profile in reception. A slice of the PSF along the z axis is shown for both the Rf and the envelope detected signal.

The three-dimensional PSF is not measured on clinical scanners. Most quality control phantoms consist of wire targets located at different depths and are held perpendicularly to the beam axis. Scanning is generally performed in a direction that is mutually perpendicular to both the beam axis and the wire (2). By generalizing the

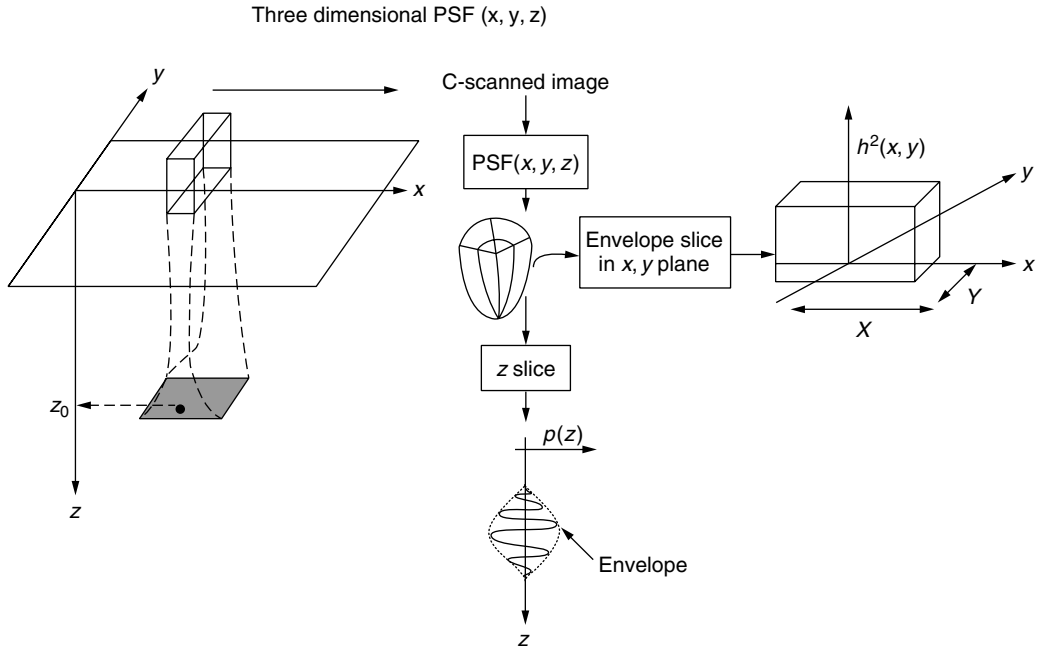


Figure 13. Components of a three-dimensional point-spread function.

two-dimensional projection concept to three dimensions, it is possible to show that the scanned image of a wire target is at best an estimate of one particular projection of the three-dimensional PSF(x, y, z). For the geometry and the example described, if the wire is along the y axis and scanning is along the x axis, the two-dimensional image $I(x, z)$ will be approximately represented as

$$I(x, z) = |p(z)| \left[\int h^2(x, y) dy \right] = |p(z)| L(x) \qquad (8)$$

where $L(x)$ is the projection of $h^2(x, y)$ along the x axis (24). Similarly if the wire targets are along the x axis and scanning is along the y axis, we will get a projection of $h^2(x, y)$ along the y axis. Either way, it is not the cross section but the projection of the beam pattern that is measured. Note that for a linear array scanner that uses an array of rectangular elements lined up, say, along the x axis, y is referred to as the elevation direction, and the beam width in that direction represents the thickness of the slice of the selected plane. $I(x, z)$ or $I(x, t)$, the image of a wire target located at $z = z_o$ shows up as a smudged blob in the brightness plot (B-scan image in gray scale), is generally accepted as the two-dimensional PSF(x, z) defined for the depth z_o.

Figure 14 shows a B-scan image of a phantom that contains several wire targets at different depths. The scanning was perpendicular to the wire target along what we may consider as the x axis. Disregard the faint speckle pattern. Note the finite shape and size of the individual bright blobs. They represent the B-scan image of the wire targets in the phantom. Each may be considered to represent the system PSF $= I(x, z)$ for the appropriate depth z_o. Two spatial resolution parameters can be extracted from $I(x, z)$. The 6-dB width (one-half amplitude) of $L(x)$ represents the lateral resolution. Similarly, the 6-dB width of the time pulse envelope $|p(t)|$ or $|p(z)|$ represents the axial resolution. Two wire targets

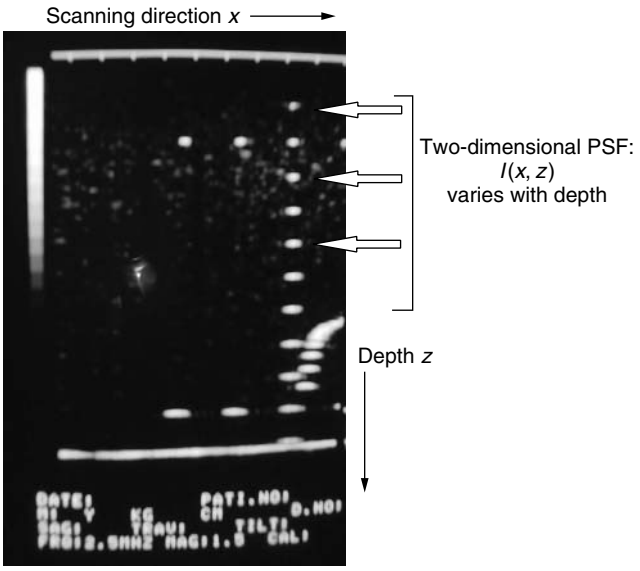


Figure 14. B-scan image of a wire target phantom.

separated by a distance smaller than the lateral resolution distance in the scanning direction will not appear as two separately distinguishable blobs. Similarly, two wire targets separated by a distance smaller than the axial resolution in the beam propagative direction will not be imaged as two separate distinguishable blobs. Typically, the axial resolution distance is much smaller (0.3 to 1 mm) than the lateral resolution (1 to 5 mm). In noncircularly symmetrical geometry, a third parameter that corresponds to the thickness of the slice may also be defined and measured.

In summary, we observe that the PSF does not have a symmetrical shape because its shape and size in the two orthogonal directions, the axial and lateral, depend on different factors. The axial resolution depends on the time width of the short pulse. Recalling from our earlier discussion, this time width is inversely proportional to the transducer bandwidth. Due to the frequency-dependent attenuation as the disturbance propagates in the tissue, the bandwidth of the pulse may change from that excited at the transducer. Therefore, axial resolution may change with depth. The lateral resolution depends on the transducer beam width at any given depth. Consequently it depends on the transducer element shape, dimensions, and the diffractive phenomenon that unfolds as the beam propagates in the medium. Thus we observe that, in general, the PSF is spatially variant. In other words, it is different for different locations of the wire target. Typical values for the resolution can range from few millimeters at low megahertz frequencies to a fraction of a millimeter at 10 MHz and above.

Contrast Resolution

This is a metric that attempts to evaluate the spatial extent of the side lobes in the PSF which may be 20 to 60 dB below the central peak value. Therefore, it requires examination of the PSF or the lateral and axial profiles on a logarithmic scale plot (27,30). We will examine the structure and magnitude of side lobes in a later section (see Fig. 17 for example). The reason for such emphasis on side-lobe structure becomes clear if we recall two major points. First, we know that speckle comes from low-level echoes from all scatterers located within the three-dimensional region spanned by the PSF, including its side lobes. These echoes are preferentially amplified in a nonlinear amplifier during B-scan signal processing. Second, the B-scan imaging scheme maps all of the echoes that it receives along its line of sight. Therefore, for example, if the side lobes in the transducer beam profile are large, then a strong scatter that is sufficiently off-axis is registered and imaged at an incorrect location laterally. For example, take the barely discernible circular lesion in the image in Fig. 10. The internal speckle echoes inside the disk image are hypoechoic, that is, they are slightly weaker than the surrounding background. If several very strong scatterers were to be located near this "lesion" and the PSF had 20-dB side lobes, then it is entirely possible that the signal from side lobes can reduce the contrast within the disk because we know from the discussion in the scattering section that there is usually a dynamic range of 20–30 dB between the

weakest and strongest echoes. Therefore, poor side-lobe structure can decrease the low-contrast detectability of the lesion. On the other hand, if the side lobes are kept below 50–60 dB, a marked improvement in lesion detectability is possible (28). This problem was acutely realized when it first became feasible to use an array of transducer elements to synthesize the beam. The side lobes became a major problem in the initial design stage. Although this is not a universally accepted definition, but one that is adapted by some manufacturers, the lateral distance where the side lobes first reach 50 dB may be used to define contrast resolution. Consideration of this image quality metric has become more and more widespread in recent years. The recent implementation of harmonic imaging technology in scanners (to be considered later) derives its benefits partly from superior performance in contrast resolution.

Spatial Uniformity

The B-scan image in Fig. 14 is from a scanner that has a linear phased array transducer. Note that the PSF shape and size vary with depth. They become larger as depth increases; the lateral resolution especially is worse with depth. As we shall see later, fixed focusing of the transducer can improve the lateral resolution in the focal zone but at a further distance, it becomes worse due to beam divergence. Therefore, nonuniformity implies that optimal resolution at all depths in such systems is not possible and there is a spatially variant PSF. This is referred to as spatial nonuniformity. It is a direct consequence of the beam shape. When linear array technology is combined with computerized beam forming, it is now possible to focus the beam dynamically at various depths. The PSF can be made more uniform spatially, thus improving on this image quality metric.

Temporal Resolution

Temporal resolution in two-dimensional B-scan imaging, is defined and is limited by the image frame rate. If the imaging objective is to visualize the rapidly moving part of the body such as a beating heart, temporal resolution becomes important. In such cases, a loss of temporal resolution could contribute to a loss of spatial resolution due to the blurring effects of motion. As we discuss later, a frame rate of 15–30/s is possible in medical ultrasound by using real-time scanners.

To summarize, the objective image quality metrics that depend on engineering system design have been defined here. The transmitted pulse length or, equivalently, the pulse bandwidth determines axial resolution. Lateral resolution, slice thickness, contrast resolution, and spatial uniformity are functions mainly of transducer size, shape, center frequency, and the role of the associated diffractive phenomenon as the beam propagates. The basic theory that governs the last aspect will be considered in the next section. The reader should bear in mind that, ultimately, a task-dependent evaluation, for example, detectability of a lesion of some size and contrast, may depend on many other factors. These factors may include tissue properties, the adjustable settings on the scanner used by the sonographer, and last but not the least, the human visual system (28). Nevertheless if these factors are held fixed, the objective metrics do provide a meaningful input to the subjective evaluation process. Most importantly, these metrics serve a purpose in system intercomparison and in evaluating the efficacy of technological advancements.

DIFFRACTION AND ITS ROLE IN ULTRASOUND IMAGING

The transducer field pattern plays an important role in determining lateral resolution and the PSF. The variation of the beam pattern with depth is governed by diffraction. Therefore, B-scan imaging is considered a diffraction-limited imaging modality, similar to a lens-based optical imaging system (24). For several reasons, we consider in detail here the diffraction from a circular disk transducer in a lossless medium. For an exact description, the diffractive integral [to be defined in Eq. (9)] must be solved, either numerically or analytically in closed form. Analytical closed form solutions, with or without approximations, do exist for circular geometry for both monochromatic and pulsed excitation. Although circular transducer geometry may not often be applicable to many imaging systems, its consideration here helps us to understand the general features of radiative field patterns (1,13,31). For example, one of the main features is the somewhat arbitrary division of the radiative field pattern into two regions, the near field and the far field, when the excitation is monochromatic. An analytical solution exists for on-axis field variations and for the field distribution in the far field (a large distance from the transducer), but the near field remains amenable only to numerical solution. The Lommel function-based frequency domain approach presented here is an analytical closed form solution under a much stronger approximation, namely, the Fresnel approximation (32). It contains in itself all of the general features of the field pattern, but in addition also provides an estimate of the field pattern in the near-field region. The main reason for including it here is to demonstrate its use in describing the general features of the field pattern. Its potential usefulness does not seem to have been explored by the ultrasound community. The study of diffraction from transducers that have noncircular geometries, such as rectangular elements and linear phased arrays, is also briefly outlined at the end of the section.

Circular Piston Transducers

Diffraction is a natural phenomenon associated with any wave propagation, including ultrasound. It must be given due consideration if we want to know the distribution of acoustic field parameters (pressure, intensity, etc.) in front of the transducer as a function of both space (three dimensions) and time. The field of a circular disk transducer is circularly symmetrical in any plane perpendicular to the beam axis z (see Fig. 15). Therefore, the field can be specified as a function of z and off-axis distance $\rho = \sqrt{x^2 + y^2}$. Two cases are of practical interest: (1) determination of the impulse response, $h_1(\rho, z, t)$, when the time excitation of the piston or the PZT

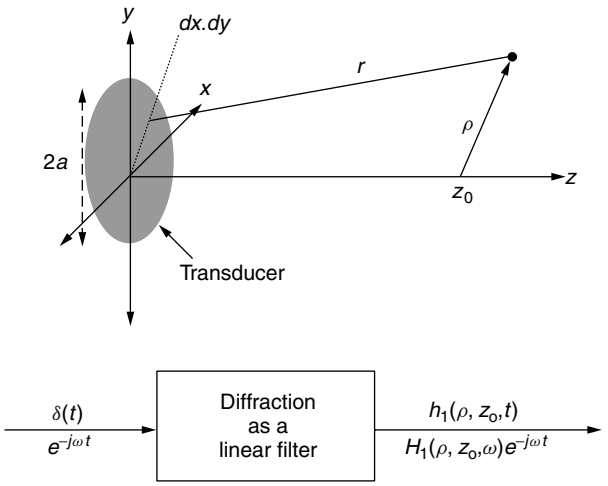


Figure 15. Geometry for diffraction from a circular transducer and the definition as a linear filter.

element is a delta function $\delta(t)$; (2) determination of the transfer function, $H_1(\rho, z, \omega)$, when the excitation is monochromatic, $e^{j\omega t}$. The exact solution in both cases is obtained by solving the appropriate diffractive integral:

$$h_1(\rho, z, t) = \frac{1}{2\pi} \int_A u(x, y) \frac{\delta\left(t - \frac{r}{c}\right)}{r}\, dx\, dy,$$

$$H_1(\rho, z, \omega) = \frac{1}{2\pi} \int_A u(x, y) \frac{e^{-jkr}}{r}\, dx\, dy, \tag{9}$$

where $k = 2\pi/\lambda = \omega/c$, $r = |\vec{r}|$, and λ is the wavelength. $u(x, y)$ defines the spatial excitation profile on the transducer face. A rigid baffle boundary condition has been assumed (20,31). In physical terms, the integral is a statement of Huygens's principle where the transducer is considered a distribution of elemental point sources. The field at any point in front of the transducer is the sum of elementary contributions from all elemental point sources of appropriate strength and phase.h_1 and H_1 are generally referred to as the velocity potential impulse response and velocity potential transfer function, respectively. In both cases, the integration must be performed over the transducer face of diameter $2a$. If spatially uniform excitation is assumed, $u(x, y) = u_0$.

An exact solution for the impulse response $h_1(\rho, z, t)$ in closed form has been derived and has been used by many researchers (33). Note that this impulse response is different from the transducer impulse $p(t)$ defined earlier for the electrical voltage drive signal input to the PZT element. From the standpoint of linear system theory, these two can be considered two separate linear subsystems in cascade (24). Their corresponding impulse responses can be convolved to predict the total system output when the PZT element excitation is an electrical delta function. For example, the shape of the short pulse disturbance received at some point (z, ρ) can be predicted by convolving $p(t)$ with $h_1(\rho, z, t)$ (30). The basic characteristics of diffraction are generally examined and studied for the monochromatic case, primarily at the center frequency of the transducer. Although the impulse response solution is exact and can be used to determine the field pattern for monochromatic excitation, computationally, it is not convenient for this purpose.

To the best of our knowledge, an exact analytical solution of the diffractive integral for monochromatic excitation at radial frequency $\omega = 2\pi f$ does not exist. Several authors have attempted an approximate solution in terms of Bessel functions (32,34). We have recently investigated the nature of this approximate solution, referred to here as $H_1^a \approx H_1$. For fixed ρ and z, the effect of diffraction on the input excitation signal can be considered a linear low-pass filter (see Fig. 15). As such, its impulse response and transfer function defined in Eq. (9) must be exact Fourier transform pairs, where t and ω are conjugate variables (24,29). Therefore $H_1^a(\rho, z, \omega)$ and $h_1(\rho, z, t)$ should constitute an approximate Fourier transform pair. This equivalence has been theoretically justified and computationally demonstrated (35). Only a short summary of results is given here, the reader is advised to consult the relevant publications for more details (35,36).

In (35), it was shown that if one makes a Fresnel approximation ($z^2 \ll a^2$) of the phase term e^{-jkr} in the diffractive integral of Eq. (9), the analytical solution can be expressed in closed form as

$$H_1^a(\rho, z, \omega) = \frac{1}{k} \exp\left[-j\left(kz + \frac{v^2}{2u} + \frac{u}{2}\right)\right]\left[U_1(u, v) + jU_2(u, v)\right], \tag{10}$$

where $u = ka^2/z$, $v = ka\rho/z$, and a stands for the transducer radius. U_n, known as Lommel functions, can be calculated as

$$U_n(u, v) = \sum_{s=0}^{\infty} (-1)^s \left(\frac{u}{v}\right)^{n+2s} J_{n+2s}(v) \tag{11}$$

where J_n stands for an order n Bessel function of the first kind. It is a convergent series, but for cautionary computational issues, the reader is referred to the published research monograph (36).

On-Axis Behavior: Near-Field and Far-Field

The diffractive integral defined by Eq. (9) can be solved exactly for on-axis conditions ($\rho = 0$). Equation (10) can also be used, but the solution will correspond to the Fresnel approximation ($z^2 \gg a^2$) (4,31). The expression for the on-axis beam intensity is

$$|H_1^a(\rho = 0, z, \omega = \omega_0)|^2 = (\text{const}) \sin^2\left[k(\sqrt{z^2 - a^2} - z)\right] \approx (\text{const}) \sin^2\left[\frac{ka^2}{4z}\right] \tag{12}$$

The first term is exact, the second term that has the approximate sign in front is the Fresnel approximation. This equation is often used to define the near- and far-field regions of a flat disk transducer. An example is

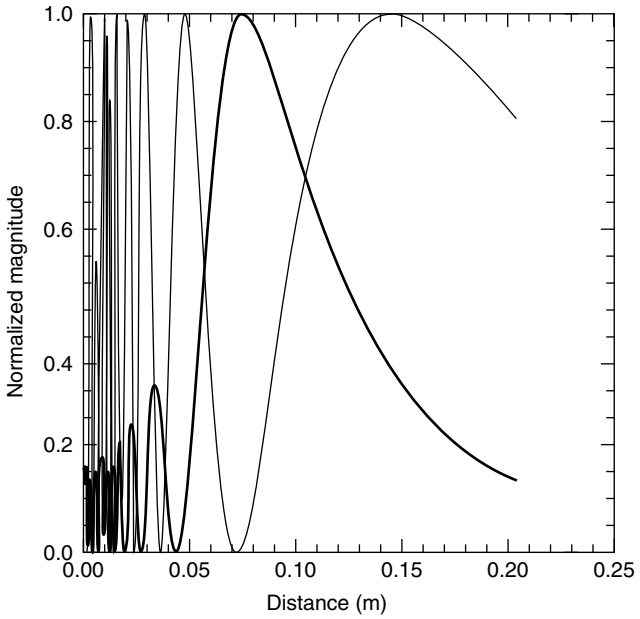


Figure 16. On-axis monochromatic signal amplitude variation due to diffraction from a circular unfocused (thin line) and focused (thick line) transducer.

given in Fig. 16, where the thin solid line represents the on-axis normalized intensity variation calculated using the Fresnel approximation. The diameter $2a = 19$ mm, the frequency $f_0 = \omega_0/2\pi = 2.46$ MHz, and the speed $c = 1,540$ m/s. The intensity value oscillates through a number of axial maxima and minima as we move away from the transducer. The last maximum occurs at an axial distance $(a^2/\lambda_0) = 14.5$ cm in this case. The region in front of the transducer up to a distance equal to (a^2/λ_0) is called the near field ($z < a^2/\lambda_0$). The region beyond this point is known as the far field ($z > a^2/\lambda_0$) (1,13). The Fresnel approximation is considered valid for any z larger than a few wavelengths. Therefore, its region of validity is most of the space in front of the transducer except the region very close to the transducer face. In Fig. 16, the number of maxima and minima become infinite as we approach the transducer face. They remain finite for the exact solution (not shown). This difference occurs because the Fresnel approximation breaks down at small z. The Fraunhofer approximation is a much weaker approximation and is valid only for $z \gg a^2/\lambda_0$, that is, only in the far field (4,31).

Beam Patterns and Beam Profiles

Beam patterns are the two-dimensional intensity or amplitude distribution in any given x, y plane defined by a fixed z_0. Due to the circular symmetry, beam profiles, defined as the beam amplitude or intensity variations as a function of the off-axis radial distance ρ, may also be used. No exact analytical solution exists for the entire space in front of the transducer. The analytical closed form approximate solution $H_1^a(\rho, z, \omega_0)$ can be used but it is based on the caveat that its validity may be questioned for planes very close to the transducer face. The top row in Fig. 17 shows the beam patterns for the same flat disk (unfocused) transducer mentioned earlier. Patterns for four different z_0 values are shown. The z_0 locations have been indicated on each pattern. Note that only the pattern at $z_0 = 18.3$ cm is in the far field of the given transducer. Here, along with the central main lobe of the beam, we can also see the surrounding weaker side lobes. In the Fraunhofer approximation, Eq. (10) reduces to a well-known formula for the magnitude of a radial beam

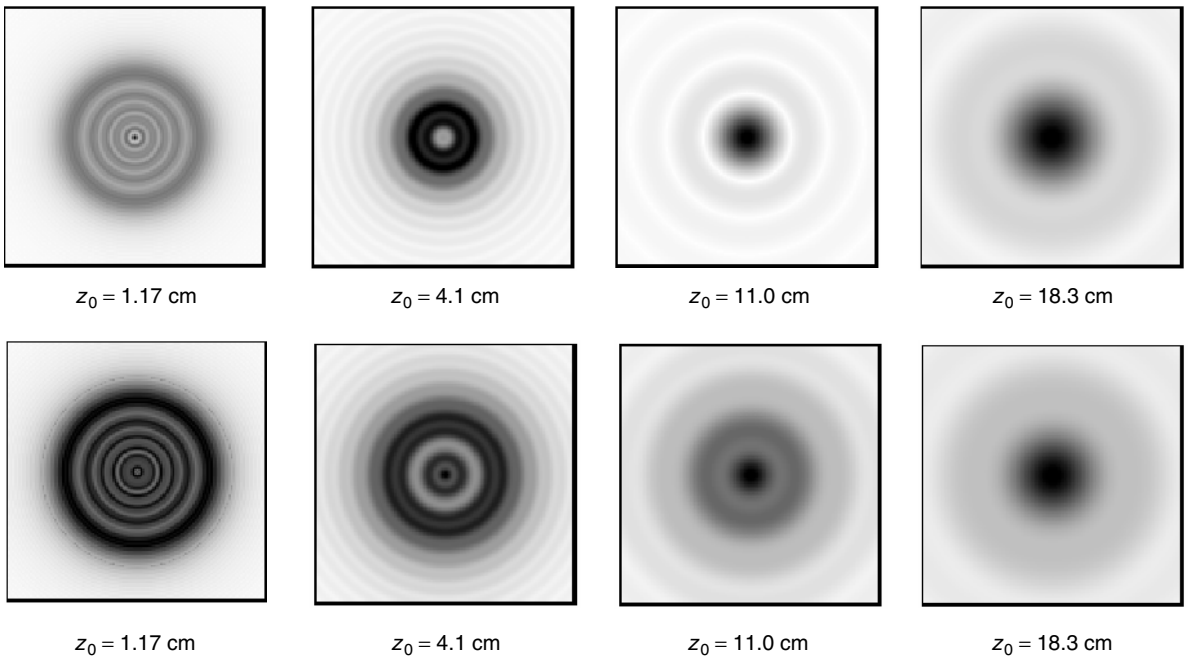


Figure 17. Beam patterns at different z_0. Top row is unfocused; bottom row is focused.

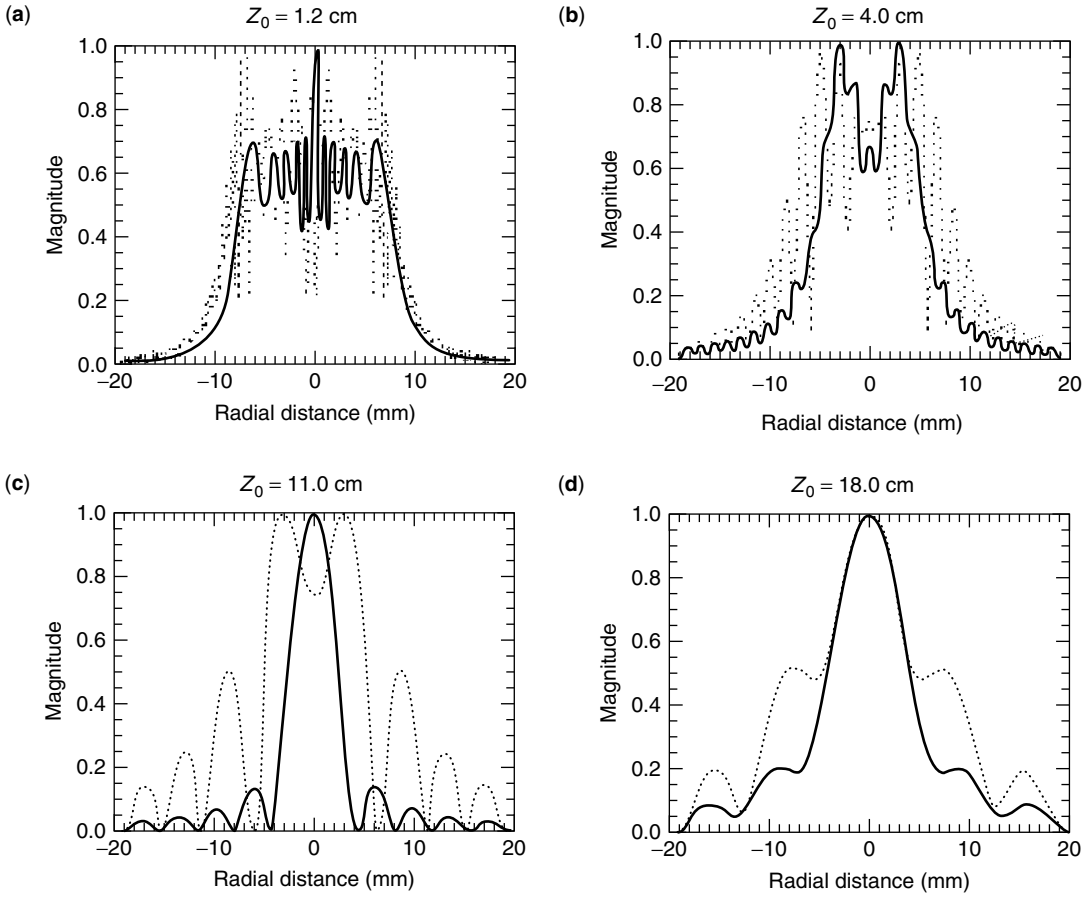


Figure 18. Beam profiles at different z_0. Dashed is unfocused, and solid is focused.

profile:

$$H_1^a(\rho, z = z_0, \omega_0) = \text{const}\left[\frac{J_1\left(\frac{ka\rho}{z_0}\right)}{\left(\frac{ka\rho}{z_0}\right)}\right]. \qquad (13)$$

This simple expression is often used to estimate the beam profile in the far field ($z \gg a^2/\lambda_0$). The four plots shown with dashed lines in Fig. 18 are the radial beam profiles of the corresponding two-dimensional beam patterns shown in Fig. 17, top row (unfocused case). Using these theoretical plots generated within the Fresnel approximation, we can summarize some general features of the beam pattern. In the near field, the beam is essentially confined within the original diameter of the transducer. The intensity varies significantly within the beam and with z. In the far field, the beam assumes a standard form. It has a maximum at the center and the intensity falls off as the radial distance increases. The width of the central lobe becomes larger as z increases, and the on-axis intensity decreases as the beam begins to diverge. The lateral resolution is generally considered to be determined by the beam confinement distance in the near field and by the -6 dB full width of the central lobe, defined by the $[J_1(\theta)/\theta]$ formula in the far field. An important aspect of lateral resolution, mainly due to diffraction, is that it is spatially nonuniform. Note also the side lobe structure that exists on either side of the central main lobe.

Focusing

The primary function of focusing is to improve the lateral resolution by reducing the central main lobe width in the focal zone. One way to achieve focusing is to curve the transducer spherically by using a radius of curvature F, as shown Fig. 19. In Fig. 16, the thick line shows the on-axis intensity for the same transducer, but the radius of curvature $F = 11$ cm. An expression similar to Eq. (12) has been used where z is replaced by ε where $\varepsilon^{-1} = z^{-1} + F^{-1}$. Though $z = 11$ cm defines the geometrical focus, a true focal point may be defined as the z location where the on-axis intensity has a maximum. Note that the true focal point is always at $z < a^2/\lambda_0$. In other words, we cannot obtain a focus in the far field of a transducer. A difference often exists between the geometrical focus and the true focus (4). Kossoff defined three different degrees of focusing: strong or short focusing implies $F \approx 0.2a^2/\lambda_0$, medium focusing implies $F \approx 0.5a^2/\lambda_0$, and weak or long focusing implies $F \approx 0.8a^2/\lambda_0$ (37).

No exact closed form solution exists for unfocused off-axis behavior. A solution in the form of an integral expression can be found in (4). However, the beam pattern only in the plane at the geometrical focal point $z = F$

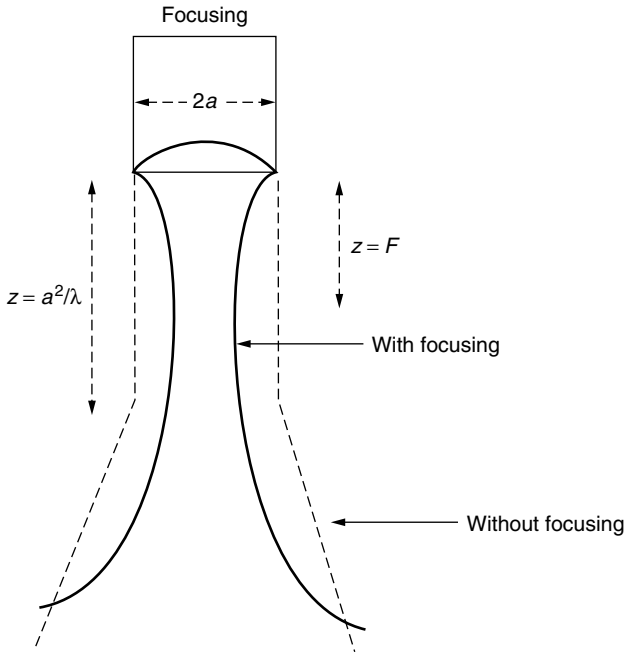


Figure 19. Effect of focusing on the beam profile.

is given analytically by Eq. (13) where z is replaced by F. Within the Fresnel approximation, a closed form expression based on the Lommel diffraction formulation for a focused piston transducer of radius a is given by (36)

$$H_1^a(\rho, z, \omega) = \frac{\varepsilon}{kz} e^{j(kz+(k\rho^2/2z)+(ka^2/2\varepsilon))} \left[U_1\left(\frac{ka^2}{\varepsilon}, \frac{ka\rho}{z}\right) + jU_2\left(\frac{ka^2}{\varepsilon}, \frac{ka\rho}{z}\right)\right]. \quad (14)$$

The circularly symmetrical beam patterns, calculated by using Lommel functions for the focused case at the same four z locations, are shown in the bottom row of Fig. 17, and the top row is the corresponding unfocused case. The solid lines in Fig. 18 represent the beam profiles for the focused case, and the dashed lines represent the unfocused case. Note a marked decrease of -6 dB width of the central lobe and the reduction in the side-lobe level at the geometrical focal point. Figure 19 illustrates the changes in the beam profile due to focusing. The beam width is reduced significantly near the focal point and also within a distance range called the focal zone. However, the beam diverges more rapidly beyond the focal zone. The on-axis intensity increases many fold at the focal point. This type of focusing can improve lateral resolution and reduce side lobes but only within the focal zone. This zone is at a fixed distance from the transducer. As we will see later, introducing phased array transducers has made it possible to achieve variable focusing at many different depths.

Noncircular Transducer Geometries

For an exact solution to diffraction from any arbitrary geometry, the integral defined in Eq. (9) must be solved numerically. However, an approximate understanding of the beam pattern is possible. Two methods are outlined here. The reader is urged to consult the references for more detail (38). The first method uses the angular spectrum technique when the excitation is monochromatic (39). If $w(x, y, z = z_1)$ is the complex (magnitude and phase) of the two-dimensional monochromatic wave field at frequency ω_0 defined in one plane at $z = z_1$, then it is possible to propagate the field to a different plane at $z = z_2$. The first step of the process is to take a two-dimensional Fourier transform of the wave field that exists at z_1. Let us call this Fourier transform $W(f_x, f_y, z = z_1)$. Here f_x and f_y are the spatial frequency variables conjugate to x and y, respectively. Then, we multiply the Fourier transform by a phase filter $e^{j2\pi(z_2-z_1)\sqrt{\lambda_0^2 - f_x^2 - f_y^2}}$. The final step is to take an inverse two-dimensional Fourier transform of the product to obtain the field $w(x, y, z = z_2)$ for the plane at $z = z_2$. The transducer geometry can be specified as the initial wave field at $z = 0$. The solution is valid in the near and the far fields of the transducer. The method can be extended to finite bandwidth (short duration) excitation signals used in medical ultrasound (40).

The second method can provide only an estimate of the beam pattern in the far field of the transducer, that is, in a plane at any $z \gg D^2/\lambda_0$, where D stands for the maximum length in the full specification of the transducer geometry. It has been shown that when the Fraunhofer approximation is applied to the exact diffractive integral for any specified transducer geometry $f(x, y)$, the resulting field pattern in the far field is simply a two-dimensional Fourier transform of the transducer geometry, followed by a scaling of the spatial frequency variables f_x and f_y by the term $\lambda_0 z$ (29,39). Another alternative is to perform a three-dimensional finite-element modeling of the wave propagation. These methods can be useful in designing and understanding the transducer array technology that has been driving the market in recent years.

TECHNOLOGICAL ADVANCEMENTS

Real-Time B-Scan Imaging

The main objective here has been to improve temporal resolution. The mechanical movement of the transducer to vary the line of sight is a slow process. Motion artifacts can be avoided in the image if the object does not move. But in medical applications, the breathing motion time period is of the order of 1 second. If a single frame of a B-scan image is considered to be composed of N number of A-lines, each requiring $(\text{PRF})^{-1}$ seconds to acquire, then the frame rate is determined by PRF/N. Given the maximum depth $z_{\max}$ to be imaged, the PRF is set at $c/2z_{\max}$. A compromise is needed between $z_{\max}$ and N. Typically for $N = 128$ and a frame rate of 30 frames per second, a maximum imaging depth of 19.5 cm is possible. Many different schemes for fast scanning have been developed. The details can be found elsewhere (41). Figure 20 illustrates a sector scanner where a single element transducer wobbles inside a protective sheath, thereby sweeping the beam (line of sight) in an arc. The resulting B-scan image has the familiar shape of a sector of a circle.

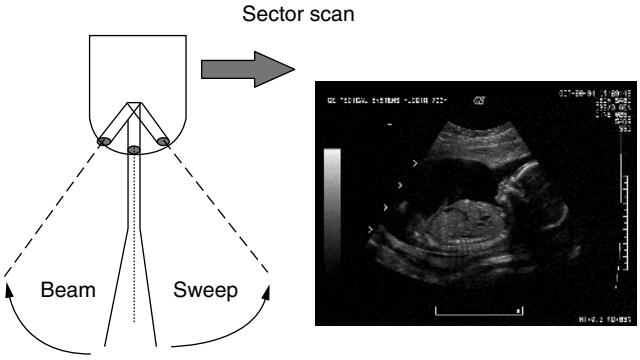


Figure 20. A sector scanning method for real-time imaging.

Focusing a single-element transducer has been used for many years to improve lateral resolution and increase the sonicated intensity in the focal zone (1,41). Equation (14) provides the beam profile for a circular disk transducer. But there is a price to pay in fixed focus operation. A loss of lateral resolution always accompanies at distances beyond the focal region due to more rapid divergence of the beam (see Fig. 19). Therefore, there is a trade-off with spatial uniformity.

Transducer Arrays

These are novel assemblies that have more than one transducer element. A linear array generally has rectangular elements all arranged in a line, as illustrated in Fig. 21, or sometimes across a curved line (a convex array). Annular arrays (concentric rings of elements) have also been developed (1,28,29). The number of elements has steadily increased since the late 1970s when they were first introduced, and it is common now to have 128 or more elements. The technical challenges that had to be surmounted include the development of dicing and design methodology, the development of multichannel signal processing that has minimum cross talk, and the reduction of grating side lobes that originate in the beam due to the inherent periodicity introduced by the array (29). It is necessary to understand the diffractive pattern produced by a given array setup to estimate and improve the image quality metrics defined earlier. Although this is a computationally difficult task, as outlined in the last section, some simple guidelines can be used to define the pattern in the far field. It is convenient to think of each element as a point source. In the operation of a phased array scanner, the effect of a large continuous-face transducer is synthesized from these periodic point sources. The aperture size (D) involved in forming the beam and the wavelength (λ_0) at the dominant frequency determine the central beam width and hence the lateral resolution (Fig. 21). Due to the periodicity of the elements, replicas of the central lobe are created in the beam. These replicas of the central lobe are called grating side lobes. Their amplitudes are not negligible and often are comparable to central beam amplitudes. As stated before, such large side lobes are unacceptable for medical imaging because they compromise contrast resolution. The element spacing d and λ_0 determine the locations of the grating side lobes. The width of the elements, w, determines the amplitudes of the grating side lobe relative to the main lobe. Many different approaches to reducing side lobes and improving the main lobe lateral resolution are possible, but none is perfect.

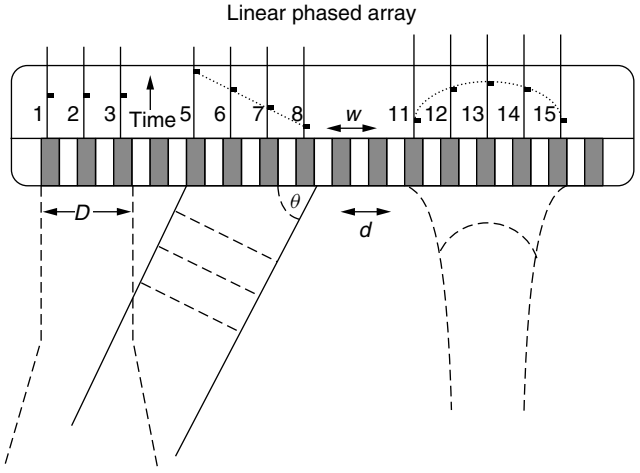


Figure 21. Linear phased array: example of electronically timed excitation for scanning (left), beam steering (middle), and variable degree of focusing (right).

Despite the side lobe problems, the advantages are many, and they have been slowly realized in practice during the last decade. First, mechanical scanning is replaced by electronically controlled sequential excitation of a group of elements in a linear array. The beam that is generated is illustrated in Fig. 21; it has excitation of element numbers 1, 2, and 3 followed by 2, 3, and 4, etc. Alternatively, the beam can be steered to any angle θ by applying an appropriate linear time delayed excitation to a group of elements (see, for example, elements 5, 6, 7, and 8 in Fig. 21). Second, a variable degree of focusing can be achieved electronically by imposing a curvature on the excitation time delay that is applied to a group of selected elements (see, for example, elements 11, 12, 13, 14, and 15 in Fig. 21). Third, there is the flexibility of reducing the side-lobe structure by what is known as beam apodization. This refers to nonuniform shaping of the amplitude of the excitation signal across the face of the transducer array or a subset of elements that form the beam. A predetermined weighting function controls the amplitude of the excitation signal in each element. The side lobes can be reduced significantly at the expense of a slight reduction in lateral resolution determined by the main lobe width (27). The fourth advantage stems from the potential power of digital signal processing that can be applied to the multichannel echo signals. One example is considered in the next section. Perhaps only a fraction of that potential has been realized so far.

Digital Processing of Multichannel Signals

This is integrating the full power of computer technology with the multichannel signal generation/reception capability afforded by the transducer arrays (27). It has been responsible for steady improvement in image quality

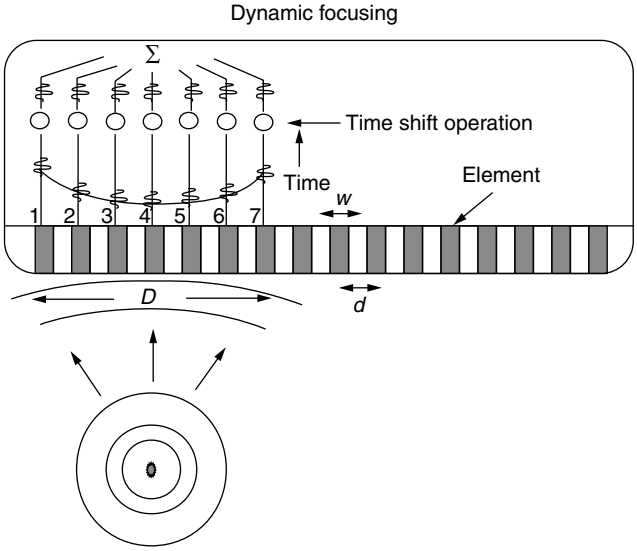


Figure 22. Signal processing of different channels for dynamic focusing of received echo signal.

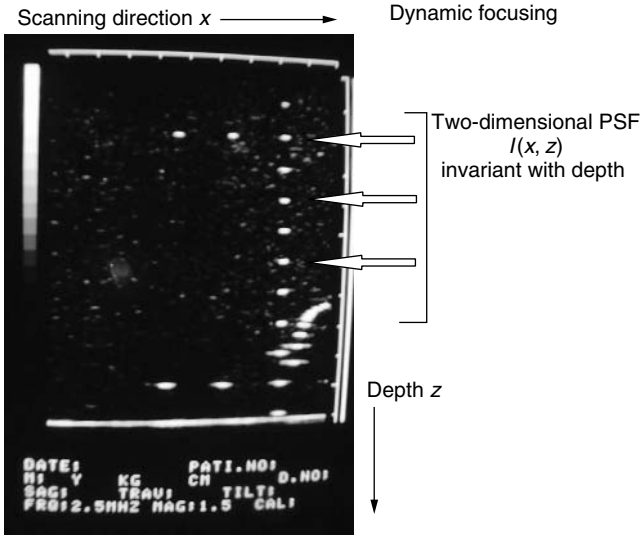


Figure 23. B-scan image of a wire target phantom where dynamic focusing is turned on.

offered by clinical scanners in the last decade. Each manufacturer follows its own proprietary technology. Hybrid analog/digital computers, specially designed to accommodate the data processing requirements of real-time image formation, are used in a dedicated and adaptive mode. Although the details remain proprietary, we can get a general idea of what is being done here. The PSF for B-scan imaging depends on the depth and the beam steering angle (28). The challenge has been to exploit the flexibility and speed of computer hardware and software to improve spatial resolution, the spatial uniformity of the PSF, and the contrast resolution (reduced side lobes), simultaneously. The lateral resolution can be improved and side-lobe reduction can be achieved at each depth by *focusing* the multichannel echo signal in the software upon reception. Consider an on-axis point scatterer at depth z that sends a short pulse back toward a linear transducer array (Fig. 22). The signal received by each element will have a travel time delay that can be easily calculated by assuming a constant speed of propagation. By applying a time shift in the software before summing up the signal in all of the channels, dynamic focusing for depth z can be achieved. Because dynamic focusing is performed in the software and on the echo signal, there is the flexibility to tailor the time delays for each depth. This processing is usually done as fast as the echoes are received, so that real-time imaging capability can be maintained. Furthermore, the receiver aperture size can also be changed for different depths (arrival times) by adding more channels to the signal summation algorithm. The apodization in transmit and/or in receive can be further adapted to different depths and imaging objectives. Significant improvement in spatial and contrast resolution can now be maintained across a large portion of the image. In other words, a high degree of spatial uniformity is possible now simultaneously along with improved spatial and contrast resolution (27). This is demonstrated in Fig. 23 by the image of wire targets at different depths in a tissue-mimicking phantom taken by a 64-element phased array scanner when the dynamic focusing was turned on. Compare the PSF with those shown in Fig. 14 where the dynamic focusing is turned off. We can visualize the improvement in the lateral resolution at each depth in the two images. At the same time, now the PSF is spatially uniform.

Doppler Ultrasound

Three types of instruments are currently used for blood flow measurements in heart, arteries and veins based on the Doppler principle: continuous-wave (CW), pulsed-wave (PW), and color-flow (CF) (42). CW is the simplest and least expensive. Two tilted transducers that have nonparallel beams produce an overlap region at some fixed range distance. One acts as a transmitter, and the other as a receiver. The echo signal from blood flow (moving blood cells) is continuously received, and the Doppler shift is estimated by a two-stage process. First, the echo signal is mixed or multiplied by the transmitted frequency signal. Sum and difference frequencies are produced at the output of the mixer. A subsequent low-pass filter removes the sum frequency but preserves the difference frequency. According to Eq. (6), the Doppler shift is proportional to the flow velocity and can be calibrated if the Doppler angle is known. An audio amplifier and speaker are often used to hear the blood flow. If the moving blood cells have a distribution of velocities, a spectrum of Doppler frequencies is displayed as a function of time. CW suffers from a lack of depth range and region of interest (ROI) size selectivity. PW combines the velocity determination of a CW system and the range discrimination of pulse-echo imaging systems. Figure 24 illustrates the processing of a signal for a pulsed Doppler system. Similar to B-scan imaging, only one transducer is used in the pulse-echo mode. The spatial pulse length for Doppler probing is usually longer than that used for imaging. A rate of 5–25 cycles/pulse is not uncommon for achieving the narrowband characteristics necessary for Doppler studies. The pulse repetition frequency (PRF) is under operator

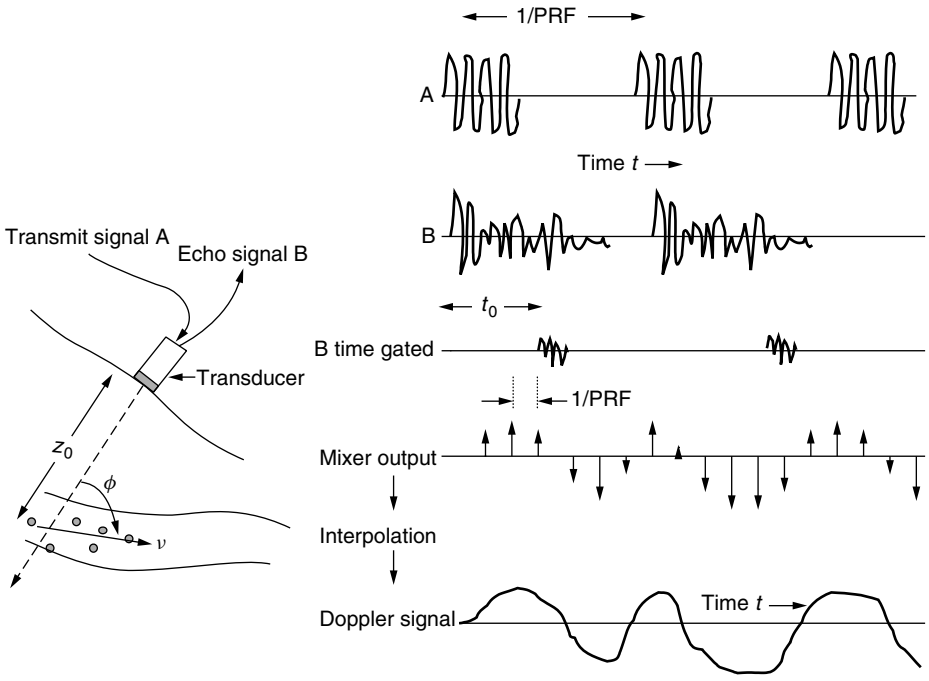


Figure 24. Processing of a pulsed wave Doppler signal. (From Bamber et al. (13)).

control (see transmit signal A in Fig. 24). The region of interest at some depth z_0 is selected by applying a time gate to the echo signal B centered at $t_0 = 2z_0/c$. Unlike CW, the output of the combined mixer and low-pass filter in PW becomes a sampled version of the Doppler-shifted frequency waveform. The sampling interval is $(\text{PRF})^{-1}$. After some sort of interpolation, this signal is sent to the audio amplifiers. The sampling aspect of PW results in a Nyquist frequency $= (1/2)\text{PRF}$. Consequently, any Doppler-shift frequency above the Nyquist limit is aliased. If the PRF is increased to reduce aliasing, the unambiguous range depth of the ROI is compromised. On the other hand, the lowest Doppler frequency requires data acquisition time equal to at least one complete cycle of the Doppler signal. This, in turn, increases the number of pulses that need to be fired before a single image frame that has velocity information can be displayed. In other words, the frame rate or temporal resolution for a fixed PRF has to be compromised to accommodate smaller Doppler shifts. The system design and optimization problem is full of challenges, notwithstanding the fact that the echo signal from blood is usually very weak and consequently the SNR is generally poor. Quadrature demodulation is often employed at the mixer stage to separate the positive and negative Doppler frequencies. A fast Fourier transform is often used to determine the amplitudes of the Doppler-shifted frequencies in a spectrum. The resolution is determined by the Doppler sample volume (the region from which Doppler-shifted echoes return and are presented audibly or visually). This is determined by the transducer beam width in the lateral and receiver gate length plus half the pulse length in the axial direction. Multiple gates can also be selected in some instruments. Firing of short pulses (for imaging) and long pulses (for Doppler) can be interleaved so that B-scan image generation, image guided gate selection, and Doppler spectrum information can be presented simultaneously. For obvious reasons, such instruments are called duplex scanners.

Color Doppler imaging is a more recent development and has found widespread clinically utility in the last few years. It presents two-dimensional, cross-sectional, real-time blood velocity or tissue motion information in color, along with the two-dimensional, cross-sectional, gray-scale anatomic B-scan image. The near real-time (high temporal resolution) requirement makes the design problem even more challenging. The signal processing chain does not use the full Doppler shift spectrum information because of a lack of time and/or lack of parallel channels necessary for real-time imaging. Instead, the A-line echo signal is split into 100–400 time gates. Consider that the scattering medium seen by a given A-line and a given time gate is interrogated by two pulses separated by a time interval equal to $\tau = (\text{PRF})^{-1}$. If the scatterer has remained stationary in that time τ, the normalized correlation factor between the first demodulated echo wave train and the second one shifted by a time delay τ will be near or equal to one. Any possible motion will change the magnitude and phase of this correlation factor. In 1985, Kasai et al. established the relationship between the magnitude and phase of this complex correlation factor and the mean and the variance of the Doppler-shifted spectrum (43). As many as eight pulses are used to determine this correlation factor accurately at each time gate center. The summation duration for correlation determination becomes 8τ. The beam must be held stationary for this duration. Therefore, the image frame rate has to be sacrificed. A mean Doppler shift can be estimated at each time gate location. If it falls below a threshold for motion detection, the information in that particular time gate is considered to be coming from a stationary tissue, and the echo signal is processed like any normal B scan and is displayed in gray scale.

If any motion is inferred at any gate location, then, the motion data are color mapped by using a color scale and that pixel location is displayed in color. A blue color is generally used to depict flow away from, and red is used to depict flow toward the transducer. There are several problems in color flow mapping systems. Noise and signal from slowly moving body parts can overwhelm the small signal from moving blood. An analog "wall filter" is often employed that passes Doppler frequencies above a set level and eliminates strong but low-frequency Doppler shifts from moving body parts such as heart or vessel walls. The spatial resolution of color Doppler is much worse than the gray-scale B-scan image, and full Doppler spectral information is not depicted. Yet the ability to "see" the flow information simultaneously in color in real time along with the anatomy depicted by the B-scan image has found many applications in medical diagnosis.

A more recent development known as amplitude or power Doppler has a signal-to-noise ratio four to five times better and therefore has higher sensitivity for visualizing blood flow in small vessels. Here, instead of using the mean of the Doppler shift and its sign, a number that is proportional to the total area under the Doppler power spectrum (positive and negative shifts included) is used and depicted by a color scale. This information is considered proportional to the density of blood cells in the sampled volume and hence to the blood flow volume. Now, blood flow patterns in small vascular structures can be visualized by power Doppler. There are numerous opportunities for further development of color-flow imaging, particularly in conjunction with the use of contrast agents. Compromises between frame rate, Doppler information content, spatial resolution, and aliasing artifacts still remain major issues and must be carefully taken into consideration while operating these instruments. Time-domain cross-correlation processing, which is a broadband technique used to extract velocity information, is also, it has been shown, of practical value. The reader may refer to a review article by Wells for further details on Doppler and color flow imaging (44).

Harmonic Imaging

The concepts of harmonic imaging are based on an understanding of finite amplitude distortion of a waveform as the wave propagates in a medium that has a significant nonlinear response. The theory has a rich history of application in underwater and airborne sound. Only recently, the biomedical community has paid clinically significant attention to this phenomenon (45). Simply put, because the propagative velocity of points on a sound wave increases with pressure, the pressure maximum of a finite-amplitude wave travels slightly faster than the pressure minimum. This leads to a continuous distortion of the waveform, and harmonics of the source frequency are generated in the medium *during* wave propagation. The generation of harmonics becomes more pronounced as peak pressure amplitudes of the interrogating pulse used for imaging, frequency, distance from the transducer, and the magnitude of the nonlinearity parameter B/A of the medium increase [see Eq. (5) and Ref. 40]. Because of low attenuation, significant second- and third-harmonic amplitude generation is possible in water, even for interrogating pulse pressures as small as 0.5 MPa (46). However, there is ample evidence that diagnostic scanners operating at peak pressure fields near or in excess of 1 MPa can generate harmonics in human body tissue imaging (47). The harmonics can be 10–30 dB lower in amplitude than the fundamental. Latest scanners have incorporated a band-pass filter in the processing chain that removes the first harmonic (or fundamental) but keeps the second harmonic in the Rf echo signal for B-scan image formation. It is available as an option and is generally referred to as harmonic imaging, as opposed to conventional imaging at the fundamental when the filter is turned off (45). The transducers operating in pulse-echo mode need sufficient bandwidth so that a moderate first-harmonic signal can be transmitted, but a weak second-harmonic can still be received within the available bandwidth. The frequency overlap between the two bands has to be minimized, and this limits the bandwidth of the transmitted signal. Therefore, there is an inherent compromise between axial resolution and the efficacy of harmonic imaging.

Significant improvement in low-contrast lesion detection results when one switches from the fundamental to the harmonic imaging mode. The improvement in image quality has been particularly noted in obese patients. This improvement in contrast resolution is due to two factors. First, the harmonic beams generated during pulse propagation possess a narrower central lobe and also have lower side-lobe levels than the fundamental beam simply because of the higher frequency content. Therefore the point-spread function, specially the lateral resolution, for harmonic imaging turns out to be superior to fundamental imaging. Second, the harmonic component generated has an approximate quadratic relationship to the wave amplitude at the fundamental. Thus most of the harmonic energy originates from the strongest part of the beam in the center. The weaker parts of the beam, (i.e., the side lobes) participate to a lesser extent in the harmonic generation process. Third, the harmonics are absent in the near field of the transducer because the propagation has to take place over a finite distance before a sizable harmonic component can be generated. Consequently, the degradation of the point-spread function normally seen in fundamental imaging due to phase aberrations introduced by propagation through thick body fat layers under the skin is less severe in harmonic imaging. Some of these inferences have been theoretically examined (40,48) and experimentally confirmed (46).

Contrast Imaging

The development of contrast agents has played a significant role in enhancing the potential of ultrasonic imaging in medicine in the last decade. The original concept was somehow to inject some sort of air bubbles/solid particles into the bloodstream. Due to the large acoustic impedance difference between air and blood/tissue, one would expect enhanced scattering that would result in increased contrast. The challenge was to produce small enough microbubbles that could pass through the lung capillary circulation and would be stable enough to reach

the left side of the heart after an intravenous injection. A summary of the use of ultrasound contrast agents (UCA) in medical imaging is given elsewhere (49). To this date, dozens of UCAs have been developed, and many have been approved for clinical use by the Food and Drug Administration (FDA) (22). Now, most UCAs can reach the left heart ventricle cavity; they are either air-filled or contain gases that dissolve poorly in the blood and have mean diameters smaller than 7 μ. The mechanisms that produce contrast enhancement and its potential use still continue under investigation. We present here a summary of the current state of its imaging applications.

To this date, all three regimes of bubble response discussed earlier have been exploited for contrast imaging. Initially it was thought that B-mode imaging at the fundamental accompanied by the injection of an UCA would provide sufficient contrast enhancement for many situations. This was the case in imaging the heart cavity filled with blood, but the improvement was only a few decibels in tissue perfusion. The reason is that the blood vessels in body tissue are small, and therefore the bubble density is low. Echoes from tissue can easily mask any increase from an UCA. This is often quantified as the agent-to-tissue ratio. New imaging techniques are continuously being developed to improve this ratio where UCA response to ultrasound is exploited. Transient enhanced scattering has also been explored at higher peak pressure amplitudes. Power Doppler imaging has proven more sensitive for detecting flow in small blood vessels. The introduction of an UCA enhances the signals received from blood further, and the detectability of flow from small vessels is increased further. However, color Doppler methods are susceptible to tissue motion that generates clutter or noise in Doppler information display. A method where harmonic imaging is combined with power Doppler can overcome this limitation and is currently considered one of the most sensitive techniques for tissue enhancement. At very high ultrasound peak pressure amplitudes, the transient destruction of the bubble and its signature has also been exploited for myocardial tissue enhancement (23). In another new technique called pulse-inversion imaging, a sequence of two short pluses is transmitted into the body (50). The second is an inverted replica of the first. In a linear medium, the two responses cancel each other when added. But in a nonlinear system such as bubbles, the sum is not zero and can be used as a signature of the UCA. The excitation in pulse-inversion imaging need not be narrowband. Therefore, it can overcome the small bandwidth limitation that is inherent in harmonic imaging and promises improved axial resolution. In addition to the exploitation of the harmonic response of the UCA in techniques such as harmonic B-mode, harmonic Doppler, and pulse inversion, new opportunities arise when we consider the subharmonic component of the bubble response (51). A potential advantage here is that the contribution from tissue is minimal. However, the application potential is still being explored.

3-D Imaging

Theoretically, it is possible to generate 3-D ultrasound images by a simple extension of 2-D imaging. Two possible avenues have been explored. (*1*) One can acquire multiple 2-D images of known spatial orientation and then display the 3-D volumetric data set (28). For example, different planes can be scanned by moving the entire one-dimensional transducer array along the y axis in the geometry depicted in Fig. 3. Alternatively, a sector scanner may be rotated through 180° in small angular steps to generate 2-D images along different radial planes. The reader is referred to a recent review article for more information on the computational aspects (52). Many ultrasound manufacturers have already introduced such modifications in their scanners. (*2*) To avoid mechanical scanning motion, two-dimensional transducer arrays have been explored (53). The practical problems center around the complexity of fabrication, connection, and integrated circuit design required for real-time imaging by using a matrix of elements as large as 64×64, not to mention the high cost involved.

Current ultrasound systems now allow near real-time image data acquisition, registration and 3-D reconstruction, display, and manipulation. This powerful new imaging capability has clinical utility in many areas such as accurate tumor volume estimation for radiation treatment planning, fetal face rendering, mapping of major vessels in 3-D, and in ultrasound guided surgery and biopsy. Medical simulators are interactive virtual reality environments used to improve human performance in procedures that require visual correlation of three- or higher dimensional information. When fully developed, such tools will serve as a possible alternative to training and performance evaluation for physicians in minimally invasive procedures. The concept is similar to flight simulator training in aviation. In the coming years, 3-D ultrasound is expected to play a significant role in medical imaging.

Elasticity Imaging

This is a new imaging modality that has potential use in tumor detection and differential diagnosis. Palpation, where a physician can feel the hardness of a tumor, has been common a diagnostic practice, but it is not quantitative. The basic physics of elasticity can be used here for developing a more quantitative imaging modality that provides a measure of the elastic modulus of the tissue (54). When an external stress is applied to an elastic material, the material can undergo deformation. Strain is a quantitative measure of that deformation. For a simple one-dimensional case, strain is given by the fractional change in length $\Delta L/L$. The stress to strain ratio is called Young's modulus of the material. The elasticity imaging method for biological tissue is based on external tissue compression and subsequent computation of a strain profile along the transducer axis. The deformation ΔL that is needed for strain calculation is generally extracted from cross-correlation analysis of pre- and postcompression A-line signal pairs. At the very basic level, the processing involves dividing the precompressed reference A line into many overlapping segments, each several millimeters long. The transducer then applies an axial static compression to the tissue by a small amount dz. Another A line is acquired at this compression and

divided into same number of segments. Separate cross-correlation is performed for each segment, resulting in a set of time shifts for each segment. These time shifts can be converted into an estimate of ΔL. A one-dimensional strain image can be generated by mapping the strain value in each segment into gray scale and displaying it as a function of depth in the axial direction. By moving the transducer laterally and repeating the compression step processing, it is possible to generate a two-dimensional strain image, similar to a B-scan image. Using appropriate correction and calibration, it is also possible to map a profile of Young's modulus in the tissue. The axial resolution depends on the size of the segment chosen for cross-correlation processing. The time shift estimates become less noisy and more robust when the segment window is large. So, there is a trade-off here between resolution and SNR. Generally the resolution is much worse compared to a B-scan. But its utility lies in the possible contrast. For example, it may not be possible to visualize a tumor on a B-scan if the backscatter from the interior of the tumor and the surrounding tissue happen to be equal. But the tumor may become visible on the elasticity image if there is a measurable difference in Young's modulus between the tumor and the surrounding tissue. The reader is referred to publications by several researchers who have done work in this area in the past decade (55–57).

Safety Issues

As in any diagnostic test, there may be some risk (probability of damage or injury) in using medical diagnostic ultrasound. The biological effects and safety of diagnostic ultrasound have received considerable attention during the past few years (58). After a considerable review, the American Institute of Ultrasound in Medicine (AIUM) and the World Health Organization (WHO) both issued statements attesting to the safety of diagnostic ultrasound. AIUM issued a modified statement in 1997: "No confirmed biological effects on patients or instrument operators caused by exposure at intensities typical of present diagnostic ultrasound instruments have ever been reported." They do advise a prudent approach to minimize the risk by minimizing the exposure time and output levels. This is known as the ALARA (as-low-as-reasonably achievable) principle.

The mechanisms of action by which ultrasound could produce biological effects have been studied extensively. They can be divided into two groups: heating and mechanical [2]. Ultrasound can produce a temperature rise as it propagates in tissue, primarily due to absorption. Heating can increase with ultrasound intensity and frequency. A quantitative measure of the *probable* thermal activity has now been developed to provide some assistance in risk/benefit decision making. It is called the thermal index (TI). It is equal to transducer acoustic output power divided by the estimated power required to raise the tissue temperature by 1°C. Temperature rises are considered significant if they exceed 2°C. For further details on calculating the thermal index, the reader is referred to a monograph published by the National Council on Radiation Protection and Measurement (NCRP) (59). The mechanical risk is related to cavitation. Cavitation is the production and dynamics of bubbles in a liquid medium. Two types of cavitation are recognized. Stable cavitation is used to describe bubbles that oscillate in diameter in response to the incoming pressure wave. Transient cavitation occurs when bubble oscillations are so large that the bubble collapses during a compression cycle. Localized extremely high temperatures and shock waves are created in the immediate vicinity. Transient cavitation has the potential for significant destructive effects. A theory has been developed based on experiments in water that yields a predictable dependence of cavitation threshold on pressure and frequency (58). A mechanical index (MI) has been formulated to assist users in evaluating the likelihood of cavitation-related adverse biological effects for diagnostically relevant exposures. MI is equal to the peak rarefactional pressure (in MPa) divided by the square root of the center frequency (in MHz) of the pulse bandwidth. Thus far, biologically significant, adverse, nonthermal effects and the corresponding threshold MI have been identified with certainty only for medically relevant exposures in tissue that have well-defined populations of stabilized gas bodies. Many manufacturers now provide a display of the TI and MI on the scanner. They are calculated on the basis of the power settings that one chooses at any given time.

In summary, this article provides an overview of medical diagnostic ultrasound imaging systems and their status at the turn of the century. These systems are primarily of the pulse-echo type. The ultrasound properties of biological media were considered first so that their implications for system design and image characteristics could be understood. Image quality metrics were defined for the B-scan modality (60,61). Diffraction or beam divergence is a physical phenomenon that occurs as a result of propagation. A basic theory of diffraction from circular disk transducers has been presented. The main objective was to use it to describe the general features of the beam pattern and their implications for B-scan imaging. Technological developments have been fast paced in the last decade. Those that have become clinically successful or hold great potential, for example, harmonic imaging, computer processing of multichannel signals, contrast imaging, 3-D imaging, and elasticity imaging, have been considered here. The Doppler ultrasound technique for blood flow estimation has witnessed major advancements in recent years, therefore, it was included in this section. Most of the advance that have taken place are within the context of the B-scan imaging format (62). Finally a current understanding of the risks of ultrasound is summarized now in terms of two recently formulated indexes, TI and MI.

BIBLIOGRAPHY

1. K. K. Shung, M. B. Smith, and B. Tsui, *Principles of Medical Imaging*, Academic Press, San Diego, 1992, pp. 78–159.
2. F. W. Kremkau, *Diagnostic Ultrasound Principles and Instruments*, 5th ed., W.B. Saunders, Philadelphia, 1998.

3. H. Krautkramer, *Ultrasonic Testing of Materials*, 4th ed., Springer-Verlag, Berlin, 1990.
4. L. S. Schmerr, *Fundamentals of Ultrasonic Nondestructive Evaluation: A Modeling Approach*, Plenum, NY, 1998.
5. R. C. Waag, *IEEE Trans. Biomedical Eng.* **31**, 884–892 (1984).
6. C. R. Hill, ed., *Physical Principles of Medical Ultrasonics*, John Wiley & Sons, Inc., New York, NY, 1986.
7. S. Fields and F. Dunn, *J. Acoust. Soc. Am.* **54**, 809–812 (1973).
8. P. M. Morse and K. U. Ingard, in *Theoretical Acoustics*, McGraw-Hill, NY, 1968.
9. J. F. Chen, A. Zagzebski, and E. L. Madsen, *IEEE Trans. Ultrasonics Ferroelectrics Frequency Control* **41**, 435–440 (1994).
10. V. M. Narayanan and P. M. Shankar, *IEEE Trans. Ultrasonics Ferroelectrics Frequency Control* **41**, 845–852 (1994).
11. L. F. Joynt, Ph.D. Thesis, Stanford University, 1979.
12. M. Helguera, Ph.D. Thesis, Rochester Institute of Technology, 1999.
13. J. C. Bamber and M. Tristam, in S. Webb, ed., *The Physics of Medical Imaging*, Adam Hilger, Bristol, 1988.
14. B. Hete and K. K. Shung, *IEEE Trans. Ultrasonics Ferroelectrics Frequency Control* **40**, 354–365 (1993).
15. P. N. T. Wells and M. Halliwell, *Ultrasonics*, **19**, 225–229 (1981).
16. E. J. Feleppa et al., *IEEE Ultrasonic Symposium Proceedings*, **2**, 1413–1417, 1999.
17. R. F. Wagner, M. F. Insana, and D. G. Brown, *SPIE Appl. Opt. Instrum. Med. XII* **535**, 57 (1985).
18. J. M. Thijssen et al., *Ultrasound Med. Biol.* **19**, 13–20 (1993).
19. G. E. Sleefe and P. P. Lele, *Ultrasound Med. Biol.* **14**, 181–189 (1988).
20. Z. Cho, J. P. Jones, and M. Singh, in *Foundations of Medical Imaging*, John Wiley & Sons, Inc., New York, NY, 1993, Chap. 14.
21. R. T. Bayer and S. V. Lecher, *Physical Ultrasonics*, Academic Press, NY, 1969.
22. P. J. A. Frinking et al., *Ultrasound Med. Biol.* **26**, 965–975 (2000).
23. K. Wie et al., *Circulation* **97**, 472–483 (1998).
24. J. D. Gaskill, *Linear Systems, Fourier Transform and Optics*, John Wiley & Sons, Inc., New York, NY, 1978.
25. P. N. T. Wells, *Biomedical Ultrasonics*, Academic Press, London, 1977.
26. R. M. Gagliardi, *Introduction to Communication Engineering*, 2nd, ed., John Wiley & Sons, Inc., NY, 1988.
27. S. H. Maslak, in R. C. Sander and M. C. Hill, eds., *Ultrasound Annual 1985*, Raven Press, NY, 1985, pp. 1–16.
28. M. F. Insana and T. J. Hall, in P. N. T. Wells, ed., *Advances in Ultrasound Techniques and Instrumentation*, Churchill Livingstone, NY, 1993, pp. 161–181.
29. A. Macovski, in *Medical Imaging Systems*, Prentice-Hall, NJ, 1983.
30. N. A. H. K. Rao, S. Mehra, J. Bridges, and S. Venkataraman, *Ultrasonic Imaging* **17**, 114–141 (1995).
31. G. S. Kino, *Acoustic Waves: Devices, Imaging, and Signal Processing*, Prentice-Hall, Englewood Cliffs, NJ, 1987.
32. H. Seki, A. Granato, and R. J. Truell, *J. Acoust. Soc. Am.* **28**, 230–238 (1956).
33. P. Stepanishen, *J. Acoust. Soc. Am.* **49**, 1,627–1,638 (1971).
34. G. Harris, *J. Acoust. Soc. Am.* **70**, 10–19 (1981).
35. C. J. Daly and N. A. H. K. Rao, *J. Acoust. Soc. Am.* **105**, 3,067–3,077 (1999).
36. C. J. Daly and N. A. H. K. Rao, *Scalar Diffraction from a Piston Aperture*, Kluwer Academic, Boston, 2000.
37. G. Kossoff, *Ultrasound Med. Biol.* **5**, 359–365 (1979).
38. D. Liu and R. C. Waag, *IEEE Trans. Ultrasonics Ferroelectrics Frequency Control* **44**, 1–13 (1997).
39. J. W. Goodman, *Introduction to Fourier Optics*, McGraw-Hill, NY, 1968, Chap. 3.
40. K. W. Ayer, Ph.D. Thesis, Rochester Institute of Technology, 2000.
41. A. J. Angelsen, in *Ultrasound Imaging — Waves, Signals, and Signal Processing*, 1st, ed., Emantec, Trondheim, Norway, 2000. \http://www.ultrasoundbook.com.
42. F. W. Kremkau, in *Diagnostic Ultrasound: Principles and Instruments*, 5th, ed., W.B. Saunders, Philadelphia, 1998, Chaps. 5 and 7.
43. C. Kasai, K. Namekawa, A. Koyano, and R. Omoto, *IEEE Trans. Sonics Ultrasonics* **32**, 458–463 (1985).
44. P. N. T. Wells, *Phys. Med. Biol.* **39**, 2,113–2,145 (1994).
45. T. S. Desser and R. S. Jeffrey, *Diagnostic Imaging*, **21**, No. 9, 63–69, 1999.
46. M. A. Averkiou, D. N. Roundhill, and J. E. Powers, *Proc. 1997 IEEE Ultrasonic Symp.*, 1997, pp. 1,561–1,566.
47. G. Wojcik, J. Mould, S. Ayter, and L. Carcione, *Proc. 1998 IEEE Ultrasonic Symp.*, 1998, pp. 1,583–1,588.
48. T. Christopher, *IEEE Trans. Ultrasonics Ferroelectrics Frequency Control* **45**, 158–162 (1998).
49. J. Ophir and K. J. Parker, *Ultrasound Med. Biol.* **15**, 319–333 (1989).
50. S. D. Hope, C. T. Chin, and P. N. Burns, *IEEE Trans. Ultrasonics Ferroelectrics Frequency Control*, **46**, 372–382 (1999).
51. P. M. Shankar, K. P. Dala, and V. L. Newhouse, *Ultrasound Med. Biol.* **24**, 395–399 (1998).
52. A. Fenster and D. B. Downey, in *Handbook of Medical Imaging: Physics and Psychophysics*, vol. 1, 1st, ed., The International Society of Optical Engineering (SPIE), 2000, Chap. 7.
53. O. T. von Ramm, S. W. Smith, and H. G. Pavy, *IEEE Trans. Ultrasonics Ferroelectrics Frequency Control* **38**, 100–108 (1991).
54. J. Ophir et al., *Ultrasonic Imaging*, **13**, 111–134 (1991).
55. L. S. Taylor, B. C. Porter, D. J. Rubens, and K. J. Parker, *Phys. Med. Biol.* **45**, 1,477–1,494 (2000).
56. S. Y. Emelianov et al., *Proc. IEEE Ultrasonics Symp.*, 1997, pp. 1,123–1,126.
57. J. Ophir et al., *Int. J. Imaging Syst. Technol.* **8**, 89–103 (1997).
58. American Institute of Ultrasound in Medicine (AIUM): *Bioeffects and Safety of Diagnostic Ultrasound in Medicine*, Laurel, MD, 1993.
59. National Council on Radiation Protection and Measurements (NCRP) Report No. 113, *Exposure Criteria for Medical Diagnostic Ultrasound: I. Criteria Based on Thermal Mechanisms*, Bethesda, MD, 1992.
60. C. R. Hill et al., *Ultrasonics Med. Biol.* **17**, 559–575 (1991).
61. R. A. Harris et al., *Ultrasonic Med. Biol.*, **17**, 547–558 (1991).
62. J. Lowers ed., *Diagnostic Imaging Special Supplementaion on Advanced Ultrasound*, pp. 1–32, (Nov 2000).

V

VIDEO RECORDING

PETER D. LUBELL
Rainbow Network Communications
Woodbury, NY

Long before television became a household word, man endeavored to capture and record visual images. By the time the first *electronic* television was introduced at the 1939 World's Fair, both still and motion photography were well established in black-and-white and in color. Motion pictures, such as *The Wizard of Oz*, had full color images and complete audio sound tracks. The following section will provide an overview of video recording progress (both visual and aural signals) over almost 50 years since that 1939 World's Fair demonstration (see Fig. 1).

CHRONOLOGY

After the development hiatus caused by the diversion of scientific talent and engineering skills to the war effort during World War II, 1947 saw the introduction of commercial broadcast television, and by the mid-1950s color broadcast TV was on the scene as well.

Virtually from its beginning, the television industry required a means of recording and storing the live video images that it was producing. In his famous *three wishes* speech in the fall of 1951, Brigadier General David Sarnoff, then chairman of RCA and founder of the National Broadcasting Company, called for the development of a device capable of recording television signals with a magnetic tape medium (1).

In 1951 the fledgling television networks relied on a photographic film process known as kinescope recording. This involved the transferring live television programming to motion picture film by photographing a high-intensity cathode ray tube and then reconstructing the video image by scanning the filmed image with a high-resolution camera tube called a *flying-spot* scanner. The resulting *kines* were often of poor visual quality due to the joint problems of loss of image definition during the filming and the reconstruction process and to synchronizing discrepancies caused by the difference between television's frame rate of 30 per second and the 24 frames per second rate of the normal movie camera. Notwithstanding these problems, the use of *kines* was immense. It was estimated that, during the 1950s, the television industry was a larger user of movie film than the movie industry itself (2).

1939	Introduction of *electronic* television
1947	End of wartime hiatus (WW II)
1948	B&W recording via kinescope
1956	First commercial videotape recorder (VTR)—Ampex
1970	First videocassette recorder (VCR)—Sony U-Matic
1975	First consumer videocassette recorder—Sony Betamax
1976	Introduction of VHS format consumer VCR—JVC et al.
1977	Videodisk players—RCA and Philips
1978	Laserdisk players—Philips et al.
1980	Commercial HDTV tape recorder (VTR)—Sony Hi-Vision
1985	Commercial, digital, component VCR—Sony, D-1 Format
1989	Commercial, digital, composite VCR-Sony, D-2 format
1995	Digital versatile disk (DVD) Player—Sony, et al.
199?	HDTV digital disk (DVD) player

Figure 1. Video recording milestones — timeline.

Soon after Sarnoff's speech, in December 1951 a small group of engineers at the Ampex Corporation began a project to develop the prototype of the videotape recorder as we know it today. By 1956, Ampex had introduced a reel-to-reel, commercial model, using magnetic tape media supplied by the 3M company. For the next 15 years Ampex continued to provide the broadcast industry with ever more potent videotape recorders (VTRs) (3).

In 1970 the Sony Corporation, under license from Ampex, introduced the three-quarter-inch U-Matic, color videocassette player for the commercial market. In 1975 Sony established the basis for the home video recorder market by introducing the one-half-inch Betamax (SL-6300) consumer videocassette recorder (VCR). Both of these machines utilized tapes wholly contained within a cartridge, the cassette, which precluded the need to thread tape, as was the case then with the reel-to-reel machines (4).

In 1976 the Japan Victor Corporation (JVC), and others introduced a competitive and incompatible cassette recording format (VHS) for consumer use. Within a few years the VHS format became the dominant format, virtually eliminating Beta in the consumer marketplace. Despite this, as we shall see, Beta remains alive and well in the commercial realm to this day (5).

In 1977, the RCA Corporation and Philips Electronics each introduced a videodisk player to compete with the VCR in the consumer arena. Both of these machines employed analog recording schemes not to be confused with the digital laserdisk, introduced in 1978. None of these three formats permitted recording by consumers at home. They were offered only as prerecorded material, and they therefore had a difficult time competing with the VCR. Within a few years the first two fell by the wayside, leaving only the laserdisk, produced by Philips and others, with its inherently superior picture and audio quality to retain even a small foothold in the VCR-dominated consumer marketplace (6).

For almost a decade, until the 1995 announcement of the digital versatile disk, consumer video recording technology remained static.

Not so the commercial arena, which saw the transition of analog tape recording to digital component and then to digital composite formats. This has been followed by the application of high-density magnetic recording technology which has brought forth hard disk drives for video recording, in turn, leading to computer-based, nonlinear,

virtual video tape recorders for commercial recording, editing and playback, and origination applications. (A brief description of this *nonlinear* video recording technology can be found at the end of this article, entitled *nonlinear video*. A more detailed discussion can be found elsewhere in this publication under the same title.)

It took almost 2 years from the initial announcement of the high-density digital versatile disk (DVD) for it to appear in 1997 in the consumer marketplace. Now it is only available in the *prerecorded* format, although the technology currently exists to produce a *recordable* disk, once the DVD format has become an established format. Even now, development is continuing on an enhanced DVD capable of handling a digital, high-definition television signal (7).

DESCRIPTION OF THE TECHNOLOGIES

Throughout this article, the term *video* represents (except as specifically noted) a television signal consisting of program video, program audio (1, 2, or 4 channels) and a time-code channel.

For historical purposes we include a brief description of the kinescope process to provide a perspective on the more sophisticated magnetic and optical technologies which followed it and which dominate the video recording field.

Kinescope Recording

In its simplest form, kinescope recording involves filming a television image from a high-intensity picture tube and then playing it back by scanning the film with a *flying-spot* scanner to convert the filmed image to a standard video format (Fig. 2). This process allowed for time-delayed rebroadcast of television programming throughout the nation. It did so, however, at a high cost and significantly reduced picture quality.

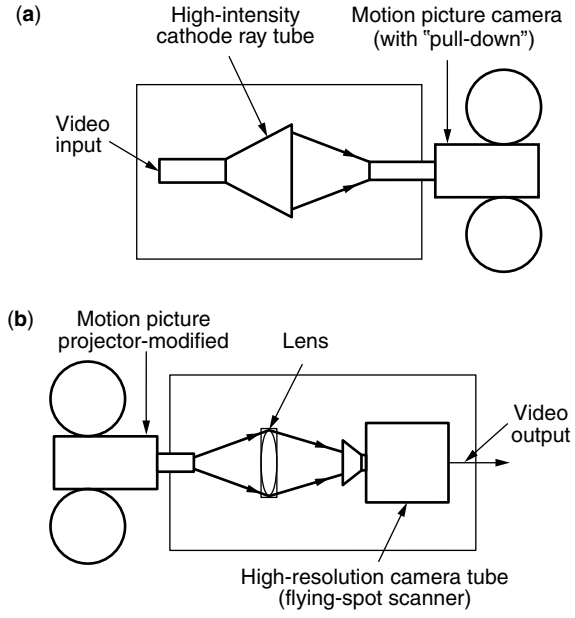


Figure 2. Basic kinescope (**a**) recording and (**b**) playback.

Picture quality suffered for two reasons; the required image-to-film and film-to-image transfers were done with limited light levels, producing grainy images of low contrast. A second cause was the frame-rate difference between television (30 frames per second) and cinematography (24 frames per second). This latter problem was circumvented by using a "pull-down" mechanism in the motion picture camera which allowed exposing one film frame to multiple television fields (two or three) so that the 24/30 ratio was reversed. The resultant product was low in contrast and image stability.

Although the broadcast use of kinescope recording disappeared with the introduction in 1956 of the Ampex videotape recorder, there are still special situations where kinescopic recording is used. Most of the timing problems have been overcome through electronic time-base correction circuits and frame-storage techniques. As we shall see, however, the paramount technology in television storage from 1956 to the present has been magnetic tape recording (8).

Magnetic Tape Recording-Analog, Linear Transport

The first commercially practical video recorder was developed in the United States by the Ampex Corporation during the period 1951 through 1956. It recorded monochromatic (black-and-white) images from a National Television Standards Committee (NTSC), 525-line, 60 Hz signal with monaural sound. This machine, designated the VR-1000, achieved its breakthrough performance by using four moving recording heads, rotating in a plane transverse to the direction of (two-inch wide) tape motion. A second contributor to the success of the VR-1000 was its application of frequency modulation (FM) in the recording process (9).

Earlier efforts to produce a videotape recorder (VTR) were based on extensions of the magnetic tape technology developed for audio recording. This approach involved moving the magnetic tape longitudinally past one or more fixed recording heads and also the direct transfer of signal level through amplitude modulation of the magnetic field being applied to the tape (Fig. 3). These early systems failed to pass muster practically because they required excessive tape speed, as high as 20 feet per second, which resulted in limited recording times, and because their use of direct AM level transfer produced unstable record and playback performance due to the nonuniform magnetic characteristics of the tapes available then. Lack of stability

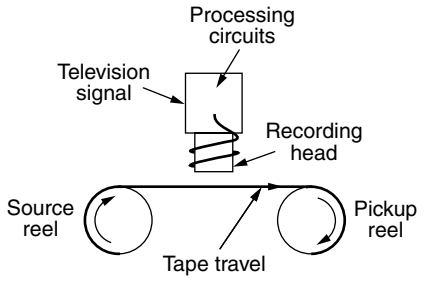


Figure 3. Linear transport; direct tape recording.

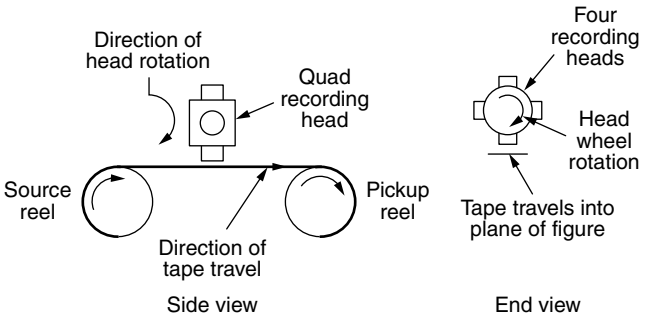


Figure 4. Quad-head tape recording.

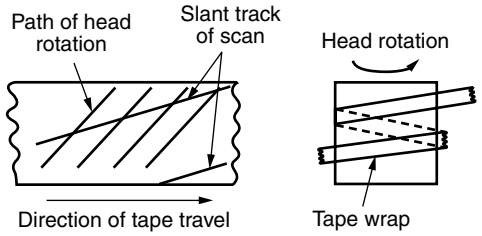


Figure 5. Helical scan (transverse) tape recording.

in their tape transport mechanisms also contributed to recording artifacts, such as *jitter, flutter* and *wow* (10).

The Ampex machine resolved these issues in the following manner (Fig. 4). Four recording heads were mounted on a cylindrical drum which rotated at high speed (14,400 rpm) in a plane whose axis was at right angles to the motion of the tape. The tape speed was only 15 inches per second, a substantial reduction from the earlier British Broadcasting Company (BBC) and RCA designs which employed tape speeds of 200 and 3,650 inches per second, respectively. In the Ampex machine, each of the four video recording heads laid down 16 horizontal television lines per revolution of the revolving head drum. Thus one drum revolution recorded 64 lines and only eight tape passes were required to complete the entire 525-line NTSC video frame.

Another significant feature of the VR-1000 was its use of FM technology for the signal-to-tape information transfer. This approach avoided the pitfalls of the direct methods used in the earlier VTRs because now the information was recorded in the frequency domain using incremental changes about a fixed carrier. The tape stock available then was much more uniform in its frequency-dependent characteristics than in its amplitude response to changing magnetic bias (11).

For almost two decades this *Ampex Quad* VTR, with minor refinements, was the commercial standard for television recording.

Helical Scan and the Cassette Format

During the next 15 years, companies in the United States and Japan focused their attention on better media (tape stock), improved tape utilization (higher density recording), and a more convenient transport mechanism.

The Minnesota Mining and Manufacturing (later 3M) company was a principal provider of media for television recording in the United States during this period. The early product was a paper-backed tape with a magnetic coating on one side. By late 1947, 3M had converted to acetate film for base stock. By the time the VR-1000 was first commercially used by CBS on November 30, 1956), the tape was an improved acetate-based 3M product called *Scotch 179 Videotape* (12).

The next step in the evolutionary process focused on improvements in tape scanning. To improve the performance of its Quad machine, Ampex and its licensees (RCA and the Victor Corporation of Japan [JVC]) worked to develop a *helical* scanning format which effectively lengthened the scan distance per head pass by using a tangential rather than a simple longitudinal course (Fig. 5). This type of scanning has become the accepted standard for VTR in most applications (13).

In 1962 machines were introduced by Ampex, RCA, and JVC utilizing helical scanning with one-inch wide tape stock. The VR-1000 Quad machines used longitudinal scanning with two-inch wide tape. By the end of the 1960s, a helical scan VTR produced by the International Video Corporation (the IVC-9000), using two-inch tape, for the first time exceeded the performance of the Ampex Quad machine, then the broadcast industry standard. The final nail in the coffin of the Quad process came when digital time-base correction circuits were devised by the Japanese in 1968, which allowed the long-scan, helical process machines to exceed the stability performance of the shorter track Quad system (14,15).

The resulting improved helical scan VTRs became widely used in broadcast applications in 1970. These systems employed three-quarter-inch wide tapes as opposed to the one-inch tapes of their predecessors.

Until this time even the improved helical scan machines utilized open reel-to-reel tape formats. In 1970, Sony made a major breakthrough in VTR design by introducing the prototype U-Matic color video recorder. This was an analog, helical-scan machine which utilized three-quarter-inch wide tape packaged in a self-contained cassette. No longer did the tape installation process require careful and time-consuming threading of loose tape by a highly experienced technician. The cassette package contained both source and pickup reels with the tape installed at the factory. Tapes could now be changed in seconds by someone with virtually no prior experience (16).

An improved, production version of the U-Matic recorder (the VP-1000) was offered for sale by Sony in 1972. Thus was born the video cassette recorder (VCR) and the groundwork was laid for a consumer product (17).

Video Cassette Recording (VCR)

The Ampex Quad machines had always been too complex and expensive to serve as a basis for translation to the consumer market. The Sony VP-1000 was initially conceived as just such a product because its cassette could be readily installed and removed and its use of advanced, lower cost solid-state technology was a further step in this direction. (Ampex had introduced the first fully transistorized VTR, the VR-1100 in 1962,

but this machine was still large and cumbersome and had setup requirements too complex to be considered for consumer use.)

This prototype U-Matic machine was a cooperative effort of three Japanese companies, Sony, JVC, and Matsushita (known by its US trade name, Panasonic). For this reason, this three-quarter-inch format was termed 3*C* in Japan. For its efforts in developing the VCR, Sony was awarded an EMMY in 1976 by the Motion Picture Academy of Arts and Science for technical achievement. In actuality, all three companies made significant contributions to the development of the VCR (18).

The early U-Matic helical scan machines were not an unqualified success for commercial applications. Their performance, as measured by resolution, signal-to-noise ratio (SNR), and stability, was not equal to the Quad machines then used by all broadcasters. The U-Matics, however, were compact, simple to operate, and inexpensive enough for field use, such as news gathering and other remotes. On the other hand, they were neither compact enough nor inexpensive enough to support their introduction as a consumer product. The three-quarter-inch cassette was still rather large and expensive. Another problem was their need for sophisticated time-base correction circuits to overcome the timing errors introduced by helical scanning of relatively narrow tape. Broadcasters eventually overcame these problems, first by introducing electronic time-base correction (TBC) in 1972 and later by using frame synchronizers in 1974 (19).

At this stage in 1974, the helical format machines were potential replacements for Quad machines in commercial applications, but not in the consumer market because they still cost too much.

Introduction of the Consumer VTR. Having achieved a somewhat competitive footing in the commercial marketplace, Sony and the other Japanese manufacturers again focused their attention on developing a consumer VCR. In 1970, Ampex and Toshiba jointly developed a one-half-inch cassette, monochromatic machine for the consumer market. It did not catch on! Neither did comparable machines introduced in 1972 by Panasonic and Philips (20).

During this period, JVC approached Ampex with a proposal to jointly develop a one-half-inch color VCR specifically for consumer use. To support their proposal, JVC had constructed a one-quarter-scale (one-half-inch format) working model based on the Ampex VR-1500 two-inch helical scan machine. The Ampex management was reportedly impressed with the prototype, but they gracefully declined to enter the world of consumer electronics at that time. Ampex, however, licensed JVC to use their technology in such an application. This was to prove a successful financial decision of momentous proportions for Ampex (21).

In 1975, Sony introduced a one-half-inch color consumer VCR using the Sony proprietary Beta (helical-scan) format. This machine, the Betamax SL-6300, was relatively compact and simple to operate. The Beta nomenclature was based on considering it a derivative of the U-Matic or Alpha format (22).

Within a year in 1976, JVC refined its prototype one-half-inch VCR, using its own, unique helical-scan format and proceeded to introduce it as the Video Home System (VHS). Thus did the *format wars* begin for the consumer market (23).

Although many technically astute persons believed that the Beta format was capable of performance superior to that of the VHS format, within a few years VHS had gained supremacy in the consumer marketplace. Many believe that this success was primarily due to JVC's decision to broadly license its format so as to encourage availability of a wide spectrum of prerecorded material to the consumer in a short time during which Beta titles were somewhat limited. Another drawback of the Beta format was its limited recording time compared to VHS. Even so, more than 20 years after their commercial introduction, U-Matic machines (Beta's origin) are still being produced and are widely used in the broadcast industry.

Early Disk (Nonmagnetic) Recording

In 1977, within two years after the introduction of the Sony Betamax, both RCA and Philips introduced a videodisk player for the consumer market. Both were play-only systems. The RCA version (called SelectVision) employed a capacitance pickup which monitored variations in a grooved, 12-inch diameter disk, rotating at 450 rpm. Each side provided up to 60 minutes of television programming (24).

The Philips version of the videodisk was quite different. It used a small laser to retrieve television information from a previously etched, grooveless, spiral track on a 12-inch disk. Philips called it the Video Long Play (VLP) system after the LP audio records current then. VLP played back previously programmed disks with a maximum program duration of two hours with the disk rotating at 150 rpm. A commercial version of the VLP system, offered by JVC and others as the Video High Density (VHD) system is still in use occasionally for special applications, such as institutional training programs and national sales promotions in the automobile business (25).

Because both SelectVision and VLP were play-only with no home recording capability, they did not receive significant consumer acceptance and were soon withdrawn from the marketplace.

The following year (1978) saw the introduction of the optical laserdisk, as we know it today. These units by Philips and others were playback-only machines using factory prerecorded material. However, they were digitally encoded and their reproduction performance was much better than that of the earlier VLP (analog) system and well above the performance of even today's VHS one-half-inch VCRs. A simplified description of the optical laserdisk recording and playback process is shown in Fig. 6. Many believe that this process led to the development of the audio Compact Disk introduced by Sony and Philips in 1982 (26).

For almost 20 years until the 1995 introduction of the DVD, the laserdisk has reigned as the gold standard for consumer video playback performance. Despite this reputation, the high cost of both the hardware and the software (discs) has limited its penetration into the

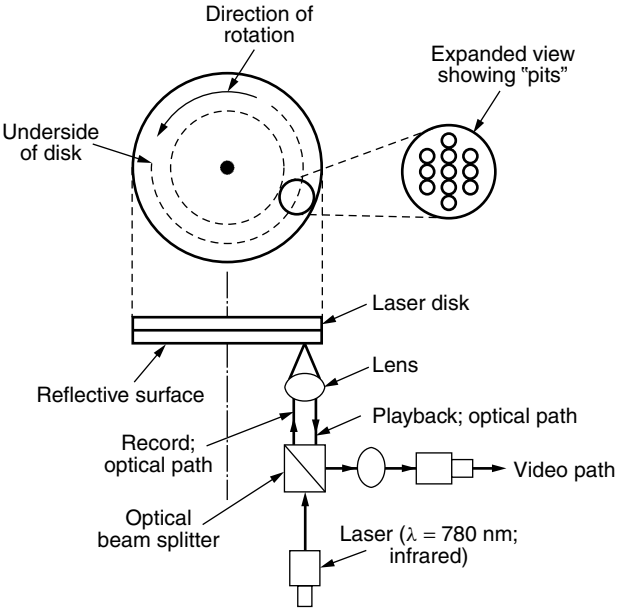


Figure 6. Laserdisk recording and playback (simplified).

U.S. consumer marketplace. This is not true in Japan, where virtually all new programming is made available simultaneously in both VHS cassette and laserdisk formats. In Japanese electronics outlets, both products remain equally competitive and laserdisk products are widely available. In the United States, after cumulative sales of more than 140 million VHS VCRs, the laserdisk has been virtually drowned in a sea of VHS tape as has the Beta tape format.

COMMERCIAL VTR EVOLUTION-COMPONENT AND COMPOSITE, DIGITAL TAPE RECORDING

Component Digital Tape Recording

Even as the consumer video industry was undergoing attempts to improve the playback performance of one-half-inch VCRs with optical and digital technologies, the commercial videotape recording field was experiencing a revolution of its own.

Until the mid 1970s, commercial videotape recording focused on handling the composite video signal which contained all elements, luminance (brightness), chrominance (color information), and synchronization and timing pulses, within a single envelope (Fig. 7). About the same time in 1978 as the consumer market saw the introduction of the digital laserdisk, the same signal processing technologies were being applied to commercial recording, first by separately processing the luminance and chrominance signals in the analog domain and then, subsequently, by applying digital techniques to this process.

Recording these two analog signal elements on separate video channels was termed *component* signal recording. It offered the advantage of reducing color artifacts during playback. This process eliminates the use of the 3.58 MHz color subcarrier and therefore results in reduced

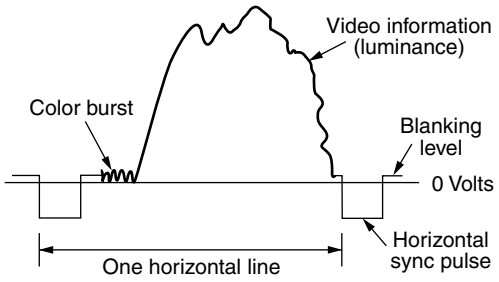


Figure 7. Composite television waveform.

susceptibility to cross modulation between component signals.

During the early 1980s two one-half-inch recording formats competed for market share for professional recording, the Sony Betacam SP (superior performance) and the Matsushita M-II formats. They were incompatible with each other because of significantly different signal processing techniques. The situation was reminiscent of the consumer format conflict (Beta versus VHS) of a decade earlier.

The Sony Betacam SP employed time-division multiplexing (TDM) to combine color component signals and record them alternately on a single color channel along with two stereo audio subcarriers at 310 kHz and 540 kHz, respectively. The luminance signal is recorded separately on a second channel (27).

The M-II format first sums the quadrature I and Q color elements into a single signal which is then recorded as the color track together with a pair of stereo audio subcarriers at 400 Hz and 700 kHz, respectively. A second channel contains the luminance information (28).

Composite Digital Tape Formats

As the analog component technology evolved, progress was being made in applying digital techniques, first employed in the consumer laserdisk, to commercial videotape recording.

In 1980, an international standards body, the International Radio Consultative Committee (CCIR), drafted Recommendation #601 which sought to define the characteristics of component digital recording. CCIR 601 was intended to be the recording standard for studios throughout the world. For many years thereafter, CCIR 601 was altered and modified. Finally, by 1985 the standards recommendations were firm enough for final approval. At this time Sony announced a digital component tape recording system based on the 601 recommendations. This D-1 system employed a 19-mm wide (approximately three-quarter-inch) tape in a cassette (similar to that employed in the U-Matic) and scanned in a helical manner with six tracks per field, two containing video data and four to support two pairs of stereo audio (29).

In 1987 Sony and Ampex working together brought forth the D-2 composite digital recording system. Like the D-1 format, the D-2 utilized a 19-mm tape and the same cassette as D-1 and provided four channels (two stereo pairs) of digital audio in addition to the video data. The scanning is also helical with six tracks per field (30).

The principal difference between the D-2 and D-1 formats is that D-2 is designed for a composite input signal which contains all of the color signal elements within a single envelope. Therefore the D-2 format is primarily for field recording whereas D-1 is for both studio and postproduction editing.

CONSUMER VIDEO-DIGITAL VERSATILE DISC (DVD)

For almost 20 years following its introduction in 1976, the VHS analog tape format remained preeminent in the consumer marketplace for video recording and playback. This was true, notwithstanding the ability of the laserdisk to provide much superior playback performance than even the enhanced VHS format, Super-VHS (S-VHS).

In 1995, barely 10 years after the highly successful introduction of the compact disk (CD) which provided digital, laser-scanned playback, an analogous versatile digital disk format (DVD) became available for video recording. Like the laserdisk before it, the DVD, as currently available, is a playback-only format.

The DVD is built on experience gained from the audio CD. However, it uses both a higher frequency (and shorter wavelength, red laser at 635 nm) and much more dense physical packing parameters (track spacing, groove width and length) to support an almost 6 : 1 increase in information storage capacity.

These improvements were not enough in themselves to provide the 10 : 1 storage improvement needed to go from 74 minutes of audio data (640–680 MB of information) to the 4,000 MB needed to provide over two hours (typically, 135 minutes) of standard resolution (NTSC) television programming. Data compression is used as well.

As is the case for the audio CD, the laser scans the reflective surface of the DVD disk from below, and like the audio CD, the DVD (as available today) uses factory prerecorded programming on a CD-like media (12 cm = 5.25 in. in diameter) (Fig. 8) (31).

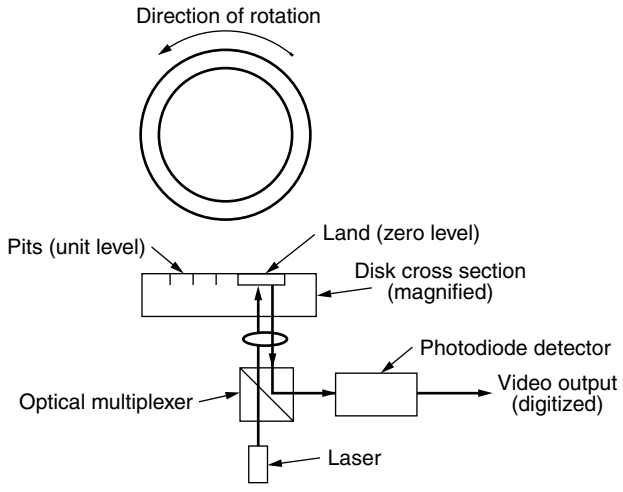


Figure 8. Digital versatile disk scanning (simplified).

FUTURE DVD ENHANCEMENTS

Even as DVD is being initially deployed in the marketplace, companies are investigating enhancement technologies to provide increased (double, quadruple, or higher) information storage capacity. One possible advance, derived linearly from present technology, seeks to replace the present binary digital encoding scheme with a multilevel scheme (nine states) which interprets the depth of the recorded *pits* to allow for coding beyond simply one or zero. The ability to read nine levels (from zero to eight) allows for a CD-sized disk capable of storing over two hours of HDTV-formatted television programming including three distinct stereo audio channels (32).

A second potential avenue of increased storage capability focuses on substituting a blue light laser at a wavelength of 390 nm for the red light laser at 635 nm currently employed in DVD recording. Because data storage capacity follows an inverse-square relationship with wavelength, this blue light technology could provide more than a 2.5 : 1 increase. Combining this with the use of smaller pit dimensions and closer track spacing, the total improvement could reach 4 : 1. This would lead to a storage capacity of over 15 GB on a 12 cm disk, enough capacity to provide five to seven hours of standard resolution (480 lines, NTSC) television or as much as two hours of HDTV (at 1,080 lines) (33).

DVD HOME RECORDING

Even as the aforementioned investigations of increased capacity are being pursued, attention is also being given to technologies which address the ability of the consumer to record at home, thereby creating a real replacement for the VCR.

At least two main approaches are being followed at this time; optical phase change (OPC) and magneto-optical (MO). OPC technology exploits the heating effect of laser energy to alter the reflectivity of the DVD disk so that this change can be read optically as a phase difference. The current DVD technology reads the signal amplitude digitally as one of a pair of binary states, either one or zero.

Magneto-optical (MO) technology is not really new. It has been used for digital data storage for some time, and it is the basis of the recordable MiniDisc audio system introduced by Sony in 1992. As used in the MiniDisc (MD) system, virtually simultaneous erasure and recording is possible by having both a magnetic recording head and a *heating* eraser laser coaxially located, respectively, above and below the recording disk (Fig. 9).

To record, the laser first heats the disk to its *Curie point* (about 400 °F). This causes the magnetic orientation of that portion of the disk to become totally random. Next, the magnetic recording head above the disk reorganizes the magnetic particles on the disk to represent one of the two binary states, one or zero. Then this orientation can be read back during playback, analogously to the digital optical process used in the current DVDs (33).

Once the disk has cooled down, its magnetic orientation is firmly set, even in the presence of significant magnetic

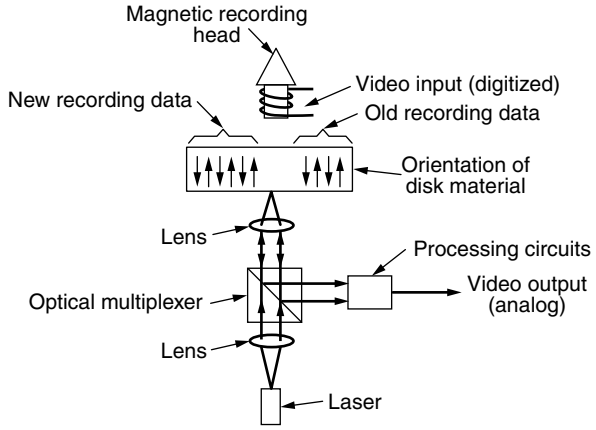


Figure 9. Magneto-optical recording.

fields. It can be altered only by applying the equivalent of the laser heating cycle.

The OPC approach is quite attractive from a manufacturing perspective because it employs many of the components and circuits of the current DVD system. However it is slightly limited in the number of write/rewrite cycles (to between 100,000 and one million), whereas the MO system has virtually no limit on its rewrite capability (34).

NONLINEAR RECORDING-COMMERCIAL VIDEO SERVERS

A *video server* is a computer-based system which provides storage and playback of a variety of video formats. It consists of two major components (Fig. 10), the computer, which serves as an interface between the video environment and the storage device, and the hard disk storage device (35).

During recording, the computer converts the input video information (both video and audio) into a digital data stream formatted for storage on a hard disk storage array. The input can be either analog (NTSC or component) or digital (uncompressed or compressed) in format. The resulting digital data stream, delivered to the hard drive, can be either uncompressed or compressed. Compressing the data prior to storage provides for efficient use of storage capacity at the cost of minor reduction in video quality during playback.

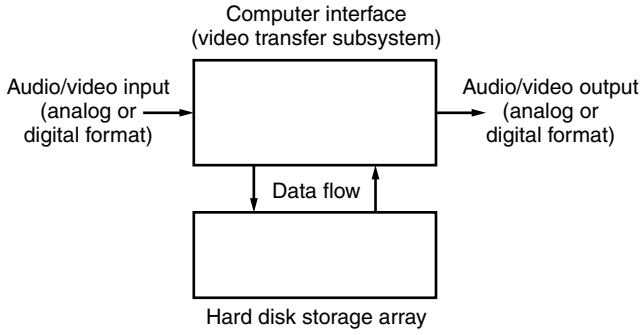


Figure 10. A basic video server.

During playback, the data file is retrieved from the storage device and converted back into a appropriate television format. This playback format may be either the same as, or different from, the input as the operator may desire. The process is termed *nonlinear* because video information can be accessed from any segment of the complete file rather than linearly, starting from the beginning of the file, as is the case for tape.

Video server applications include broadcast editing, commercial insertion, and preparation of interstitial insertion material (clip reels). An aggressive adopter of this technology is the television news sector which employs the nonlinear instant access capability of the VS systems to enhance its creation of timely news programming.

Video servers can also be used to provide pay per view and near-video-on-demand (NVOD) type service to subscribers on a delivery system which has feedback capability from the subscriber's location to the server site (36).

Now, video server technology is in a period of rapid growth with the application of constantly improved technology. Philips, Sony, and Hewlett-Packard are some of the companies currently active in this field (37).

The storage array concept allows multiple diskrete disks to act as a single unit. The use of such multiple disks and drives provides redundancy in the event a single device (either disk or drive) fails. Even using compression of the storage data stream requires the use of multiple parallel drives to support the data transfer rate. A typical disk drive supports a data rate of three to five MB/s. A full digital component data stream can produce a rate of 35 MB/s. Multiple disk drives combined with parallel data paths can support up to 40 MB/s. Use of compression techniques can reduce the 35 MB/s stream to an effective rate of less than 20 MB/s which then can be handled by a single path storage array (38).

Because current storage technology is evolving at such a rapid pace, it is virtually impossible to predict the eventual structure of these nonlinear systems, even in the near future.

BIBLIOGRAPHY

1. S. Wolpin, *Invention Technol.* **10**(2), 52 (1994).
2. A. Abramson, *SMPTE J.* **64**, 72 (1955).
3. C. Ginesburg, in The birth of video recording, *SMPTE 82nd Convention*, San Francisco, CA, October 5, 1957.
4. R. Mueller, *AV Video* **7**, 72 (1985).
5. J. Lardner, in *Fast Forward*, New York, W. W. Norton, 1987.
6. D. Mennie, *IEEE Spectrum* **12**, 34, August 1975.
7. P. Lubell, *IEEE Spectrum* **32**, 32, August (1995).
8. A. Abramson, *SMPTE J.* **90**, 508 (1981).
9. Ref. 3, p. 74.
10. Reference 5, pp. 51–52, and N. Heller, Getting a grip on tape, *Video Systems*, vol. 12, 26, 1986.
11. Ref. 3, pp. 74–75.
12. D. Rushin, HiFi to high definition: Five decades of magnetic tape, dB Magazine, Jan./Feb. 1991.
13. R. Warner Jr., Earl Masterson, *IEEE Spectrum* **33**, 51, February (1996).
14. A. Harris, *SMPTE J.* **70**, 489 (1961).

15. C. Bentz, *Video Systems* **11**, 49 (1985).
16. Ref. 4, p. 73.
17. S. Stalos, *AV Video*, **7**, 75 (1985).
18. Ref. 17, p. 76.
19. C. Bentz, *Broadcast Engineering* **27**, 80 (1985).
20. Ref. 2, p. 74.
21. Ref. 13, p. 57.
22. N. Kihara et al., *IEEE Trans. Consum. Electron.* **26**, 56 (1976).
23. Ref. 5, pp. 145–149.
24. W. Workman, The videodisk player, 25/4, Session 25, *ELECTRO/81 Prof. Program Session Record*, New York, April 1981.
25. K. Compaan and P. Kramer, *SMPTE J.* **83**, 564 (1974).
26. P. Rice and R. Dubbe, *SMPTE J.* **91**, 237 (1982).
27. K. Sadashige, *SMPTE J.* **98**, 25–31 (1989).
28. J. Roizen, *TV Broadcast* **8**, 68 (1985).
29. R. Hartman, *SMPTE J.* **95**, 12–15 (1986).
30. D. Brush, *SMPTE J.* **95**, 12–15 (1986).
31. Ref. 7, p. 33.
32. T. Wong and M. O'Neill, in 15 GB per side and no blue laser, *Data Storage*, April 1997, pp. 80–83.
33. MD Rainbow Book Sony Standard, San Francisco, CA, 1991.
34. *Phase-Change Rewritables*, Sony Product Release, Palo Alto, CA, September 1997.
35. C. Bernstein, *SMPTE J.* **106**, 511–518 (1997).
36. P. Hejtmanek, *Broadcast Eng.* **38**, 27 (1996).
37. P. Hejtmanek, *Broadcast Eng.* **39**, 36 (1997).
38. Digital video servers for on-air broadcast use, *SMPTE Seminar NAB Convention*, Las Vegas, NV, April 1997.

W

WAVELET TRANSFORMS

RAGHUVEER RAO
Rochester Institute of Technology, RIT
Rochester, NY

INTRODUCTION

In a 1984 paper, Grossman and Morlet (1) investigated the problem of representing square-integrable functions by using shifts and dilations of a single function called a wavelet. The motivation lay in the fact that there is evidence for the usefulness of such representations in analyzing seismic traces (2). Although earlier articles dealing with similar problems can be found, for example (3), the Grossman–Morlet paper provides an integral transform along with conditions under which one can invert the transform. Thus was born the field of wavelet transforms. Subsequent developments in the field such as the formulation of discrete basis functions, the relationship to subband filtering, and the application to signal/image processing have contributed to widespread interest in the subject (4). The purpose of this article is to provide a short overview of wavelet transforms and their application to certain image processing problems.

WHAT ARE WAVELETS?

Wavelets are continuous time functions that integrate to zero and are square-integrable. Thus, a one-dimensional wavelet $\psi(t)$ satisfies

$$\int_{-\infty}^{\infty} \psi(t)dt = 0 \tag{1}$$

and

$$\int_{-\infty}^{\infty} |\psi(t)|^2 dt < \infty. \tag{2}$$

A wavelet can be real or complex-valued. The function argument t is usually regarded as time. An example of a wavelet is shown in Fig. 1. The reason for using the term wavelet is attributed to the resemblance of the function to a decaying wave. Wavelets are band-pass functions in the sense that they have a Fourier transform magnitude that vanishes at the origin and that peaks at a finite frequency.

Consider

$$\psi_{a,b}(t)\frac{1}{\sqrt{|a|}}\psi\left(\frac{t-b}{a}\right). \tag{3}$$

where $a > 0$ and b are real-valued and the symbol stands for "defined as." The function $\psi_{a,b}(t)$ is a *mother wavelet* $\psi(t)$ dilated by a factor a and shifted in time by an amount

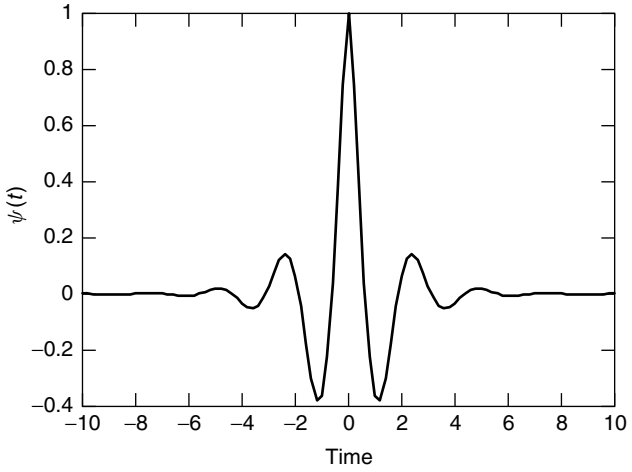


Figure 1. A wavelet.

b. The family of wavelets $\psi_{a,b}(t)$ parameterized by the *dilation* or *scale* a and the *shift* b are sometimes referred to as the *daughter wavelets* of $\psi(t)$. The energy or squared norm of each member of this family is the same because of the normalization factor $1/\sqrt{|a|}$ in Eq. (3). Figure 2 shows two daughter wavelets of the mother wavelet of Fig. 1.

CONTINUOUS WAVELET TRANSFORM

The continuous wavelet transform (CWT) of a finite-norm, one-dimensional signal $s(t)$ with respect to a mother wavelet $\psi(t)$ is given by

$$S(a,b) = \int_{-\infty}^{\infty} s(t)\psi_{a,b}^*(t)dt, \tag{4}$$

where the asterisk denotes complex conjugation. Defining the inner product of two functions $x(t)$ and $y(t)$ as $\langle (t), y(t)\rangle \int_{-\infty}^{\infty} x(t)y^*(t)dt$, the CWT becomes

$$S(a,b) = \langle s(t), \psi_{a,b}(t)\rangle. \tag{5}$$

Thus, the wavelet transform of a signal with respect to a mother wavelet is simply the collection of projections of the signal onto the set of daughter wavelets as a function of a and b.

Due to the effectively finite duration and bandwidth of wavelets, these transforms localize the signal information in both time and frequency. A *scalogram*, which is the squared magnitude of the wavelet transform is usually plotted to view the time versus scale information. As an example, consider the chirp waveform illustrated in Fig. 3. Its scalogram with respect to the *Mexican hat wavelet*

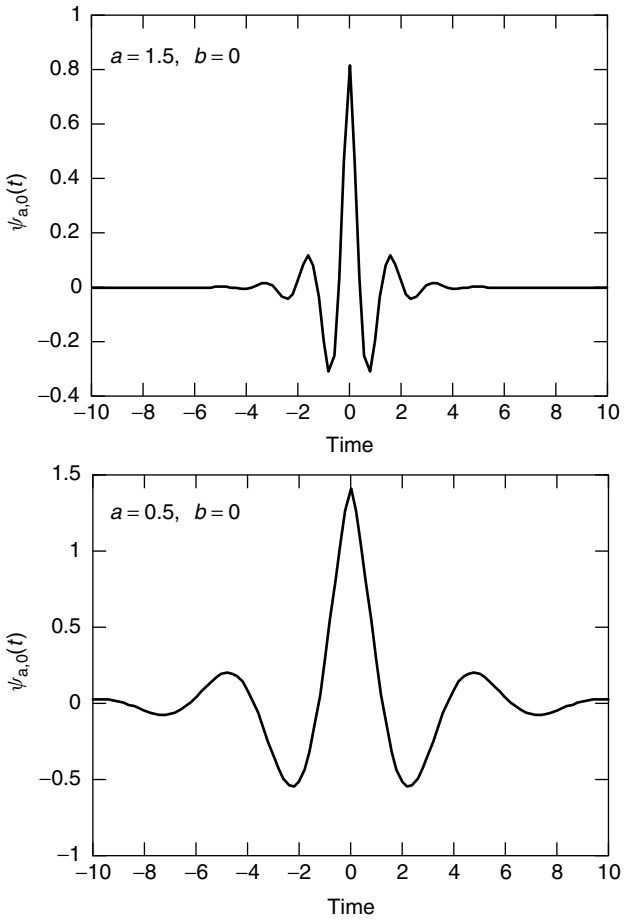


Figure 2. Wavelet dilations.

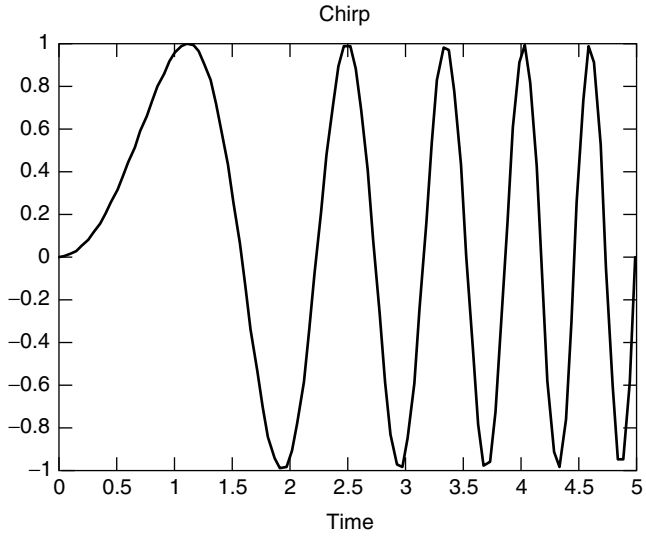


Figure 3. Chirp function $\sin(0.4\pi t^2)$.

(shown in Fig. 4),

$$\psi(t) = (1 - 2t^2)\exp[-t^2], \tag{6}$$

is displayed in Fig. 5.

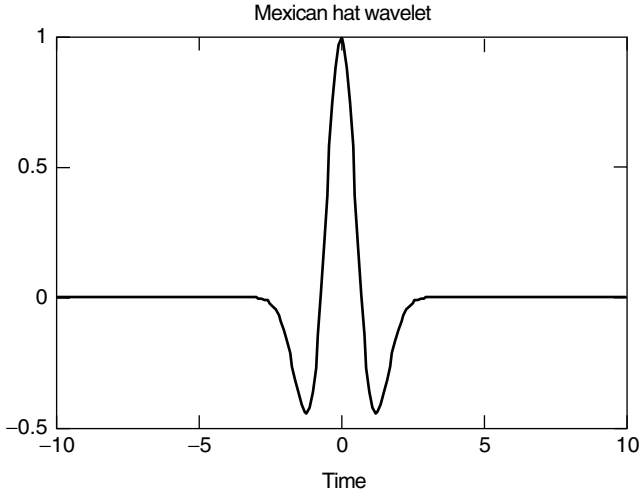


Figure 4

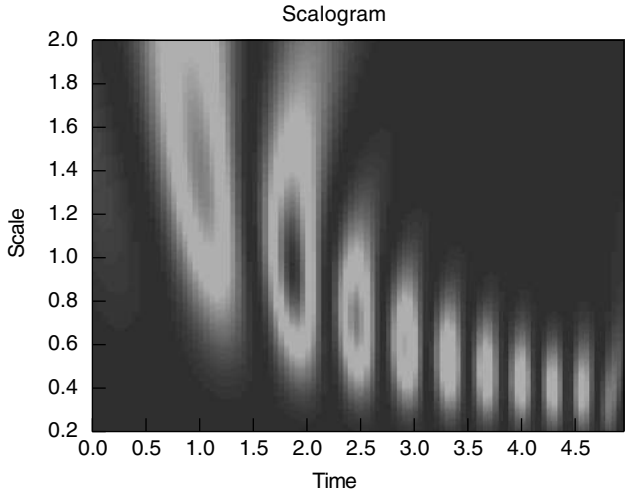


Figure 5. Scalogram of chirp. See color insert.

The scale variable bears an inverse relationship to frequency. As dilation increases, the zero crossing rate of the daughter wavelet decreases. Therefore, the wavelet transform at large values of a provides information about the lower frequency spectrum of the analyzed signal. Conversely, information contained at the higher end of the frequency spectrum is obtained for very small values of a. Thus, as seen in Fig. 5, the scalogram peaks occupy continually lower positions along the scale axis, as time progresses. This corresponds to increasing frequency of the chirp with time, as seen in Fig. 3.

An elegant expression exists for the inverse wavelet transform if a wavelet satisfies the admissibility condition. Suppose that

$$\Psi(\omega) = \int_{-\infty}^{\infty} \psi(t)\exp(-i\omega t)dt \tag{7}$$

is the Fourier transform of a wavelet $\psi(t)$. Then, the wavelet is said to satisfy the *admissibility condition* if

$$C\int_{-\infty}^{\infty}\frac{|\Psi(\omega)|^2}{|\omega|}d\omega<\infty \tag{8}$$

When the admissibility condition is satisfied, the *inverse continuous wavelet transform*, is given by

$$s(t)=\frac{1}{C}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\frac{1}{|a|^2}S(a,b)\psi_{a,b}(t)\,da\,db. \tag{9}$$

DISCRETE WAVELET TRANSFORM

It is possible to find mother wavelets so that one can synthesize functions from a set of their daughter wavelets whose dilation and shift parameters are indexed by the set of integers. In particular, dilations and translations from a *dyadic set*,

$$\psi_{a,b}(t): a=2^k, b=2^k i, i \text{ and } k \text{ are integers}, \tag{10}$$

provide the daughter wavelets.

Consider finite-norm functions $\phi(t)$ and $\tilde{\phi}(t)$ that have nonvanishing integrals satisfying

$$\phi(t)=2\sum_k h(k)\phi\,(2t-k) \tag{11}$$

$$\tilde{\phi}(t)=2\sum_k \tilde{h}(k)\tilde{\phi}(2t-k) \tag{12}$$

and

$$\left\langle \phi(t),\tilde{\phi}(t-k)\right\rangle=\delta(k), \tag{13}$$

where

$$\delta(k)\begin{cases}1 & \text{k = 0}\\ 0 & \text{otherwise.}\end{cases}$$

The functions $\phi(t)$ and $\tilde{\phi}(t)$are called scaling functions. Equations (11) and (12), called the *dilation equations* or *two-scale* relations, indicate that the scaling functions are generated as linear combinations of their own dyadic shifts and dilations. Equation (13) imposes orthogonality between one scaling function and integral translations of the other.

Two conditions that follow from Eqs. (11)–(13) are

$$\sum_k h(k)=\sum_k \tilde{h}(k)=1 \tag{14}$$

and

$$\sum_k h(k)\tilde{h}(k+2n)=\frac{\delta(n)}{2}. \tag{15}$$

Defining

$$\begin{aligned}H(\omega)&\sum_n h(n)\exp(-i\omega n)\\ \tilde{H}(\omega)&\sum_n \tilde{h}(n)\exp(-i\omega n).\end{aligned} \tag{16}$$

In the frequency domain, Eqs. (13) and (15) become

$$H(0)=\tilde{H}(0)=1, \tag{17}$$

$$H(\omega)\tilde{H}^*(\omega)+H(\omega+\pi)\tilde{H}^*(\omega+\pi)=1. \tag{18}$$

If the sequences $h()$ and $h^{\sim}()$ are viewed as impulse responses of discrete-time filters, then $H(\omega)$ and $H^{\sim}(\omega)$ are their frequency responses. It follows from Eqs. (17) and (18) that the product $H(\pi)\tilde{}H^*(\pi)=0$. Although this condition is satisfied by letting just one of the frequency responses be zero at a frequency of π radians, the interesting and practical cases require that both frequency responses be zero at π radians. Consequently, the filters assume low-pass characteristics.

Wavelets are constructed from scaling functions by forming sequences $g(n)$ and $g^{\sim}(n)$ as

$$\begin{aligned}g(n)&=(-1)^{1-n}\tilde{h}(1-n),\\ \tilde{g}(n)&=(-1)^{1-n}h(1-n),\end{aligned} \tag{19}$$

and then using the two scale relations given by Eqs. (11) and (12).

$$\begin{aligned}\psi(t)&=2\sum_k g(k)\phi\,(2t-k)\\ \tilde{\psi}(t)&=2\sum_k \tilde{g}(k)\tilde{\phi}\,(2t-k)\end{aligned} \tag{20}$$

The resulting wavelet $\psi(t)$ is orthogonal to $\tilde{\phi}(t)$, and its integer translates. Similarly, the wavelet $\tilde{\psi}(t)$ is orthogonal to $\phi(t)$, and its integer translates. The wavelet pair $\psi(t)$ and $\tilde{\psi}(t)$ are said to be duals of each other. Along with their dyadic dilations and translations, they form a *biorthogonal basis* for the space of square-integrable functions. Given a square-integrable function $f(t)$,

$$f(t)=\sum_i\sum_k a_{k,i}\psi_{2^k,2^k i}(t), \tag{21}$$

where

$$a_{k,i}=\left\langle f(t),\tilde{\psi}_{2^k,2^k i}(t)\right\rangle. \tag{22}$$

The special case of an *orthogonal wavelet basis* arises when the filter coefficients and scaling functions are their own biorthogonal duals, that is, $h()=\tilde{h}()$ and $\tilde{\phi}(t)=\phi(t)$. This results in a single wavelet $\psi(t)$ that is orthogonal to its own integral translates, as well as all of its dyadic dilations. The earliest known orthogonal wavelet is the *Haar wavelet* which appeared in the literature several decades before wavelet transforms (5) and is given by

$$\psi(t)=\begin{cases}1 & 0\le t<1/2\\ -1 & 1/2\le t<1\\ 0 & \text{otherwise.}\end{cases} \tag{23}$$

The Haar wavelet is a special case of an important class of orthogonal wavelets due to Daubechies (6). The coefficient sequence $h(k)$ in the dilation equation for this class of wavelets consists of a finite and even number of coefficients. The resulting scaling function and wavelet are

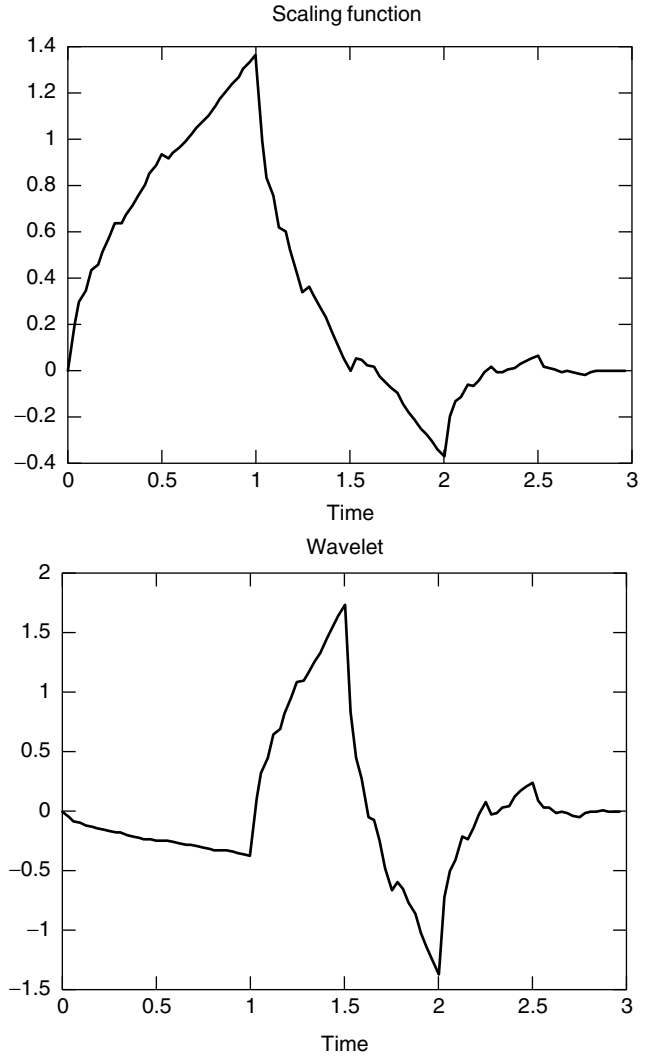


Figure 6. Daubechies four-tap filter. Top: scaling function. Bottom: wavelet.

compactly supported. A particularly famous example[1] is one involving four coefficients, $h(0) = (1 + \sqrt{3}/8)$, $h(1) = (3 + \sqrt{3}/8)$, $h(2) = (3 - \sqrt{3}/8)$, and $h(3) = (1 - \sqrt{3}/8)$. The corresponding scaling function and wavelet are shown in Fig. 6.

MULTIRESOLUTION ANALYSIS AND DIGITAL FILTERING IMPLEMENTATION

Let V_k denote the vector space spanned by the set $\{\phi\ (2^{-k}t - \ell) : \ell\}$ integer for an integer k. By virtue of Eq. (11), the vector spaces display the nesting

$$\ldots \subset V_1 \subset V_0 \subset V_{-1}. \tag{24}$$

There are other interesting properties:

$$x(t) \in V_k \Leftrightarrow x(2t) \in V_{k-1}, \tag{25}$$

[1] The Haar wavelet has two coefficients $h(0) = 1/2$ and $h(1) = 1/2$.

$$\lim_{k \to -\infty} V_k = L^2(R), \tag{26}$$

$$\bigcap_{k=-\infty}^{\infty} V_k = 0(t), \tag{27}$$

where $0(t)$ is the function whose value is identically zero everywhere.

The properties given by Eqs. (24)–(27) make the set of vector spaces $\{V_k : k = 0, \pm 1, \pm 2, \ldots\}$, a multiresolution analysis (MRA). Obviously, an MRA is also generated by $\{\tilde{V}_k : k = 0, \pm 1, \pm 2, \ldots\}$ where $\tilde{V}_k$ is the vector space spanned by $\{\tilde{\phi}(2^{-k}t - \ell) : \ell$ integer$\}$.

Suppose that $x_k(t)$ is some function in the vector space V_k. Expanded in terms of the basis functions of V_k.

$$x_k(t) = \sum_n x(k, n)\phi\ (2^{-k}t - n) \tag{28}$$

where, (by Eq. (13)),

$$x(k, n) = \left\langle x_k(t), 2^{-k}\tilde{\phi}(2^{-k}t - n) \right\rangle.$$

Notice how the coefficients of the expansion are generated by projecting $x_k(t)$ not onto V_k but rather onto $\tilde{V}_k$. Now, suppose that we have a square-integrable signal $f(t)$ that is not necessarily contained in any of the vector spaces V_k for a finite k. Approximations to this signal in each vector space of the MRA can be obtained as projections of the $f(t)$ on these vector spaces as follows. First, inner products $f_{k,n}$ are formed as

$$f_{k,n} = \left\langle f(t), 2^{-k}\tilde{\phi}(2^{-k}t - n) \right\rangle. \tag{29}$$

Then, the approximation $f_k(t)$ of $f(t)$ in V_k, referred to as the approximation at level k, is constructed as

$$f_k(t) = \sum_n f_{k,n}\phi\ (2^{-k}t - n). \tag{30}$$

For any k, $f_k(t)$ is a coarser approximation to the signal $f(t)$ than $f_{k-1}(t)$. Thus, there is a hierarchy of approximations to the function $f(t)$. As k decreases, we get increasingly finer approximations. Hence, the term multiresolution analysis.

The detail function $g_k(t)$ at level k is defined as the difference between the approximation at that level and the next finer level

$$g_k(t) = f_{k-1}(t) - f_k(t). \tag{31}$$

The detail represents the information lost in going from one level of approximation to the next coarser level and can be represented as a wavelet expansion at dilation 2^k as

$$g_k(t) = \sum_i a_{k,i}\psi_{2^k, 2^k i}(t), \tag{32}$$

where the coefficients $a_{k,i}$ are exactly as in Eq. (22). Because

$$f_k(t) = \sum_{j=k+1}^{\infty} g_k(t), \tag{33}$$

as can be seen by repeated application of Eq. (31), and

$$f(t) = \lim_{k \to -\infty} f_k(t), \tag{34}$$

Eqs. (32)–(34) lead to the result in Eq. (21).

There is a simple digital filtering scheme available to determine the approximation and detail coefficients, once the approximation coefficients are found at a given resolution. Suppose that we have determined coefficients $f_{0,n}$ at level 0. Then, the next coarser or lower[2] level approximation coefficients $f_{1,n}$ are found by passing the sequence $f_{0,n}$ through a digital filter whose impulse response is given by the sequence $2\tilde{h}(-n)$ and then retaining every other sample of the output.[3] Here $\tilde{h}(n)$ is the coefficient sequence in the dilation equation (12). Similarly, the detail coefficients $a_{1,n}$ are found by passing the sequence $f_{0,n}$ through a digital filter whose impulse response is given by the sequence $2\tilde{g}(-n)$ and once again retaining every other sample of the filter; we define $\tilde{g}(n)$ as the coefficient sequence in Eq. (19).

The block diagram in Fig. 7 illustrates these operations. The circle that has the down arrow represents the downsampling by two operations. Given an input sequence $x(n)$, this block generates the output sequence $x(2n)$ that is, it retains only the even-indexed samples of the input. The combination of filtering and downsampling is called decimation. It is also possible to obtain the approximation coefficients at a finer level from the approximation and detail coefficients at the next lower level of resolution by the digital filtering operation of interpolation, for which the block diagram is shown in Figure 8. The circle that has the up arrow represents upsampling by a factor of 2. For an input sequence $x(n)$, this operation results in the output sequence that has a value zero for odd n and the value $x(n/2)$ for even n. The process of filtering followed by upsampling is called interpolation.

In most signal processing applications where the data samples are in discrete time, wavelet decomposition has come to mean the filtering of the input signal at multiple stages of the arrangement in Fig. 7. An N-level decomposition uses N such stages yielding one set of approximation coefficients and N sets of detail coefficients. Wavelet reconstruction means the processing of approximation and detail coefficients of a decomposition through multiple stages of the arrangement in Fig. 8. A

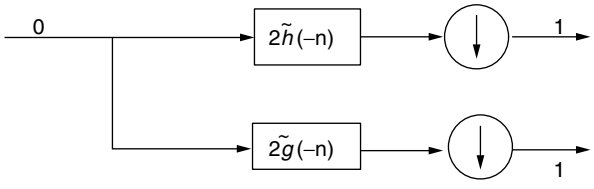


Figure 7. Block diagram of decimation.

[2] Lower because of the lower resolution in the approximation. Thus, approximation levels go lower as k gets higher.

[3] The reader unfamiliar with digital filtering terminology may refer to any of a number of textbooks, for example, Mitra (7).

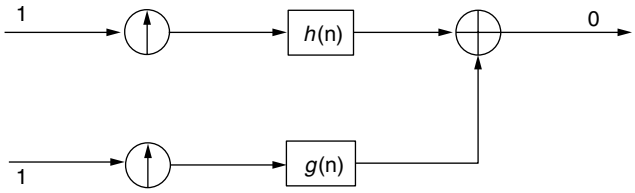


Figure 8. Block diagram of interpolation.

block diagram of a two-stage decomposition-reconstruction scheme is shown in Fig. 9.

Application to Image Processing

Wavelet applications of image processing are based on exploiting the localization properties of the wavelet transform in space and spatial frequency. Noise removal, or what has come to be known as *image denoising*, is a popular application of the wavelet transform, as is image compression. Other types of two-dimensional (2-D) wavelet constructs are possible, but most applications involve separable wavelet basis functions that are relatively straightforward extensions of 1-D basis functions along the two image axes. For an orthogonal system, these basis functions are

$$\begin{aligned} \psi_a(x,y) &= \psi(x)\phi(y) \\ \psi_b(x,y) &= \psi(y)\phi(x) \\ \psi_c(x,y) &= \psi(x)\psi(y). \end{aligned} \tag{35}$$

The scaling function

$$\phi_d(x,y) = \phi(x)\phi(y) \tag{36}$$

also comes in handy when the wavelet is transformed or decomposed is across a small number of scales.

Because images are of finite extent, there are a finite number of coefficients associated with the 2-D wavelet expansion on any dyadic scale. The number of coefficients on a given scale is one-quarter the number of coefficients on the next finer scale. This permits arranging the wavelet coefficients in *pyramidal form* as shown in Fig. 10. The top left corner of the transform in an *N-level* decomposition is a projection of the image on $\phi_d(x,y)$ at a dilation of 2^N and is called the *low-resolution* component of the wavelet decomposition. The other coefficients are projections on the wavelets $\psi_a(x,y)$, $\psi_b(x,y)$ and $\psi_c(x,y)$ on various scales and are called the *detail coefficients*. At each level, the coefficients with respect to these three wavelets are seen in the bottom-left, top-right and bottom-right sections respectively.

As can be seen from the figure, the detail coefficients retain edge-related information in the input image. Wavelet expansions converge faster around edges than Fourier or discrete cosine expansions, a fact exploited in compression and denoising applications. Most types of image noise contribute to wavelet expansions principally in the high-frequency detail coefficients. Thus, wavelet transformation followed by a suitable threshold zeroes

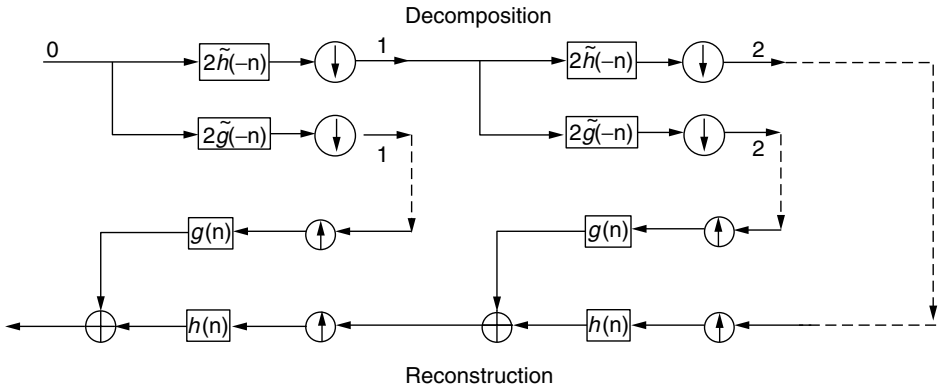


Figure 9. Two-level wavelet decomposition and reconstruction.

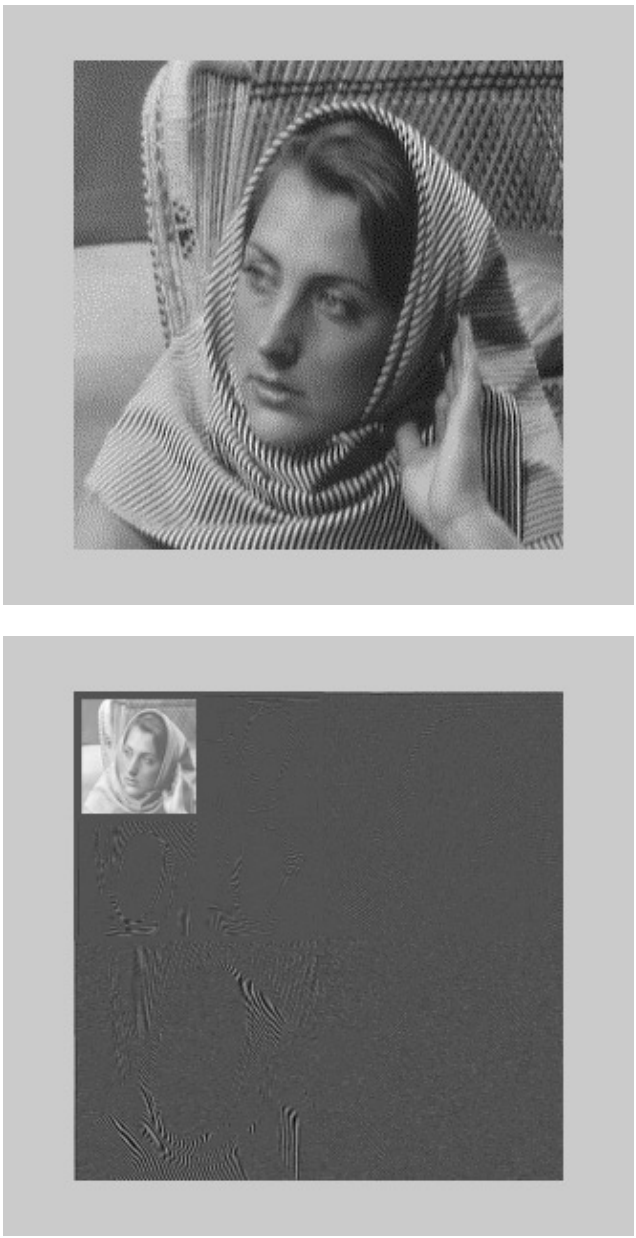


Figure 10. Top: image "Barbara." Bottom: Its two-level wavelet transform.

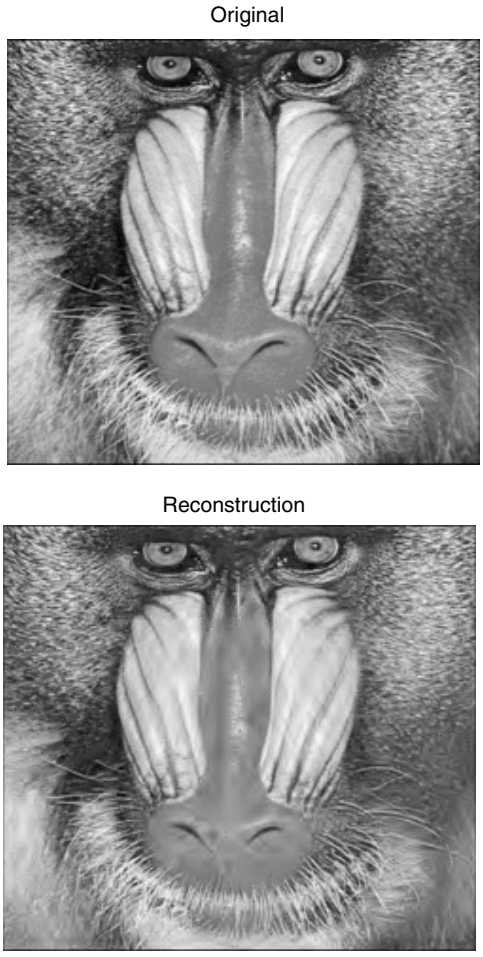


Figure 11. Illustration of image reconstruction using a fraction of its wavelet transform coefficients.

out many of these coefficients. A subsequent image reconstruction results in image denoising and minimal edge distrotion. An example is shown in Fig. 11.

Wavelets in Image Compression

Wavelet transforms have made their impact in the world of image compression, and this section provides insight

into the reason. The fast convergence rate of wavelet expansions is the key to the success of wavelet transforms in image compression. Among known linear transforms, wavelet transforms provide the fastest convergence in the neighborhood of point singularities. Although they do not necessarily provide the fastest convergence, along edges they still converge faster than the discrete Fourier or discrete cosine transforms (DCT) (8). Consequently, good image reconstruction can be obtained by retaining a small number of wavelet coefficients (9). This is demonstrated in Fig. 11 where a reconstruction is performed using only 10% of wavelet coefficients. These coefficients were chosen by sorting them in descending order of magnitudes and retaining the first 10%. This method of reducing the number of coefficients and, consequently, reducing the number of bits used to represent the data is called zonal sampling (10). Zonal sampling is only one component in achieving high compression rates. Quantization and entropy coding are additional components.

An examination of the wavelet transform in Fig. 10 reveals vast areas that are close to zero in value especially in the detail coefficients. This is typical of wavelet transforms of most natural scenes and is what wavelet transform-based compression algorithms exploit the most. An example of a compression technique that demonstrates this is the Set Partitioning in Hierarchical Trees algorithm due to Said and Pearlman (10,11). Yet another approach is to be found in the FBI fingerprint image compression standard (12) which uses run-length encoding.

Most values in the detail regions can be forced to zero by applying a small threshold. Contiguous sections of zeros can then be coded simply as the number of zeros in that section. This increases the compression ratio because we do not have to reserve a certain number of bits for each coefficient that has a value of zero. For example, suppose that we estimate the maximum number of contiguous zeros ever to appear in the wavelet transform of an image at about 32,000. Then, we can represent the number of zeros in a section of contiguous zeros using a 16-bit binary number. Contrast this with the situation where every coefficient, including those that have a value of zero after thresholding, is individually coded by using binary digits. Then, we would require at least one bit per zero-valued coefficient. Thus, a section of 1,000 contiguous zeros would require 1,000 bits to represent it as opposed to just 16 bits. This approach of coding contiguous sections of zeros is called run-length encoding.

Run-length encoding is used as part of the FBI's wavelet transform-based fingerprint image compression scheme. After wavelet transformation of the fingerprint image, the coefficients are quantized and coefficients close to zero are forced to zero. Run-length encoding followed by entropy coding, using a Huffman code, is performed. Details of the scheme can be found in various articles, for example, (12).

The advantages of wavelet transform in compression have made it the basis for the new JPEG-2000 image compression standard. Information regarding this standard may be found elsewhere in the encyclopedia.

BIBLIOGRAPHY

1. A. Grossman and J. Morlet, *SIAM J. Math. Anal.* 723–736 (1984).
2. J. Morlet, *Proc. 51st Annu. Meet. Soc. Exploration Geophys.*, Los Angeles, 1981.
3. C. W. Helstrom, *IEEE Trans. Inf. Theory* **12**, 81–82 (1966).
4. I. Daubechies, *Proc. IEEE* **84**(4), 510–513 (1996).
5. A. Haar, *Math. Annal.* **69**, 331–371 (1910).
6. I. Daubechies, *Ten Lectures on Wavelets*, SIAM, Philadelphia, 1992.
7. S. K. Mitra, *Digital Signal Processing: A Computer-Based Approach*, McGraw-Hill Irwin, Boston, 2001.
8. D. L. Donoho and M. R. Duncan, *Proc. SPIE*, **4,056**, 12–30 (2000).
9. R. M. Rao and A. S. Bopardikar, *Wavelet Transforms: Introduction to Theory and Applications*, Addison-Wesley Longman, Reading, MA, 1998.
10. N. S. Jayant and P. Noll, *Digital Coding of Waveforms: Principles and Applications to Speech and Video*, Prentice-Hall, Englewood Cliffs, NJ, 1984.
11. A. Said and W. A. Pearlman, *IEEE Trans. Circuits Syst. Video Technol.* **6**(3), 243–250 (1996).
12. C. M. Brislawn, J. N. Bradley, R. J. Onyshczak, and T. Hopper, *Proc. SPIE — Int. Soc. Opt. Eng. (USA)* **2,847**, 344–355 (1996).

WEATHER RADAR

Robert M. Rauber
University of Illinois at Urbana-Champaign
Urbana, IL

This article contains a brief overview of the history of meteorological radars, presents the operating principles of these radars, and explains important applications of radar in the atmospheric sciences. Meteorological radars transmit short pulses of electromagnetic radiation at microwave or radio frequencies and detect energy backscattered toward the radar's antenna by scattering elements in the atmosphere. Radiation emitted by radar is scattered by water and ice particles, insects, other objects in the path of the beam, and refractive index heterogeneities in air density and humidity. The returned signal is the combination of radiation backscattered toward the radar by each scattering element within the volume illuminated by a radar pulse. Meteorologists use the amplitude, phase, and polarization state of the backscattered energy to deduce the location and intensity of precipitation, the wind speed in the direction of the radar beam, and precipitation characteristics, such as rain versus hail.

HISTORICAL OVERVIEW

Radar, an acronym for radio detection and ranging, was initially developed to detect aircraft and ships remotely.

During the late 1930s, military radar applications for aircraft detection were adopted by Britain, Germany, and the United States, but these radars were limited to very low frequencies (0.2–0.4 GHz) and low power output. In 1940, the highly secret British invention of the cavity magnetron permitted radars to operate at higher frequencies (3–10 GHz) and high power output, allowing the Allies of World War II to detect aircraft and ships at long ranges.

Studies of atmospheric phenomenon by radar began almost as soon as the first radars were used. These studies were initiated because weather and atmospheric echoes represented undesirable clutter that hampered detection of military targets. Particularly problematic atmospheric echoes were caused by storms and the anomalous propagation of radar beams. Because of extreme security, studies of these phenomena went unpublished until after the war.

A key advance during World War II was the development of the theory relating the magnitude of echo intensity and attenuation to the type and size of drops and ice particles illuminated by the radar beam. Theoretical analyses, based on Mie scattering principles, predicted a host of phenomena that were subsequently observed, such as the bright band at the melting level and the high reflectivity of very large hailstones. The first weather radar equations were published in 1947. Slightly modified, these equations remain today as the foundation of radar meteorology.

The first large weather-related field project for a nonmilitary application, the "Thunderstorm Project," was organized after the war to study coastal and inland thunderstorms. Data from this and other projects stimulated interest in a national network of weather radars. Enthusiasm for a national network was spurred by efforts to estimate precipitation from radar measurements. The discovery of the "hook echo" and its association with a tornado (Fig. 1) led to widespread optimism that tornadoes may be identified by radar. Following the war, several surplus military radars were adapted for weather observation. Beginning in 1957, these were replaced by the Weather Surveillance Radar (WSR-57), which became the backbone of the U.S. National Weather Service radar network until the WSR-88D Doppler radars were installed three decades later.

Radar meteorological research in the decades between 1950 and 1970 focused primarily on studies of the physics of precipitation formation, precipitation measurement,

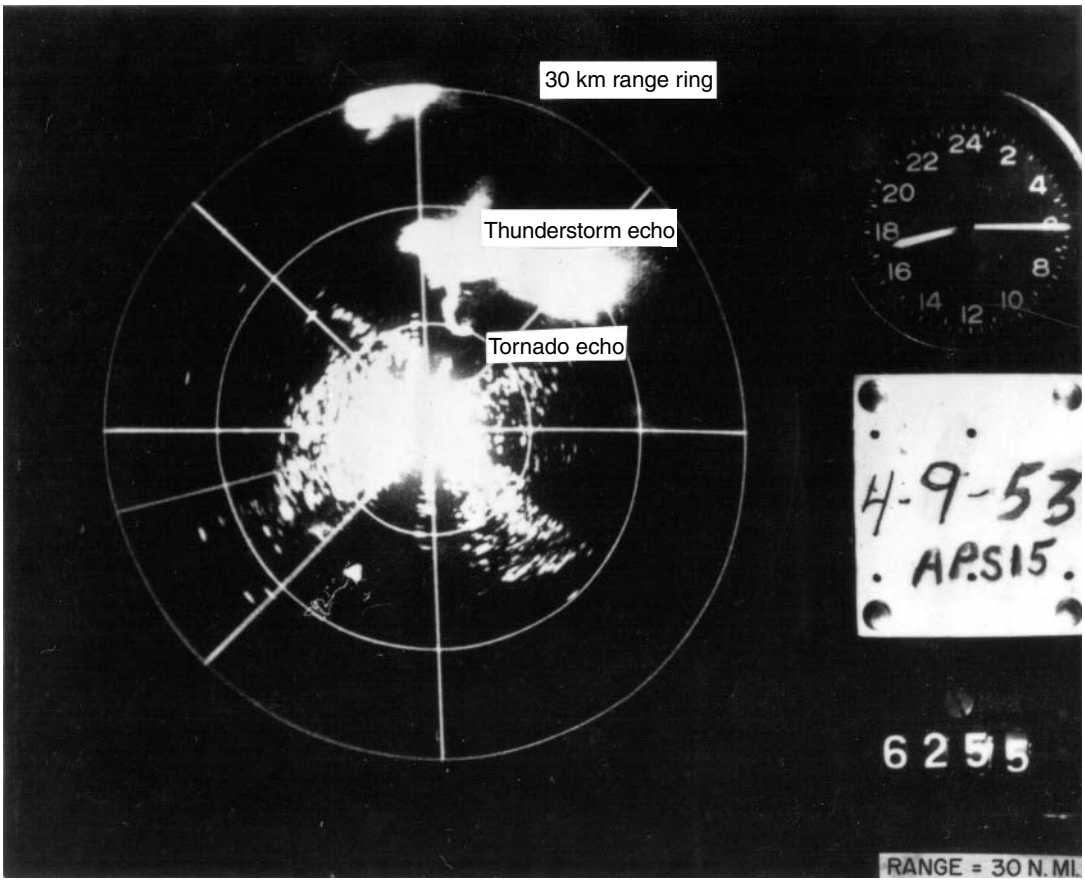


Figure 1. First published photograph of a thunderstorm hook echo observed on a radarscope. The tornado, which was located at the southern end of the hook, occurred just north of Champaign, IL, on 9 April, 1953. (From G. E. Stout and F. A. Huff, Radar records Illinois tornadogenesis. *Bulletin of the American Meteorological Society* **34**, 281–284 (1953). Courtesy of Glenn E. Stout and the American Meteorological Society.)

storm structure, and severe storm monitoring. The first coordinated research flights through large cyclonic storms were conducted in conjunction with weather radar observations. These studies led to investigations of the physics of the "bright band," a high reflectivity region in stratiform clouds associated with melting particles, "generating cells," regions near cloud tops from which streams of ice particles appeared, and other cloud physical processes. Films of radarscopes were first used to document the evolution of storms. Doppler radars, which measure the velocity of scattering elements in the direction of the beam, were first developed.

The advent of digital technology, the rapid growth in the number of scientists in the field of radar meteorology, and the availability of research radars to the general meteorological community led to dramatic advances in radar meteorological research beginning around 1970. The fundamental change that led the revolution was the proliferation of microprocessors and computers and associated advances in digital technology. A basic problem that hampers radar scientists is the large volume of information generated by radars. For example, a typical pulsed Doppler radar system samples data at rates as high as three million samples per second. This volume of data is sufficiently large that storage for later analysis, even today, is impractical — the data must be processed in real time to reduce its volume and convert it to useful forms. Beginning in the early 1970s, advances in hardware, data storage technology, digital displays, and software algorithms made it possible to collect, process, store, and view data at a rate equal to the rate of data ingest. A key advance was the development of efficient software to process the data stream from Doppler radars. This development occurred at about the same time that the hardware became available to implement it and led to rapid advances in Doppler measurements.

Doppler radars were soon developed whose antennas rotate in azimuth and elevation so that the full hemisphere around the radar could be observed. A network of Doppler radars was installed throughout the United States in the early 1990s to monitor severe weather. Other countries are also using Doppler radars now for storm monitoring. Mobile, airborne, and spaceborne meteorological research radars were developed in the 1980s and 1990s for specialized applications. Scientists currently use these radars and other types of modern meteorological radar systems to study a wide range of meteorological phenomena.

BASIC OPERATING PRINCIPLES OF RADAR

Radars transmit brief pulses of microwave energy. Each pulse lasts about 1 microsecond, and pulses are separated by a few milliseconds. The distance r to a target, determined from the time interval Δt, between transmission of the microwaves and reception of the echo, is given by

$$r = \frac{c\Delta t}{2}, \tag{1}$$

where c is the speed of light.

The pulse repetition frequency determines the maximum unambiguous range across which a radar can detect targets. After a pulse has been transmitted, the radar must wait until echoes from the most distant detectable target of interest return before transmitting the next pulse. Otherwise, echoes of the nth pulse will arrive from distant targets after the $(n+1)$-th pulse has been transmitted. The late arriving information from the nth pulse will then be interpreted as echoes of the $(n+1)$-th pulse. Echoes from the distant targets will then be folded back into the observable range and will appear as weak elongated echoes close to the radar. Echoes from distant targets that arrive after the transmission of a subsequent pulse are called *second-trip* or *ghost* echoes. The maximum unambiguous range $r_{\max}$, for a radar is given by

$$r_{\max} = \frac{c}{2F}, \tag{2}$$

where F is the pulse repetition frequency. Depending upon the application, the F chosen is generally from 400–1,500 s^{-1}, leading to $r_{\max}$ between 375 and 100 km. Although a low F is desirable for viewing targets far from the radar, there are other issues that favor a high value for F. Accurate measurement of echo intensity, for example, requires averaging information from a number of pulses from the same volume. The accuracy of the measurement is directly related to the number of samples in the average, so a high F is desired. Using Doppler measurements, F determines the range of velocities observable by the radar. A greater range of velocities can be observed by using a higher F.

The four primary parts of a typical pulsed Doppler radar, a transmitter, antenna, receiver, and display, are contained in the simplified block diagram of a Doppler radar shown in Fig. 2. The transmitter section contains a microwave tube that produces power pulses. Two kinds of transmitter tubes, *magnetrons* and *klystrons*, are in general use. The magnetron is an oscillator tube in which the frequency is determined mainly from its internal structure. The klystron, illustrated in Fig. 2, is a power amplifier. Its microwave frequency is established by mixing lower frequency signals from a stable low-power oscillator, termed the STALO, and a coherent local oscillator, termed the COHO. Microwaves are carried from the klystron to the antenna through *waveguide* plumbing that is designed to minimize energy loss.

The transmitter and receiver normally share a single antenna, a task accomplished by using a fast switch called a *duplexer* that connects the antenna alternately to the transmitter and the receiver. The size and shape of the antenna determine the shape of the microwave beam. Most meteorological radars use circular-parabolic antennas that form the beam into a narrow cone that has a typical half-power beam width between about 0.8 and 1.5°. Radar antennas typically can be rotated in both azimuth and elevation, so that the entire hemisphere around the radar can be observed. Wind profiler radars employ "phased-array" antennas, whose beam is scanned by electronic means rather than by moving the antenna. Side lobes, peaks of energy transmission that occur outside the main beam, as shown in Fig. 2, complicate the radiation pattern.

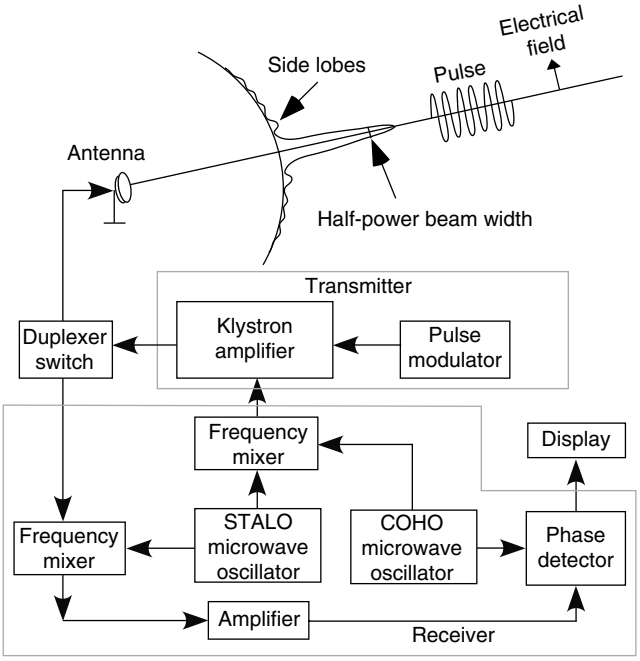


Figure 2. Simplified block diagram of a radar showing the key hardware components, the antenna radiation pattern, and a microwave frequency pulse.

Targets that lie off the beam axis may be illuminated by power transmitted into the side lobes. Echoes from these targets cannot be distinguished from echoes from targets in the main lobe. Echoes from side lobes introduce confusion and error into radar observations and are undesirable. Other undesirable echoes occur when the side lobes, or the main lobe, strikes ground targets such as trees. The echoes from these objects, called *ground clutter*, sometimes make it difficult to interpret meteorological echoes in the same vicinity.

When the microwaves strike any object in their path, a small part of the energy is reflected or scattered back toward the antenna. The antenna receives backscattered waves from all scatters in a volume illuminated by a pulse. These waves superimpose to create the received waveform, which passes along the waveguide through the duplexer into the receiver. The power collected by the antenna is small. Whereas a typical peak transmitted power might be a megawatt (10^6 watts), the typical received power might be only a nanowatt (10^{-9} watts). The receiver first amplifies the signal and then processes it to determine its amplitude, which is used to calculate the *radar reflectivity factor*. The radar reflectivity factor is proportional to the sum of the sixth power of the diameter of all of the raindrops in the radar volume and is related to the precipitation intensity. Doppler radar receivers also extract information about the phase of the returned wave, which is used to determine the *radial velocity*, the velocity of the scatterers in the direction of the radar beam. The radial velocity is related to the component of the wind in the direction of the beam and, when the beam is pointed above the horizontal, to the terminal fall velocity of the particles. Polarization diversity radars determine polarization information by comparing either the intensity or phase difference between pulses transmitted at different polarization states. This information is used to estimate the shapes and types of particles in clouds. Finally, pulse-to-pulse variations in radial velocity are used to estimate the velocity spectral width, which provides a rough measure of the intensity of turbulence.

Weather radars typically operate at microwave frequencies that range from 2.8–35 GHz (wavelength of 10.7–0.86 cm). A few research radars operate at frequencies that exceed 35 GHz, but their use is limited because of extreme attenuation of the beam in clouds. Radar wind profilers operate in ultrahigh frequency (UHF; 0.3–3.0 GHz) and very high frequency (VHF; 0.03–0.3 GHz) bands. Meteorologists typically use radar wavelength rather than frequency because wavelengths can be compared more directly to precipitation particle sizes. This comparison is important because radar-derived meteorological quantities such as the radar reflectivity factor are based on Rayleigh scattering theory, which assumes that particles are small relative to the wavelength. The factors that govern the choice of the wavelength include sensitivity, spatial resolution, the nature of the targets (e.g., thunderstorms, cirrus clouds), the effects of attenuation, as well as equipment size, weight, and cost. For shorter wavelengths, higher sensitivity can be achieved with smaller and cheaper radar systems; however, shorter wavelengths suffer severe attenuation in heavy precipitation, which limits their usefulness.

PROPAGATION OF RADAR WAVES THROUGH THE ATMOSPHERE

As electromagnetic pulses propagate outward from a radar antenna and return from a target, they pass through air that contains water vapor and may also contain water drops and ice particles. Refraction and absorption of the electromagnetic energy by vapor, water, ice, and air affect the determination of both the location and properties of meteorological targets that comprise the returned signal.

The height of a radar beam transmitted at an angle ϕ above the earth's surface depends on both the earth's curvature and the refractive properties of the earth's atmosphere. Due to the earth's curvature, a beam propagating away from a radar will progress to ever higher altitudes above the earth's surface. Refraction acts to oppose the increase in beam altitude. Electromagnetic waves propagate through a vacuum at the speed of light, $c = 3 \times 10^8\ \mathrm{m\,s^{-1}}$. When these same waves propagate through air or water drops, they no longer propagate at c, but at a slower velocity v, that is related to properties of the medium. The refractive index of air, defined as $n = c/v$, is related to atmospheric dry air density, water vapor density, and temperature. The value of n varies from approximately 1.003 at sea level to 1.000 at the top of the atmosphere, a consequence of the fact that dry air density and water vapor density decrease rapidly as height increases. In average atmospheric conditions, the

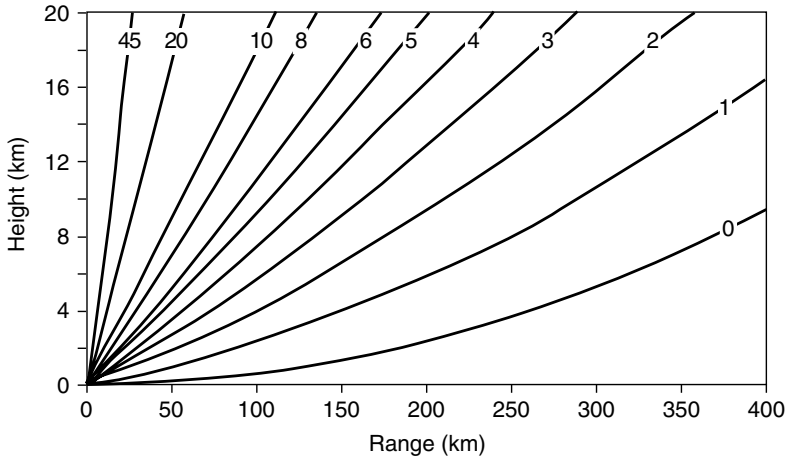


Figure 3. The height of a radar beam above the earth's surface as a function of range and elevation angle considering earth curvature and standard atmospheric refraction.

vertical gradient of n is about -4×10^{-8} m^{-1}. According to Snell's law of refraction, radar beams that pass through an atmosphere where n decreases with height will bend earthward. To calculate the height of a radar beam, radar meteorologists must consider both the earth's curvature and the vertical profile of n. Normally, the vertical profile of n is estimated from tables that are based on climatologically average conditions for the radar site. Figure 3 shows examples of beam paths for various values of ϕ. It is obvious from Fig. 3 that the earth's curvature dominates refraction under average conditions because the radar beam's altitude increases as distance from the radar increases. An important consequence of the earth's curvature is that radars cannot detect storms at long distances because the beam will pass over the distant storm tops. For example, a beam pointed at the horizon ($\phi = 0^\circ$) in Fig. 3 attains a height of 9.5 km at a 400-km range.

Any deviation of a radar beam from the standard paths shown in Fig. 3 is termed *anomalous propagation*. Severe anomalous propagation can occur in the atmosphere when n decreases very rapidly with height. Ideal conditions for severe refraction of microwaves exists when a cool moist layer of air is found underneath a warm dry layer, and temperature increases with altitude through the boundary between the layers. Under these conditions, which often occur along some coastlines (e.g., the U.S. West Coast in summer), beams transmitted at small ϕ can bend downward and strike the earth's surface. In these cases, echoes from the surface appear on the radar display and cause uncertainty in the interpretation of meteorological echoes.

Radar waves experience power loss from both energy absorption and scatter. Collectively, radar meteorologists refer to this power loss as *attenuation*. Although atmospheric gases, cloud droplets, fog droplets, and snow contribute to attenuation, the most serious attenuation is caused by raindrops and hail. Attenuation depends strongly on wavelength; shorter wavelength radars suffer the most serious attenuation. For example, a 3-cm-wavelength radar can suffer echo power losses 100 times that of a 10-cm-wavelength radar in heavy precipitation. This is a particular problem when a second storm lies along the beam path beyond a closer storm. Two-way attenuation by precipitation in the first storm may make the distant storm appear weak or even invisible on the radar display.

Were it not for attenuation, short wavelength radars would be in general use because of their greater sensitivity, superior angular resolution, small size, and low cost. Instead, radars with wavelengths shorter than 5 cm are rarely used. One exception is a class of radars called "cloud" radars, that are typically pointed vertically to study the structure of clouds that pass overhead. Aircraft meteorological radars also employ short wavelengths, despite attenuation. On aircraft, weight and space constraints limit antenna size, whereas beam width constraints required for meteorological measurements using the smaller antenna normally limit useable wavelengths to 3 cm or less. Fortunately, radars whose wavelengths are near 10 cm suffer little from attenuation in nearly all meteorological conditions. For this reason, the wavelength chosen for U.S. National Weather Service WSR-88D radars was 10 cm.

In the past, a few research radar systems have been equipped with two radars of different wavelength; two antennas point in the same direction and are mounted on the same pedestal. These dual-wavelength systems have been used for hail detection, improved rainfall measurements, and to understand cloud processes. Dual-wavelength techniques depend on the fact that energy of different wavelengths is attenuated differently as it passes through a field of precipitation particles, and is scattered in different ways by both particles and refractive index heterogeneities. Dual-wavelength radars have received less attention since the advent of polarization diversity radars, which provide better techniques for discriminating particle types and estimating rainfall.

THE WEATHER RADAR EQUATION AND THE RADAR REFLECTIVITY FACTOR

Radar is useful in meteorology because the echo power scattered back to the radar by meteorological targets such as raindrops and snowflakes is related, with some

caveats, to meteorologically significant quantities such as precipitation intensity. The basis for relating the physical characteristics of the targets to the received echo power is the weather radar equation. The radar range equation for meteorological targets, such as raindrops, is obtained by first determining the radiated power per unit area (the power flux density) incident on the target, next determining the power flux density scattered back toward the radar by the target; and then determining the amount of back-scattered power collected by the antenna. For meteorological targets, the radar range equation is given by

$$\overline{P_r} = \frac{P_t G^2 \lambda^2}{64\pi^3 r^4} V\eta, \tag{3}$$

where $\overline{P_r}$ is the average received power, P_t the transmitted power, G the antenna gain, λ the wavelength of transmitted radiation, r the range, V the scattering volume, and η the reflectivity.

The volume V illuminated by a pulse is determined from the pulse duration τ, the speed of light, c, the range r, and the angular beam width θ, which is normally taken as the angular distance in radians between the half-power points from the beam center (see Fig. 2). For antennas that have a circular beam pattern, the volume is given by

$$V = \frac{\pi c \tau \theta^2 r^2}{8}. \tag{4}$$

Typically the beam width is about 1°, and the pulse duration 1 microsecond, so at a range of 50 km, the scattering volume equals about 10^8 m^3. In moderate rain, this volume may contain more than 10^{11} raindrops. The contributions of each scattering element in the volume add in phase to create the returned signal. The returned signal fluctuates from pulse to pulse as the scattering elements move. For this reason, the returned signals from many pulses must be averaged to determine the average received power.

The radar cross section σ of a spherical water or ice particle, whose diameter D is small compared to the wavelength λ, is given by the Rayleigh scattering law

$$\sigma = \frac{\pi^5}{\lambda^4} |K|^2 D^6, \tag{5}$$

where $|K|^2$ is a dimensionless factor that depends on the dielectric properties of the particle and is approximately equal to 0.93 for water and 0.18 for ice at radar wavelengths. The radar reflectivity of clouds and precipitation is obtained by summing the cross sections of all of the particles in the scattering volume and is written as

$$\eta = \frac{\pi^5}{\lambda^4} |K|^2 Z. \tag{6}$$

The *radar reflectivity factor* Z is defined as

$$Z = \frac{\sum D^6}{V}, \tag{7}$$

where the summation is across all of the particles in the scattering volume. The quantity Z is of prime interest to meteorologists because it relates the diameter of the targets (e.g., raindrops), and therefore the raindrop size distribution, to the power received at the radar. The radar equation for meteorological targets is obtained by combining Eqs. (3),(4), and (6) and solving for Z to obtain

$$Z = \left[\frac{512(2\ln 2)}{\pi^3 c}\right]\left[\frac{\lambda^2}{P_t \tau G^2 \theta^2}\right]\left[\frac{r^2 \overline{P_r}}{K^2}\right]. \tag{8}$$

The term $(2\ln 2)$ in the numerator was added to account for the fact that most antenna systems are designed for tapered, rather than uniform illumination, to reduce the effects of side lobes.

Equation (8) is valid provided that the raindrops and ice particles illuminated by the radar beam satisfy the Rayleigh criterion. For long-wavelength radars (e.g., 10 cm), the Rayleigh criterion holds for all particles except for large hail. However, for shorter wavelength radars, the Rayleigh criterion is sometimes violated. Equation (8) assumes a single value of K. The true value of Z will not be measured for radar volumes that contain water and ice particles or in volumes where water is assumed to exist, but ice is actually present. Equation (8) was derived under the assumption that attenuation can be neglected. As pointed out earlier, this assumption is reasonable only for longer wavelength radars. When one or more of the assumptions used to derive Eq. (8) are invalid, the measured quantity is often termed the *equivalent radar reflectivity factor* (Z_e). Most problems that have invalid assumptions are minimized by selecting radars with long (e.g., 10 cm) wavelengths.

It is customary to use m^3 as the unit for volume and to measure particle diameters in millimeters, so that Z has conventional units of mm^6/m^3. Typical values of the radar reflectivity factor range from 10^{-5} mm^6/m^3 to 10 mm^6/m^3 in nonprecipitating clouds, 10 to 10^6 mm^6/m^3 in rain, and as high as 10^7 mm^6/m^3 in large hail. Because of the sixth-power weighting on diameter in Eq. (5), raindrops dominate the returned signal in a mixture of rain and cloud droplets. Because Z varies over orders of magnitude, a logarithmic scale, defined as

$$dBZ = 10\log_{10}\left(\frac{Z}{1\ \text{mm}^6/\text{m}^3}\right), \tag{9}$$

is used to display the radar reflectivity factor. Radar images of weather systems commonly seen in the media (e.g., Fig. 4) show the radar reflectivity factor in logarithmic units. Images of the radar reflectivity factor overlain with regional maps permit meteorologists to determine the location and intensity of precipitation. Meteorologists often interchangeably use the terms "radar reflectivity factor" and "radar reflectivity," although radar experts reserved the term "reflectivity" for η .

Radar data are typically collected on a cone formed as the beam is swept through 360° of azimuth at a constant angle of elevation. A series of cones taken at several angles of elevation constitutes a *radar volume*. Images of the radar reflectivity factor from individual radars are typically projected from the conical surface onto a map-like format called the "plan-position indicator," or PPI display,

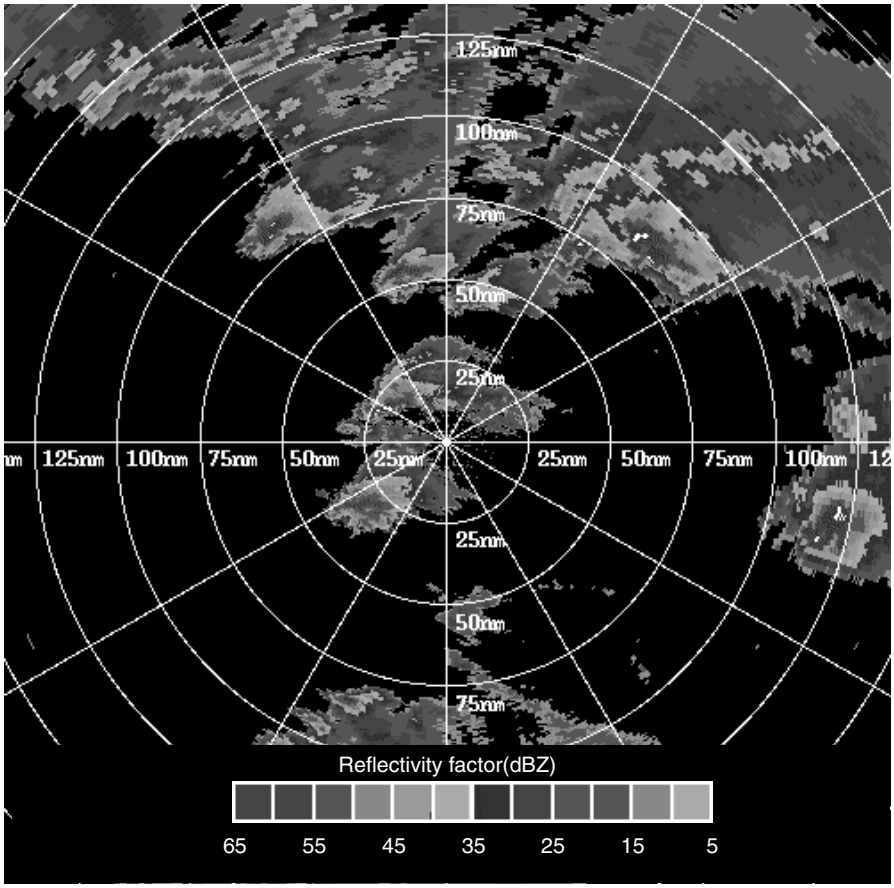


Figure 4. Plan-position indicator scan of the radar reflectivity factor at 0.5° elevation from the Lincoln, IL, radar on 19 April 1996 showing tornadic thunderstorms moving across the state. The red areas denote the heavy rainfall and hail, and green and blue colors denote lighter precipitation. See color insert.

where the radar is at the center, north at the top, and east at the right. Because of the earth's curvature, atmospheric refraction, and beam tilt, distant radar echoes on a PPI display are at higher altitudes than those near the radar.

Sometimes, a radar beam is swept between the horizon and the zenith at a constant azimuth. In this case, data are plotted in a "range-height indicator" or RHI display, which allows the meteorologist to view a vertical cross section through a storm.

Meteorologists broadly characterize precipitation as convective when it originates from storms such as thunderstorms that consist of towering cumulus clouds that have large vertical motions. Convective storms produce locally heavy rain and are characterized by high values of the reflectivity factor. Convective storms typically appear on PPI displays as small cores of very high reflectivity. Weaker reflectivity values typically extend downwind of the convective cores as precipitation particles are swept downstream by the wind. Figure 4 shows an example of the radar reflectivity factor measured by the Lincoln, IL, Doppler radar during an outbreak of tornadic convective storms on 19 April 1996. The red areas in the image, which denote heavy rain, correspond to the convective area of the storm, and the green areas northeast of the convective regions denote the lighter rain downstream of the convective regions.

Meteorologists characterize precipitation as stratiform when it originates from clouds that have a layered structure. These clouds have weak vertical motions, lighter precipitation, and generally lower values of the radar reflectivity factor. Echoes from stratiform precipitation in a PPI display appear in Fig. 5, a radar image of a snowstorm over Michigan on 8 January, 1998. Stratiform echoes are generally widespread and weak, although often they do exhibit organization and typically have narrow bands of heavier precipitation embedded in the weaker echo.

On RHI scans, convective storms appear as cores of high reflectivity that extend from the surface upward high into the storm. Divergent winds at the top of the storm carry precipitation outward away from the convective region. In strong squall lines, this precipitation can extend 50–100 km behind the convective region, creating a widespread stratiform cloud. These features are all evident in Fig. 6, an RHI scan through a squall line that occurred in Kansas and Oklahoma on 11 June 1985. The radar bright band is a common characteristic often observed in RHI displays of stratiform precipitation. The bright band is a local region of high reflectivity at the melting level (see BB in Fig. 6). As ice particles fall from aloft and approach the melting level, they often aggregate into snowflakes that can reach sizes of a centimeter or more. When these particles first begin to melt, they develop a water coating on the ice surfaces. The reflectivity increases dramatically due to both the larger particle sizes and the change in the dielectric properties of the melting snow (K in Eq. 6). Because of the dependence of Z on the sixth power of the particle diameters, these

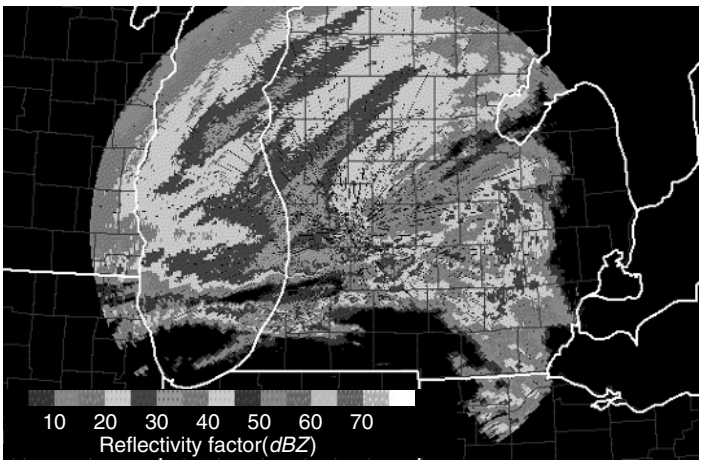


Figure 5. Plan-position indicator scan of the radar reflectivity factor at 0.5° elevation from the Grand Rapids, MI, radar on 8 January, 1998 showing a snowstorm. Bands of heavier snowfall can be seen embedded in the generally weaker echoes. See color insert.

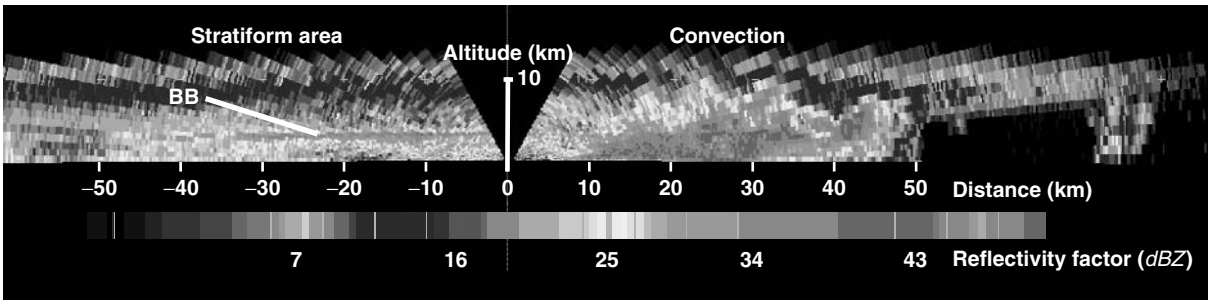


Figure 6. Range-height indicator scan of the radar reflectivity factor taken through a mature squall line on 11 June, 1985 in Texas. The bright band is denoted by the symbol BB in the trailing stratiform region of the storm on the left side of the figure. (Courtesy of Michael Biggerstaff, Texas A & M University, with changes.) See color insert.

large water-coated snowflakes have very high reflectivity. Snowflakes collapse into raindrops, reducing in size as they melt. In addition, their fall speed increases by a factor of 6, so they quickly fall away from the melting layer and reduce particle concentrations. As a result, the reflectivity reduces below the melting level. Therefore, the band of highest reflectivity occurs locally at the melting level (see Fig. 6). The bright band can also appear in PPI displays as a ring of high reflectivity at the range where the beam intersects the melting level (see Fig. 7). Another common feature of stratiform clouds is precipitation streamers, regions of higher reflectivity which begin at cloud top and descend in the cloud. Figure 8 shows an example of a precipitation streamer observed in a snowstorm in Michigan by the NCAR ELDORA airborne radar. Streamers occur when ice particles form in local regions near a cloud top, descend, grow as they fall through the cloud, and are blown downstream by midlevel winds.

The PPI and RHI displays are the most common displays used in meteorology. In research applications, radar data are often interpolated to a constant altitude and displayed in a horizontal cross section to visualize a storm's structure better at a specific height. Similarly, interpolated data can be used to construct vertical cross sections, which appear much like the RHI display. Composite radar images are also constructed by combining reflectivity data from several radars. These composites are typically projections of data in PPI format onto a single larger map. For this reason, there is ambiguity concerning the altitude of the echoes on composite images.

The radar reflectivity factor Z is a general indicator of precipitation intensity. Unfortunately, an exact relationship between Z and the precipitation rate R does not exist. Research has shown that Z and R are approximately related by Eq. (10):

$$Z = aR^b, \tag{10}$$

where the coefficient a and the exponent b take different values that depend on the precipitation type. For example, in widespread stratiform rain, a is about 200 and b is 1.6 if R is measured in mm/h and Z is in mm^6/m^3. In general, radar estimates of the short-term precipitation rate at a point can deviate by more than a factor of 2 from surface rain gauge measurements. These differences are due to uncertainties in the values of a and b, radar calibration uncertainties, and other sources of error. Some of these errors are random, so radar estimates of total accumulated rainfall over larger areas and longer times tend to be more accurate. Modern radars use algorithms to display total accumulated rainfall by integrating the rainfall rate, determined from Eq. (10), across selected time periods. For example, Fig. 9 shows the accumulated rainfall during the passage of a weather system over eastern Iowa and Illinois. Radar estimated rainfall exceeded one inch locally near

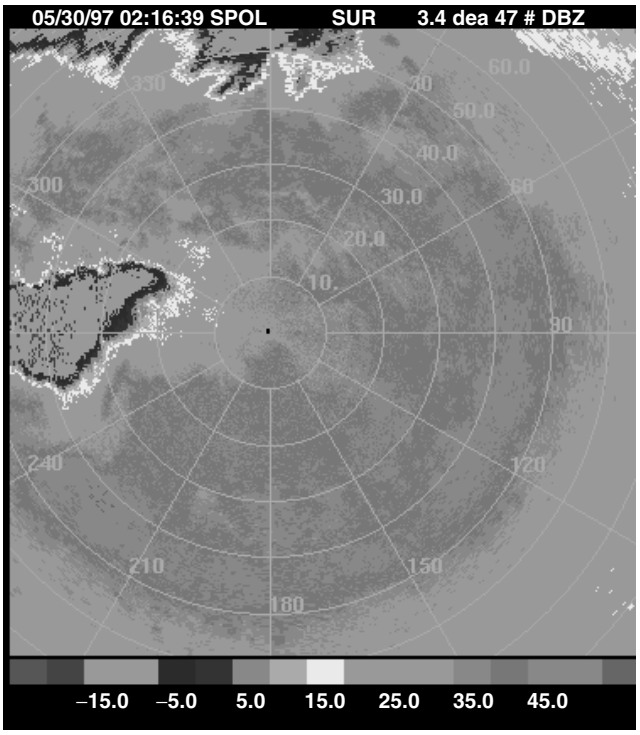


Figure 7. Plan-position indicator scan of the radar reflectivity factor through a stratiform cloud. The bright band appears as a ring of strong (red) echoes between the 40 and 50 km range. Melting snowflakes, which have a high radar reflectivity, cause the bright band. (Courtesy of R. Rilling, National Center for Atmospheric Research). See color insert.

Davenport, IA, and Champaign, IL, and other locations received smaller amounts.

DOPPLER RADARS

Doppler Measurements

A Doppler frequency shift occurs in echoes from targets that move along the radar beam. The magnitude and direction of the frequency shift provides information about the targets' motion along the beam, toward or away from the radar. In meteorological applications, this measurement, termed the *radial velocity*, is used primarily to estimate winds. The targets' total motion consists of four components, two horizontal components of air motion, vertical air motion, and the target mean fall velocity in still air. Because only one component is observed, additional processing must be done or assumptions made to extract information about the total wind field.

To obtain the Doppler information, the phase information in the echo must be retained. Phase, rather than frequency, is used in Doppler signal processing because of the timescales of the measurements. The period of the Doppler frequency is typically between about 0.1 and 1.0 millisecond. This is much longer than the pulse duration, which is typically about 1 microsecond, so only a fraction of a cycle occurs within the pulse period. Consequently, for meteorological targets, one cannot measure the Doppler frequency by just one transmitted pulse.

The Doppler frequency is estimated instead by measuring the phase ϕ of the echo, at a specific range for each pulse in a train of pulses. Each adjacent pair of sampled phase values of the returned wave, for example, ϕ_1 and ϕ_2, ϕ_2 and ϕ_3, etc., can be used to obtain an estimate of the Doppler frequency f_d from

$$f_d = \frac{(\phi_{n+1} - \phi_n)F}{2\pi}. \tag{11}$$

Conceptually, the average value of f_d from a number of pulse pairs determines the final value of the Doppler frequency. In actual signal processing, each phase measurement must be calculated from the returned signal by performing an arctangent calculation. This type of calculation is computationally demanding for radars that collect millions of echo samples each second. In practice, a more computationally efficient technique called the Pulse Pair Processor, which depends on the signal autocorrelation function, is normally used to extract the Doppler frequency.

Unfortunately, inversion of the sampled phase values to determine the Doppler frequency and the target radial velocity is not unique. As a result, velocity ambiguity problems exist for Doppler radar systems. In the inversion process, Doppler radars normally use the lowest frequency that fits the observed phase samples to determine the target radial velocity. Using this approach, the maximum unambiguous observable radial velocity $v_{r,max}$ is given by

$$|v_{r,max}| = \frac{\lambda F}{4}. \tag{12}$$

Doppler frequencies that correspond to velocities higher than $|v_{r,max}|$ are aliased, or folded, back into the observable range. For example, if $|v_{r,max}| = 20\ m\,s^{-1}$, then the range of observable velocities will be $-20\ m\,s^{-1} \le v_{r,max} \le 20\ m\,s^{-1}$,

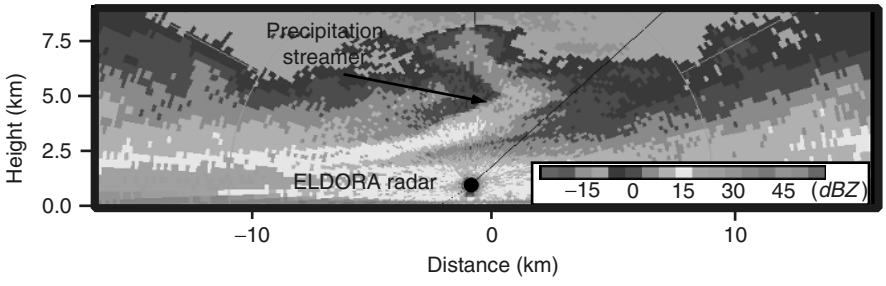


Figure 8. Range-height indicator scan of the radar reflectivity factor through a snowstorm on 21 January, 1998. The data, taken with the ELDORA radar on the National Center for Atmospheric Research Electra aircraft, show a precipitation streamer, a region of heavier snow that develops near the cloud top and is carried downstream by stronger midlevel winds. See color insert.

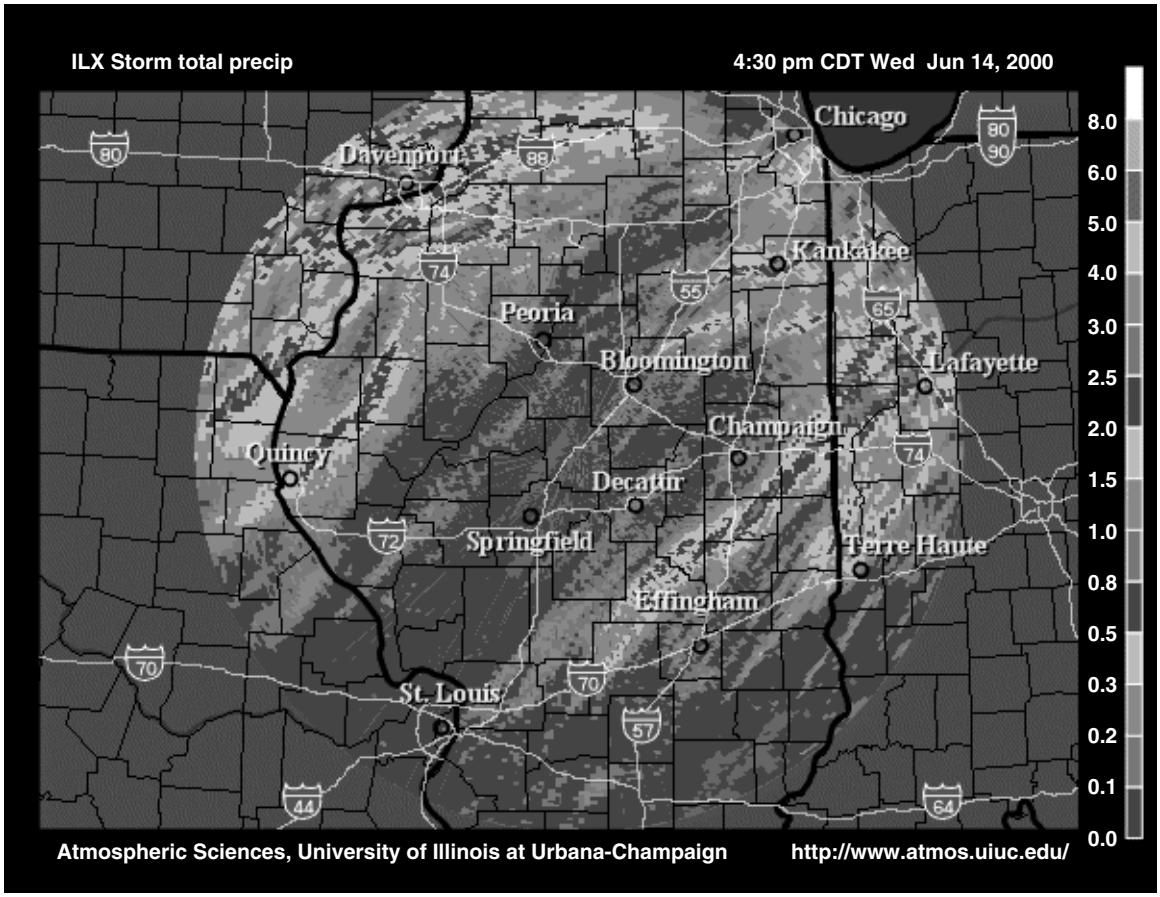


Figure 9. Total precipitation through 4:30 P.M. local time on 14 June, 2000, as measured by the National Weather Service radar at Lincoln, IL, during the passage of a storm system. See color insert.

and a true velocity of 21 $\mathrm{m\,s^{-1}}$ would be recorded by the radar system as $-19\ \mathrm{m\,s^{-1}}$. Folding of Doppler velocities is related to a sampling theorem known as the Nyquist criterion, which requires that at least two samples of a sinusoidal signal per cycle be available to determine the frequency of the signal. In a pulse Doppler radar, f_d is the signal frequency, and F is the sampling rate. The velocity $v_{r,\max}$, corresponding to the maximum unambiguous Doppler frequency, is commonly called the Nyquist velocity. The Nyquist velocity depends on wavelength (Eq. 12), so that long-wavelength radars (e.g., 10 cm) have a larger range of observable velocities for the same pulse repetition frequency. For example, at $F = 1{,}000\ \mathrm{s^{-1}}$, the Nyquist velocity will be $25\ \mathrm{m\,s^{-1}}$ at $\lambda = 10$ cm, but only $7.5\ \mathrm{m\,s^{-1}}$ at $\lambda = 3$ cm. Velocities $>25\ \mathrm{m\,s^{-1}}$ commonly occur above the earth's surface, so velocities recorded by the shorter wavelength radar could potentially be folded multiple times. This is another reason that shorter wavelength radars are rarely used for Doppler measurements. From Eq. (12), it is obvious that long wavelength and high F are preferable for limiting Doppler velocity ambiguity.

The choice of a high value of F to mitigate Doppler velocity ambiguity is directly counter to the need for a low value of F to mitigate range ambiguity [Eq. (2)]. Solving for F in Eq. (12) and substituting the result for F in Eq. (2), we find that

$$r_{\max} v_{\max} = \frac{c\lambda}{8}. \tag{13}$$

This equation, which shows that the maximum unambiguous range and radial velocity are inversely related, is called the *Doppler dilemma* because a choice good for one parameter will be a poor choice for the other. Figure 10 shows the limits imposed by the Doppler dilemma for several commonly used wavelengths.

There are two ways in common use to avoid the restrictions imposed on range and radial velocity measurements by the Doppler dilemma. The first, which is implemented in the U.S. National Weather Service radars, involves transmitting a series of pulses at a small F, followed by another series at large F. The first set is used to measure the radar reflectivity factor out to longer range, whereas the second set is used to measure radial velocities across a wider Nyquist interval, but at a shorter range. An advantage of this approach is that erroneous velocities introduced from range-folded echoes (i.e., from targets beyond the maximum unambiguous range) during the large F pulse sequence can be identified and removed. This is accomplished by comparing the echoes in the small and large F sequences and deleting all data that appears

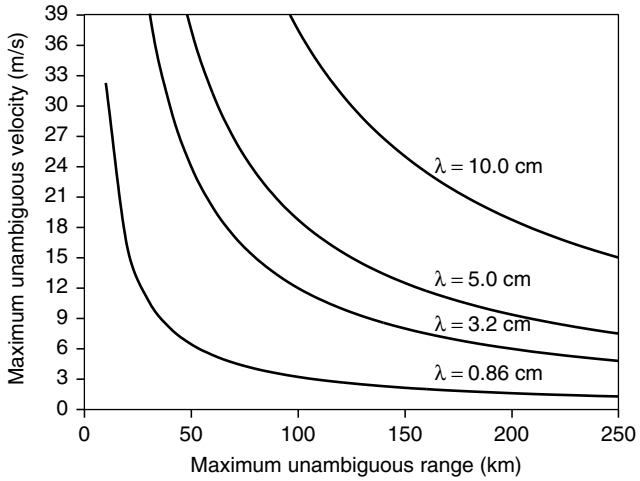


Figure 10. Relationship between the maximum unambiguous range and velocity for Doppler radars of different wavelength (λ).

only in the high F sequence. This approach has the disadvantage of increasing the dwell time (i.e., slowing the antenna rotational rate) because a sufficiently large number of samples is required to determine both the radar reflectivity factor and the radial velocity accurately. A second approach involves transmitting a series of pulses, each at a slightly different value of F, in an alternating sequence. From these measurements, it is possible to calculate the number of times that the velocity measurements have been folded. Algorithms can then unfold and correctly record and display the true radial velocities.

The Doppler Spectrum

The discussion of Doppler radar to this point has not considered the fact that meteorological targets contain many individual scatterers that have a range of radial speeds. The variation of radial speeds is due to a number of factors, including wind shear within the pulse volume (typically associated with winds increasing with altitude), turbulence, and differences in terminal velocities of the large and small raindrops and ice particles. The echoes from meteorological targets contain a spectrum of Doppler frequencies that superimpose to create the received waveform. The way that these waves superimpose changes from pulse to pulse, because, in the intervening time, the particles in the pulse volume change position relative to one another. This can best be understood by considering a volume containing only two raindrops of the same size that lie on the beam axis. Consider that these raindrops are illuminated by a microwave beam whose electrical field oscillates in the form of a sine wave. If the raindrops are separated by an odd number of half-wavelengths, then the round-trip distance of backscattered waves from the two drops will be an even number of wavelengths. The returned waves, when superimposed, will be in phase and increase the amplitude of the signal. On the other hand, consider a case where the two drops are separated by an odd number of quarter-wavelengths. In this case, the superimposed returned waves will be 180° out of phase and will destructively interfere, eliminating the signal. In the real atmosphere, where billions of drops and ice particles are contained in a single pulse volume, the amplitude of the signal and the Doppler frequency (and therefore the radial velocity) vary from pulse to pulse due to the superimposition of the backscattered waves from each of the particles as they change position relative to each other. The returned signal from each pulse will have a different amplitude and frequency (and radial velocity) which depend on the relative size and position of each of the particles that backscatters the waves. The distribution of velocities obtained from a large number of pulses constitutes the Doppler power spectrum. Figure 11, a Doppler power spectrum, shows the power S returned in each velocity interval Δv, across the Nyquist interval $\pm v_{r,max}$. The spectrum essentially represents the reflectivity-weighted distribution of particle radial speeds; more echo power appears at those velocities (or frequencies) whose particle reflectivities are greater. Some power is present at all velocities due to noise generated from the electronic components of the radar, the sun, the cosmic background, and other sources. The meteorological signal consists of the peaked part of the spectrum.

Full Doppler spectra have been recorded for limited purposes, such as determining the spectrum of fall velocities of precipitation by using a vertically pointing radar or estimating the maximum wind speed in a tornado. However, recording the full Doppler spectrum routinely requires such enormous data storage capacity that, even today, it is impractical. In nearly all applications, the full Doppler spectrum is not recorded. More important to meteorologists are the moments of the spectrum, given by

$$\overline{P_r} = \sum_{-v_{r,max}}^{v_{r,max}} S(v_r)\Delta v, \tag{14}$$

$$\overline{v_r} = \frac{\sum_{-v_{r,max}}^{v_{r,max}} v_r S(v_r)\Delta v}{\overline{P_r}}, \tag{15}$$

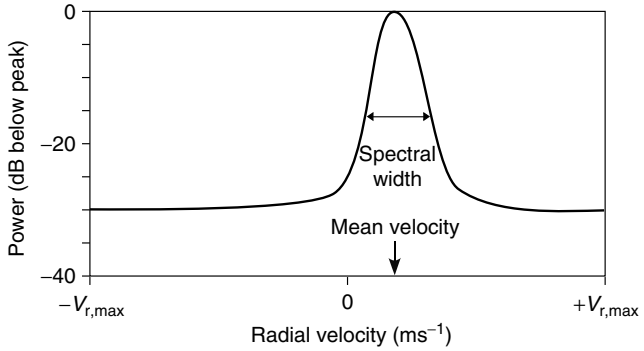


Figure 11. An idealized Doppler velocity spectrum. The average returned power is determined from the area under the curve, the mean radial velocity as the reflectivity-weighted average of the velocity in each spectral interval (typically the peak in the curve), and the spectral width as the standard deviation normalized by the mean power [see Eqs. (14)–(16)].

$$\sigma_v = \frac{\sum_{-v_{r,max}}^{v_{r,max}} (v_r - \overline{v_r})^2 S(v_r) \Delta v}{\overline{P_r}}. \quad (16)$$

The quantity $\overline{P_r}$, the area under the curve in Fig. 11, is the averaged returned power, from which the radar reflectivity factor can be determined by using Eq. (8). The mean radial velocity $\overline{v_r}$, typically the velocity near the peak in the curve, represents the average motion of the precipitating particles along the radar beam. The spread in velocities is represented by the spectral width σ_v. The spectral width gives a rough estimate of the turbulence within the pulse volume. Processing techniques used in Doppler radars extract these parameters, which are subsequently displayed and recorded for future use.

Doppler Radial Velocity Patterns in PPI Displays

Doppler radial velocity patterns appearing in radar displays are complicated by the fact that a radar pulse moves higher above the earth's surface, as it recedes from the radar. Because of this geometry, radar returns originating from targets near the radar represent the low-level wind field, and returns from distant targets represent winds at higher levels. In a PPI radar display, the distance away from the radar at the center of the display represents both a change in horizontal distance and a change in vertical distance. To determine the wind field at a particular elevation above the radar, radar meteorologists must examine the radial velocities on a ring at a fixed distance from the radar. The exact elevation represented by a particular ring depends on the angle of elevation of the radar beam.

Figure 12 shows two examples illustrating the relationship between radial velocity patterns observed on radar images and corresponding atmospheric wind profiles. The examples in Fig. 12 are simulated, rather than real images. Doppler velocity patterns (right) correspond to vertical wind profiles (left), where the wind barbs indicate wind speed and direction from the ground up to 24,000 feet (7,315 m). Each tail on a wind barb represents 10 knots (5 m s^{-1}). The direction in which the barb is pointing represents the wind direction. For example, a wind from the south at 30 knots would be represented by an upward pointing barb that has three tails, and a 20-knot wind from the east would be represented by a left pointing barb that has two tails. Negative Doppler velocities (blue-green) are toward the radar and positive (yellow–red) are away. The radar location is at the center of the display.

In the top example in Fig. 12, the wind speed increases from 20 to 40 knots (10 to 20 m s^{-1}) between zero and 12,000 feet (3,657 m) and then decreases again to 20 knots at 24,000 feet (7,315 m). The wind direction is constant. The radar beam intersects the 12,000-foot level along a ring halfway across the radar display, where the maximum inbound and outbound velocities occur. The bottom panels of Fig. 12 show a case where the wind speed is constant at 40 knots, but the wind direction varies from southerly to westerly between the ground and 24,000 feet. The innermost rings in the radar display show blue to the south and orange to the north, representing a southerly wind. The outermost rings show blue to the west and orange to the east, representing westerly winds. Intermediate rings show a progressive change from southerly to westerly as one moves outward from the center of the display. In real applications, wind speed and direction vary with height and typically vary across the radar viewing area at any given height. The radial velocity patterns are typically more complicated than the simple patterns illustrated in Fig. 12.

Of particular importance to radar meteorologists are radial velocity signatures of tornadoes. When thunderstorms move across the radar viewing area and tornadoes are possible, the average motion of a storm is determined from animations of the radar reflectivity factor or by other means and is subtracted from the measured radial velocities to obtain the *storm-relative radial velocity*. Images of the storm-relative radial velocity are particularly useful in identifying rotation and strong winds that may indicate severe conditions. Tornadoes are typically less than 1 kilometer wide. When a tornado is present, it is usually small enough that it fits within one or two beam widths. Depending upon the geometry of the beam, the distance of the tornado from the radar, and the location of the beam relative to the tornado, the strong winds of the tornado will typically occupy one or two pixels in a display. Adjacent pixels will have sharply different storm-relative velocities, typically one strong inbound and one strong outbound. Figure 13b shows a small portion of a radar screen located north of a radar (see location in Fig. 13a). Winds in this region are rotating (see Fig. 13c), and the strongest rotation is located close to the center of rotation, as would occur in a tornado. The radial velocity pattern in Fig. 13b is characteristic of a tornado vortex signature. Often, the winds will be so strong in a tornado that the velocities observed by the radar will be folded in the pixel that contains the tornado. Tornado vortex signatures take on slightly different characteristics depending on the position of individual radar beams relative to the tornado and whether or not the velocities are folded.

Single Doppler Recovery of Wind Profiles

A single Doppler radar provides measurements of the component of target motion along the beam path. At low elevations, the radial velocity is essentially the same as the radial component of the horizontal wind. At high elevation angles, the radial velocity also contains information about the targets' fall velocity and vertical air motion. Meteorologists want information about the total wind, not just the radial velocity. The simplest method for recovering a vertical profile of the horizontal wind above a radar is a technique called the *velocity-azimuth display (VAD)*, initially named because the wind was estimated from displays of the radial velocity versus azimuth angle at a specific distance from the radar, as the radar scanned through 360° of azimuth.

In the VAD technique, the radar antenna is rotated through 360° at a fixed angle of elevation. At a fixed value of range (and elevation above the radar), the sampled volumes lie on a circle centered on the radar.

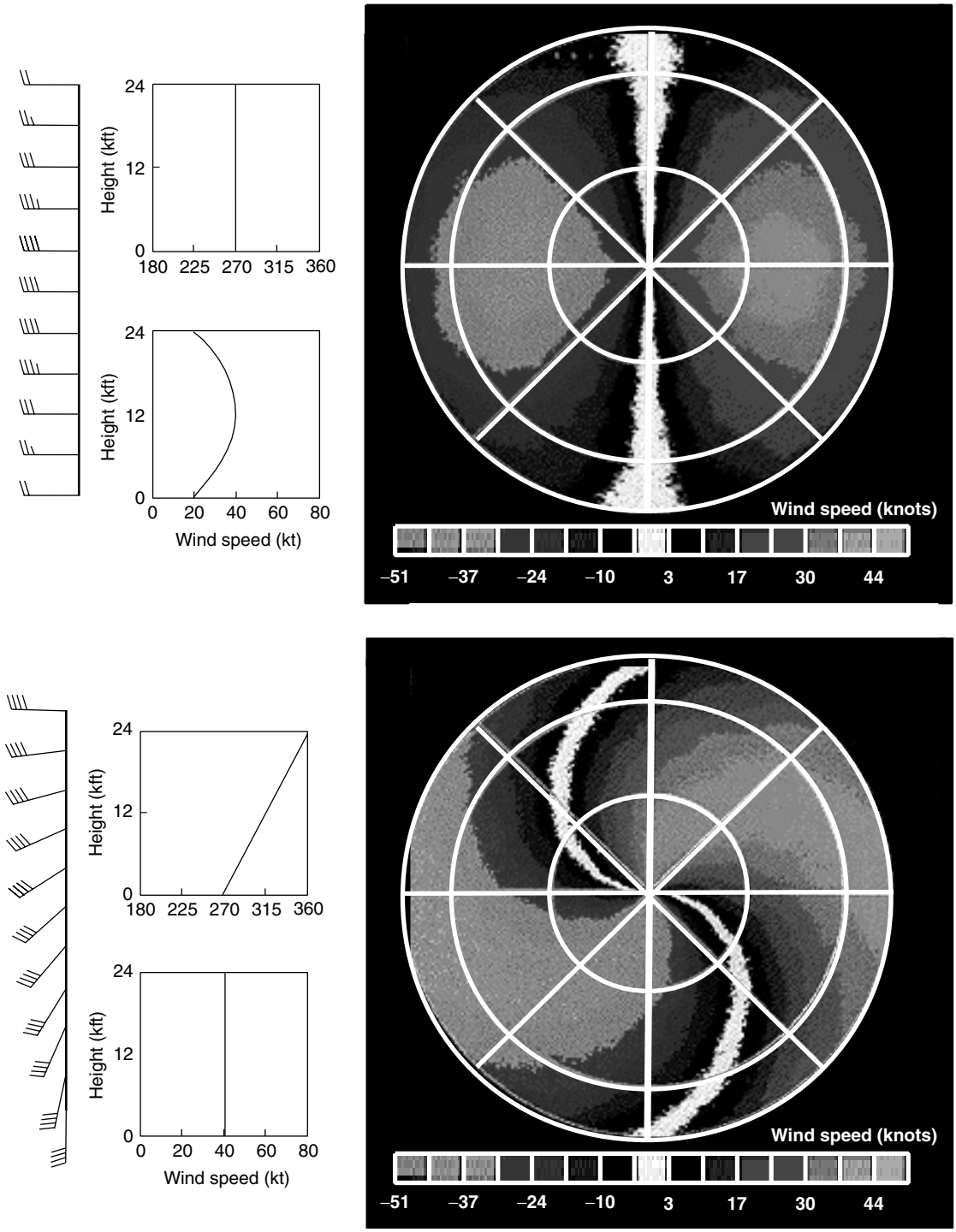


Figure 12. (Top) Doppler radial velocity pattern (right) corresponding to a vertical wind profile (left) where wind direction is constant and wind speed is 20 knots at the ground and at 24,000 feet altitude and 40 knots at 12,000 feet. Negative Doppler velocities (blues) are toward the radar, which is located at the center of the display. (Bottom) Same as top, except the wind speed is constant and the wind direction varies from southerly at the ground to westerly at 24,000 feet. (Courtesy of R. A. Brown and V. T. Woods, National Oceanic and Atmospheric Administration, with changes). See color insert.

The circles become progressively larger and higher in altitude as range increases, as shown in Fig. 14. The standard convention used for Doppler radars is that approaching radial velocities are negative and receding velocities are positive. To understand how VAD works, assume initially that the wind is uniform at 10 $m\,s^{-1}$ from the northwest across a circle scanned by the radar and that the particles illuminated by the radar along this

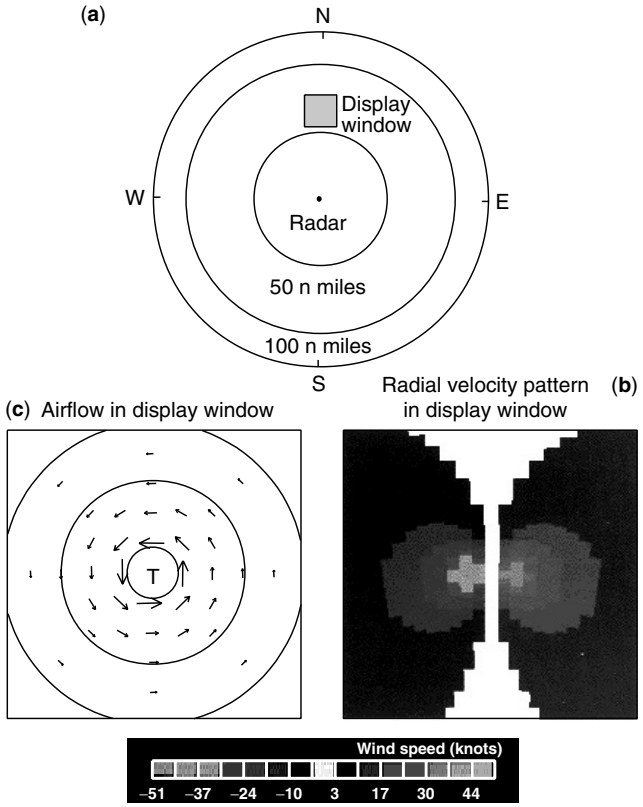


Figure 13. (**a**) Location of the 27×27 nautical mile radial velocity display window in bottom figures. The window is located 65 nautical miles north of the radar. (**b**) Radial velocity pattern corresponding to a tornado vortex signature (peak velocity = 60 kt, core radius = 0.5 nautical miles). One of the beams is centered on the circulation center. (**c**) Wind circulation corresponding to radial velocity pattern. Arrow length is proportional to wind speed, and the curved lines represent the overall wind pattern. (Courtesy of R. A. Brown and V. T. Woods, National Oceanic and Atmospheric Administration, with changes.) See color insert.

circle are falling at 5 m s^{-1}, as shown in Fig. 14a. Under these conditions, the radial velocity reaches its maximum negative value when the antenna is pointed directly into the wind and reaches its maximum positive value when pointed directly downwind. The radial velocity is negative when the antenna is pointed normal to the wind direction because of the particle fall velocity. When plotted as a function of azimuthal angle, the radial velocity traces out a sine wave (Fig. 14b), the amplitude of the sine wave is a measure of the wind velocity, the phase shift in azimuth from 0° is a measure of the wind direction, and the displacement of the center axis of the sine wave from 0 m s^{-1} is a measure of the vertical motion of the particles (Fig. 14b–d).

In reality, the flow within a circle may not be uniform. However, with some assumptions, the winds and properties of the flow related to the nonuniformity, such as air divergence and deformation, can be estimated by mathematically determining the value of the fundamental harmonics from the plot of radial velocity versus azimuthal angle. These properties are used in research applications to obtain additional information about atmospheric properties, such as vertical air motion.

National Weather Service radars used for weather monitoring and forecasting in the United States use the VAD scanning technique routinely to obtain vertical wind profiles every 6 minutes. Figure 15 shows an example of VAD-determined wind profiles from the Lincoln, IL, radar during a tornado outbreak on 19 April 1996. The data, which is typical of an environment supporting tornadic thunderstorms, shows 30-knot southerly low-level winds just above the radar, veering to westerly at 90 knots at 30,000 feet. Profiles such as those in Fig. 15 provide high-resolution data that also allow meteorologists to identify the precise position of fronts, wind shear zones that may be associated with turbulence, jet streams, and other phenomena. The VAD technique works best when echoes completely surround the radar, when the sources of the echoes are clouds that are stratiform rather than convective, and when the echoes are deep. In clear air, winds can be recovered only at lower elevations where the radar may receive echoes from scatterers such as insects and refractive index heterogeneities.

Single Doppler Recovery of Wind Fields

The VAD technique permits meteorologists to obtain only a vertical profile of the winds above the radar. In research, scientists are often interested in obtaining estimates of the total three-dimensional wind fields within a storm. Normally, such wind fields can be obtained only by using two or more Doppler radars that simultaneously view a storm from different directions. However, under certain conditions, it is possible to retrieve estimates of horizontal wind fields in a storm from a single Doppler radar. In the last two decades, a number of methodologies have been developed to retrieve single Doppler wind estimates; two of them are briefly described here. The first method, called *objective tracking of radar echo with correlations*, or TREC, employs the technique of pattern recognition by cross-correlating arrays of radar reflectivities measured several minutes apart to determine the translational motion of the echoes. One advantage of the TREC method is that it does not rely on Doppler measurements and therefore can be used by radars that do not have Doppler capability. Typically, fields of radar reflectivity are subdivided into arrays of dimensions from 3–7 km, which for modern radars, consists of 100–500 data points. The wind vector determined from each array is combined with those from other arrays to provide images of the horizontal winds across the radar scan which can be superimposed on the radar reflectivity. Figure 16 shows an example of TREC-derived winds from the Charleston, SC, WSR-57 radar, a non-Doppler radar, during the landfall of Hurricane Hugo in 1989. The TREC method in this case captured the salient features of the hurricane circulation, including the cyclonic circulation about the eye, the strongest winds near the eye wall (seen in the superimposed reflectivity field), and a decrease in the magnitude of the winds with distance outside the eyewall. The strongest winds detected, 55–60 m s^{-1}, were consistent with surface wind speeds observed by other instruments.

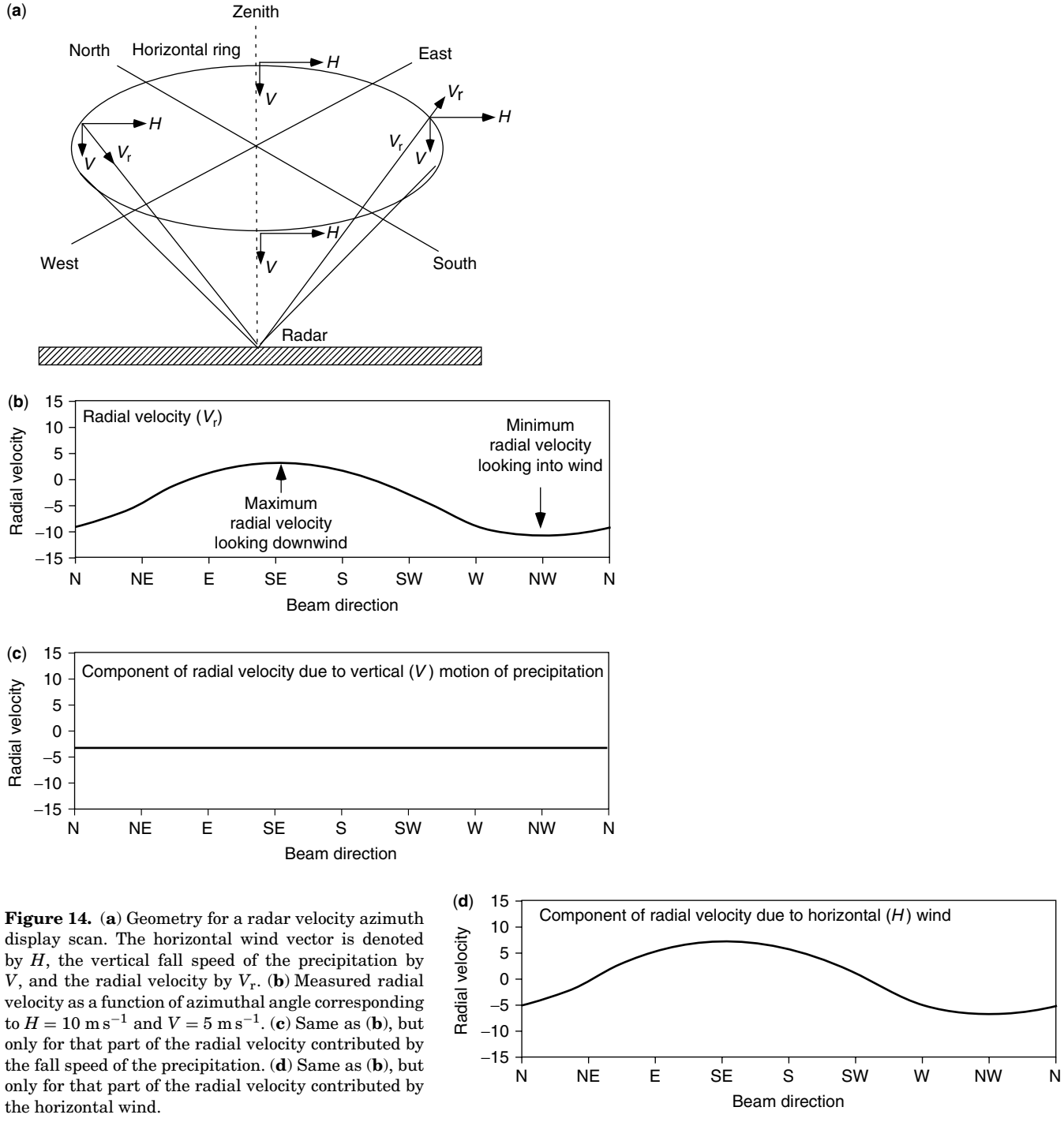


Figure 14. (**a**) Geometry for a radar velocity azimuth display scan. The horizontal wind vector is denoted by H, the vertical fall speed of the precipitation by V, and the radial velocity by V_r. (**b**) Measured radial velocity as a function of azimuthal angle corresponding to $H = 10\ \mathrm{m\,s^{-1}}$ and $V = 5\ \mathrm{m\,s^{-1}}$. (**c**) Same as (**b**), but only for that part of the radial velocity contributed by the fall speed of the precipitation. (**d**) Same as (**b**), but only for that part of the radial velocity contributed by the horizontal wind.

A second method, called the *synthetic dual Doppler (SDD) technique*, uses the Doppler radial velocity measurements from two times. This method can be used if the wind field remains nearly steady state in the reference frame of the storm between the times and the storm motion results in a significant change in the radar viewing angle with time. Figure 17 shows a schematic of the SDD geometry for (a) radar-relative and (b) storm-relative coordinates. In radar-relative coordinates, the storm is first viewed at $t - \Delta t/2$, where Δt is the time separation of the two radar volumes used for the SDD analysis and t is the time of the SDD wind retrieval. At some later time, $t + \Delta t/2$, the storm has moved a distance d to a new location, and the radar viewing angle β from the radar to the storm changes. Using radial velocity measurements from these two time periods, an SDD horizontal wind field can be retrieved for an intermediate time period t, when the storm was located at an intermediate distance $d/2$. The geometry of the storm-relative coordinates in Fig. 17b is identical to a conventional dual-Doppler (i.e., two-radar)

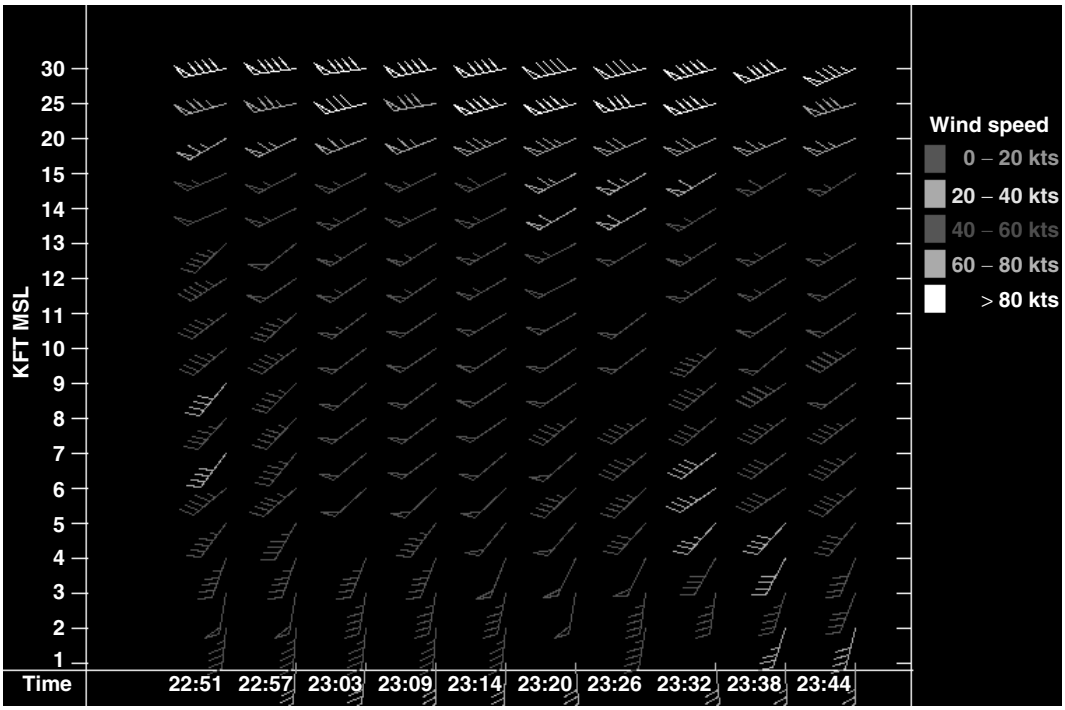


Figure 15. Wind speed and direction as a function of height and time derived using the velocity-azimuth display (VAD) technique. The data were collected by the Lincoln, IL, National Weather Service radar on 19 April 1996 between 22 : 51 and 23 : 44 Greenwich Mean Time. Long tails on a wind barb represent 10 knots (5 m s^{-1}), short tails 5 knots, and flags 50 knots. The direction in which the barb is pointing represents the wind direction. For example, a wind from the north at 20 knots would be represented by downward pointing barb that has two tails, and a 60-knot wind from the west would be represented by a right pointing barb that has a flag and a long tail. See color insert.

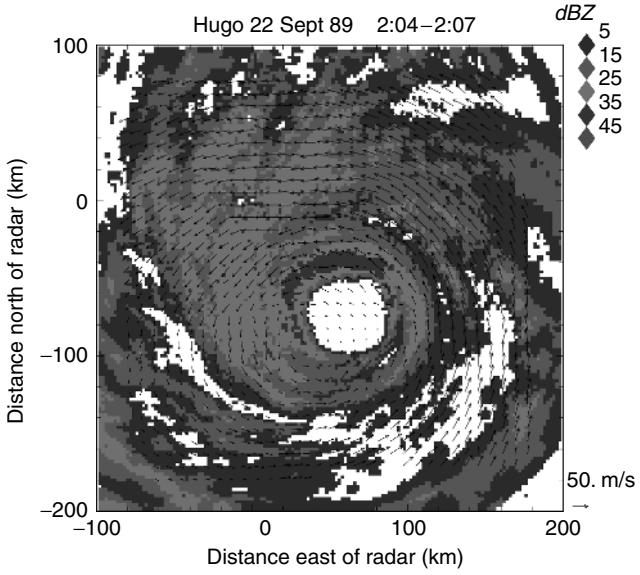


Figure 16. TREC-determined wind vectors for Hurricane Hugo overlaid on radar reflectivity. A 50 m s^{-1} reference vector is shown on the lower right. (From J. Tuttle and R. Gall, A single-radar technique for estimating the winds in a tropical cyclone. *Bulletin of the American Meteorological Society*, **80**, 653–688, 1998. Courtesy of John Tuttle and the American Meteorological Society.) See color insert.

system that is viewing a single storm during the same time period. However, when using the SDD technique for a single radar, it necessary to "shift the position" of the radar a distance $d/2$ for both time periods by using the storm propagation velocity. Figure 17b shows that using data collected at two time periods and shifting the radar position can, in essence, allow a single radar to obtain measurements of a storm from two viewing geometries at an intermediate time and location. Figure 18 shows an example of SDD winds recovered for a vortex that developed over the southern portion of Lake Michigan during a cold air outbreak and moved onshore in Michigan. The forward speed of the vortex has been subtracted from the wind vectors to show the circulation better. The bands of high reflectivity are due to heavy snow. The SDD wind retrieval from the WSR-88D radar at Grand Rapids, MI, clearly shows the vortex circulation and convergence of the wind flows into the radial snowbands extending from the vortex center, which corresponds with the position of a weak reflectivity "eye."

Multiple Doppler Retrieval of 3-D Wind Fields

Three-dimensional wind fields in many types of storm systems have been determined during special field campaigns in which two or more Doppler radars have been deployed. In these projects, the scanning techniques

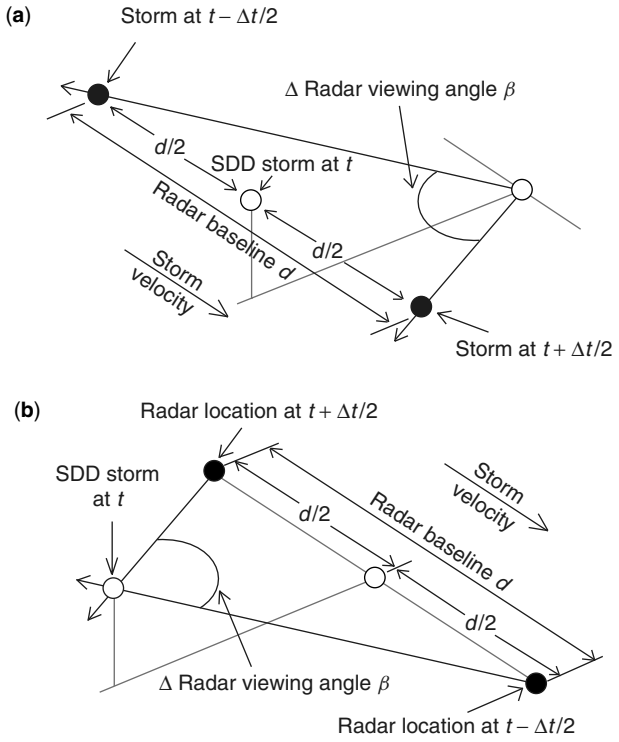


Figure 17. Schematic diagrams of the SDD geometry for (**a**) radar-relative and (**b**) storm-relative coordinates. Δt is the time separation of the two radar volumes used for the SDD analysis and t is the time of the SDD retrieval. The distance d is analogous to the radar baseline in a conventional dual-Doppler system. The solid circles represent (panel a) the observed storm locations and (panel b) the shifted radar locations. The open circles denote the location of the radar and the SDD retrieved storm position. (Courtesy of N. Laird, Illinois State Water Survey).

are optimized to cover the entire storm circulation from cloud top to the ground in as short a time as feasible. The techniques to recover the wind fields from radial velocity measurements from more than one Doppler radar are termed *multiple Doppler analysis*. Radar data are collected in spherical coordinates (radius, azimuth, elevation). Multiple Doppler analyses are normally done in Cartesian space, particularly because the calculation of derivative quantities, such as divergence of the wind field, and integral quantities, such as air vertical velocity, are required. For this reason, radial velocity and other data are interpolated from spherical to Cartesian coordinates. Data are also edited to remove nonmeteorological echoes such as ground clutter and second-trip echoes, and are unfolded to correct velocity ambiguities discussed previously. The data must also be adjusted spatially to account for storm motion during sampling.

The equation relating the radial velocity measured by a radar to the four components of motion of particles in a Cartesian framework is

$$v_r = u \sin a \cos e + v \cos a \cos e + (w + w_t) \sin e \qquad (17)$$

where u, v, and w are the west–east, north–south, and vertical components of air motion in the Cartesian system, a and e are the azimuthal and elevation angles of the radar, w_t is the mean fall velocity of the particles, and v_r is the measured radial velocity. For each radar viewing a specific location in a storm, v_r, a, and e are known and u, v, w, and w_t are unknown. In principle, four separate measurements are needed to solve for the desired four unknown quantities. In practice, four measurements of radial velocity from different viewing angles are rarely available. Most field campaigns employ two radars, although some have used more. The remaining unknown variables are estimated by applying constraints imposed by mass continuity in air flows, appropriate application of boundary conditions at storm boundaries, estimates of particle fall velocities based on radar reflectivity factors, and additional information available from other sensors.

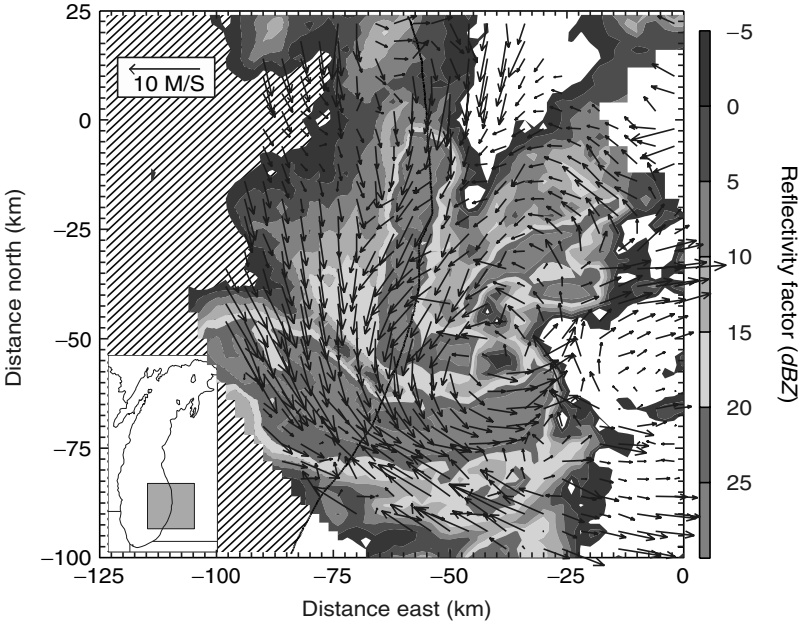


Figure 18. Radar reflectivity factor and winds (relative to the forward speed of the vortex) at a 2-km altitude within a vortex over Lake Michigan derived from two radar volumes collected by the Grand Rapids, MI, WSR-88D Doppler radar at 1,023 and 1,123 UTC on 5 December 1997. The winds were derived using the synthetic dual-Doppler technique. (Courtesy of N. Laird, Illinois State Water Survey). See color insert.

The details of multiple Doppler analysis can become rather involved, but the results are often spectacular and provide exceptional insight into the structure and dynamics of storm circulations.

For example, Fig. 19a shows a vertical cross section of the radar reflectivity and winds across a thunderstorm derived from measurements from two Doppler radars located near the storm. The forward speed of the storm has been subtracted from the winds to illustrate the circulations within the storm better. The vertical scale on Fig. 19a is stretched to illustrate the storm structure better. A 15-km wide updraft appears in the center of the storm, and central updraft speeds approach 5 $m\,s^{-1}$. The updraft coincides with the heaviest rainfall, indicated by the high reflectivity region in the center of the storm. Figure 19b shows the horizontal wind speed in the plane of the cross section. The sharp increase in wind speed marks the position of an advancing front, which is lifting air to the east of the front (right side of the figure), creating the updraft appearing in Fig. 19a. Three-dimensional Doppler wind analyses in severe storms have been used to determine the origin and trajectories of large hailstones, examine the origin of rotation in tornadoes, study the circulations in hurricanes, and investigate the structure of a wide variety of other atmospheric phenomena such as fronts, squall lines, and downbursts.

Retrieval of Thermodynamic Parameters from 3-D Doppler Wind Fields

Wind in the atmosphere is a response to variations in atmospheric pressure. Pressure variations occur on many scales and for a variety of reasons but are closely tied on larger scales to variations in air temperature. Newton's second law of motion describes the acceleration of air

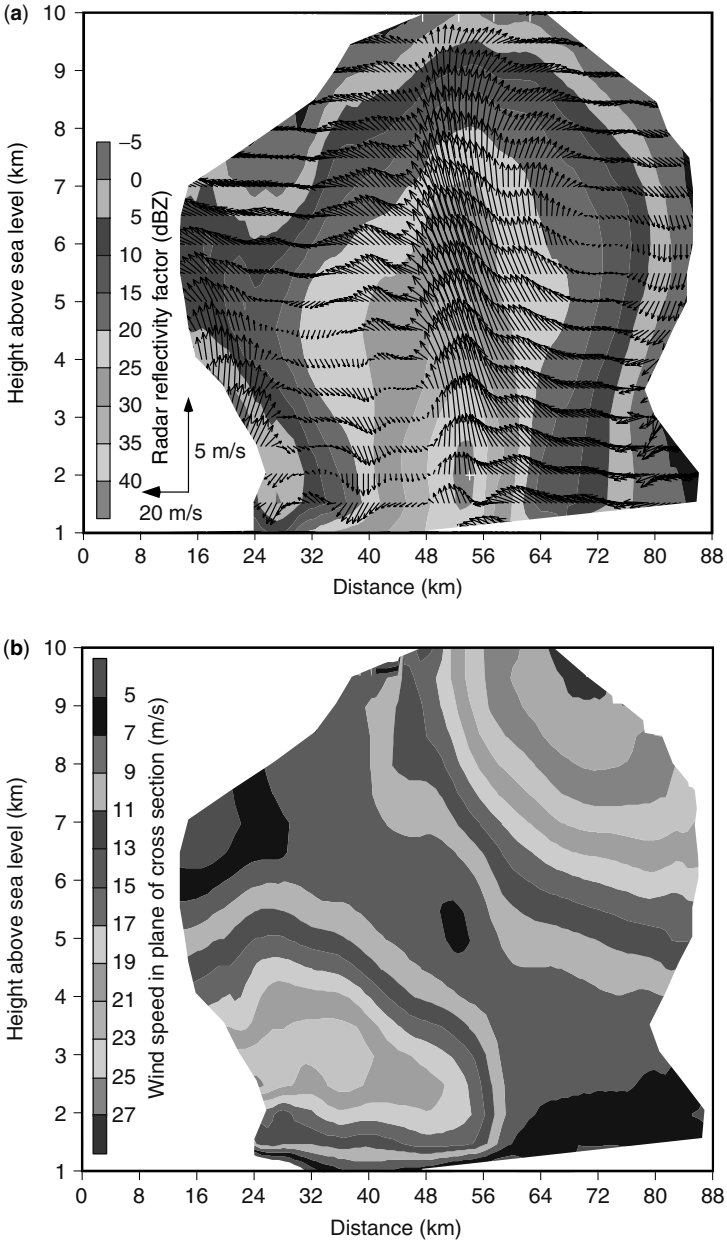


Figure 19. Vertical cross sections through a thunderstorm derived from measurements from two Doppler radars. The storm occurred in northeast Kansas on 14 February, 1992. The panels show (**a**) radar reflectivity (*dBZ*) and winds (vectors, $m\,s^{-1}$) in the plane of the cross section. The forward speed of the storm has been subtracted from the wind vectors to illustrate the vertical circulations within the storm; (**b**) the horizontal wind speed ($m\,s^{-1}$) in the plane of the cross section. Positive values of the wind speed denote flow from left to right; (**c**) the perturbation pressure field (millibars) within the storm. The perturbation pressure is the pressure field remaining after the average pressure at each elevation is subtracted from the field. See color insert.

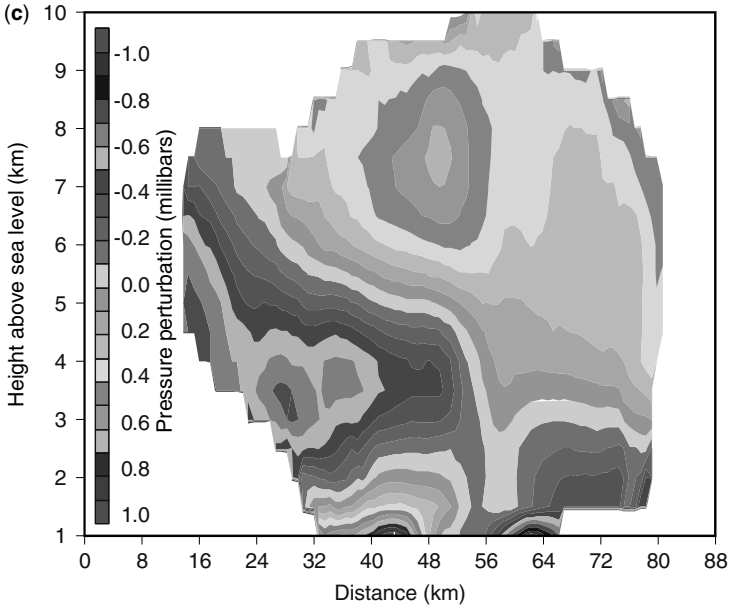


Figure 19. (*continued*)

and its relation to atmospheric pressure, thermal fields, the earth's rotation, and other factors. For atmospheric processes, Newton's law is expressed in the form of three equations describing the acceleration of the winds as a response to forces in the three cardinal directions. Doppler wind fields and the momentum equations have been used to retrieve perturbation pressure and buoyancy information from derived three-dimensional multiple Doppler wind fields. More recently, retrieval techniques have also been developed that incorporate the thermodynamic equation, which relates temperature variations in air parcels to heating and cooling processes in the atmosphere. These retrieval techniques have extended multiple Doppler radar analyses from purely kinematic descriptions of wind fields in storms to analysis of the dynamic forces that create the wind fields.

As an example, the retrieved pressure perturbations associated with the storm in Fig. 19b are shown in Fig. 19c. The primary features of the pressure field include a positive pressure perturbation in the upper part of the storm located just to the left of the primary updraft, a large area of negative pressure perturbation both in and above the sharp wind gradient in the middle and upper part of the storm, a strong positive perturbation at the base of the downdraft to the left of the main updraft and a weak negative pressure perturbation in the low levels ahead of the advancing front. Physically, the pressure perturbations are associated with two physical processes: (1) horizontal accelerations at the leading edge of the front and within the outflow at the top of the updraft, and (2) positive buoyancy in the updraft and negative buoyancy in the downdraft. Analyses such as these help meteorologists understand how storms form and organize and the processes that lead to their structures.

POLARIZATION DIVERSITY RADARS

As electromagnetic waves propagate away from a radar antenna, the electrical field becomes confined to a plane that is normal to the propagative direction. The orientation of the electrical field vector within this plane determines the wave's polarization state. For radars, the electrical field vector either lies on a line or traces out an ellipse in this plane, which means that radar waves are polarized. If the electrical field lies on a line, the condition for most meteorological radars, the waves are *linearly polarized*. A radar wave that propagates toward the horizon is *vertically polarized* if the electrical field vector oscillates in a direction between the zenith and the earth's surface, and is *horizontally polarized* if the vector oscillates in a direction parallel to the earth's surface. Polarization diversity radars measure echo characteristics at two orthogonal polarizations, typically, horizontal and vertical. This is done either by changing the polarization of successive pulses and/or transmitting one and receiving both polarizations. At this time, polarization diversity radars are used only in meteorological research. However, the U.S. National Weather Service is planning to upgrade its Doppler radar network in the near future to include polarization capability.

Polarization diversity radars take advantage of the fact that precipitation particles have different shapes, sizes, orientations, dielectric constants, and number densities. For example, raindrops smaller than about 1 mm in diameter are spherical, but larger raindrops progressively flatten due to air resistance and take on a "hamburger" shape as they become large. Hailstones are typically spherical or conical but may take on more diverse shapes depending on how their growth proceeds. Hailstones sometimes develop a water coating, while growing at subfreezing temperatures, due to heat deposited on the hailstone surface during freezing. The water coating changes the dielectric constant of the hail surface. Small ice crystals typically are oriented horizontally but become randomly oriented as they grow larger. Eventually, individual crystals form loosely packed, low-density snowflakes as they collect one another during fall. When snowflakes fall through the melting level, they develop

wet surfaces and a corresponding change in dielectric constant. Horizontally and vertically polarized waves are closely aligned with the natural primary axes of falling precipitation particles and therefore are ideal orientations to take advantage of the particle characteristics to identify them remotely.

Linearly polarized waves induce strong electrical fields in precipitation particles in the direction of electrical field oscillation and weak fields in the orthogonal direction. For particles that have a large aspect ratio, such as large raindrops, a horizontally polarized wave induces a larger electrical field and subsequently, a larger returned signal than a vertically polarized wave. In general, the two orthogonal fields provide a means of probing particle characteristics in the two orthogonal dimensions. Differences in particle characteristics in these dimensions due to shape or orientation will appear as detectable features in the returned signal. These are *backscatter effects* related to particles in the radar scattering volume. *Propagation effects* such as attenuation, which are associated with particles located between the radar and the scattering volume, also differ for the two orthogonal polarizations. Measurement of the differences in propagation at orthogonal polarization provides further information about the characteristics of particles along the beam path.

There are six backscatter variables and four propagation variables that carry meaningful information provided by polarization diversity radars that employ linear polarization. Other variables are derived from these basic quantities. The most important backscatter variables are (1) the *reflectivity factor* Z for horizontal polarization; (2) the *differential reflectivity* Z_{DR}, which is the ratio of reflected power at horizontal and vertical polarization; (3) the *linear depolarization ratio* (LDR), which is the ratio of the cross-polar power (transmitted horizontally, received vertically) to the copolar power (transmitted and received horizontally); (4) the complex *correlation coefficient* $\rho_{hv}e^{j\delta}$ between copolar horizontally and vertically polarized echo signals; and (5) the *phase* of the correlation coefficient δ, which is the difference in phase between the horizontally and vertically polarized field caused by backscattering. In the expression for the correlation coefficient, $j = \sqrt{-1}$. Propagative effects that influence polarization measurements include (1) attenuation of the horizontally polarized signal, (2) attenuation of the vertically polarized signal, (3) depolarization, and (4) the differential phase shift Φ_{DP}. A differential phase shift, or lag, occurs in rain because horizontally polarized waves propagate more slowly than vertically polarized waves. This occurs because larger raindrops are oblate and present a larger cross section to horizontally polarized waves. The variable of most interest is the specific differential phase K_{DP}, which is the range derivative of Φ_{DP}. The specific differential phase, it has been shown, is an excellent indicator of liquid water content and rain rate and may be superior to rain rates derived from standard Z–R relationships based on Eq. (10).

An example of the capabilities of polarimetric radars in identifying precipitation type appears in Fig. 20. This figure shows a cross section of (a) Z, the radar reflectivity factor for horizontally polarized radiation; (b) Z_{DR}, the differential reflectivity; and (c) a particle classification based on polarization variables. Note in panel (c) that the hail shaft, in yellow, has high Z and low Z_{DR}, and the rain to the right of the hail shaft has low Z and high Z_{DR}. These variables, combined with other polarimetric measurements, allow meteorologists to estimate the other types of particles that populate a storm. For example, the upper part of the storm in Fig. 20 contained hail and graupel (smaller, softer ice spheres), and the "anvil" top of the storm extending to the right of the diagram was composed of dry snow and irregular ice crystals. Identifying boundaries between different particle types uses techniques invoking "fuzzy logic" decisions, which take into account the fact that there is overlap between various polarization parameters for various precipitation types.

The use of K_{DP} for precipitation measurements is especially promising. The major advantages of K_{DP} over Z is that K_{DP} is independent of receiver and transmitter calibrations, unaffected by attenuation, less affected by beam blockage, unbiased by ground clutter cancelers, less sensitive to variations in the distributions of drops, biased little by the presence of hail, and can be used to detect anomalous propagation. For these reasons, research efforts using polarization diversity radars have focused particularly on verification measurements of rainfall using K_{DP}. Quantitative predictions of snowfall may also be possible using polarization diversity radars, but this aspect of precipitation measurement has received less attention in meteorological research because of the greater importance of severe thunderstorms and flash flood forecasting.

WIND PROFILING RADARS

A wind profiling radar, or *wind profiler*, is a Doppler radar used to measure winds above the radar site, typically to a height of about 15 km above the earth's surface. Wind profilers are low-power, high-sensitivity radars that operate best in clear air conditions. Profilers, which operate at UHF and VHF frequencies, detect fluctuations in the index of refraction at half the radar wavelength caused by fluctuations in the radio refractive index. These fluctuations arise from variations in air density and moisture content primarily from turbulence. Radar meteorologists assume that the fluctuations in the radio refractive index are carried along with the mean wind, and therefore the Doppler frequency shift for the motion of the scattering elements can be used to estimate the wind.

Wind profilers use fixed *phased array* antennas. In the 404.37-MHz (74-cm wavelength) profilers used by the U.S. National Weather Service, an antenna is made up of a 13×13 meter grid of coaxial cables, and the antenna itself consists of many individual radiating elements, each similar to a standard dipole antenna. If a transmitted pulse arrives at each of these elements at the same time (in phase), a beam propagates away from the antenna vertically. If the pulses arrive at rows of elements at slightly different times (out of phase), a beam propagates upward at an angle to the zenith. The phasing is controlled

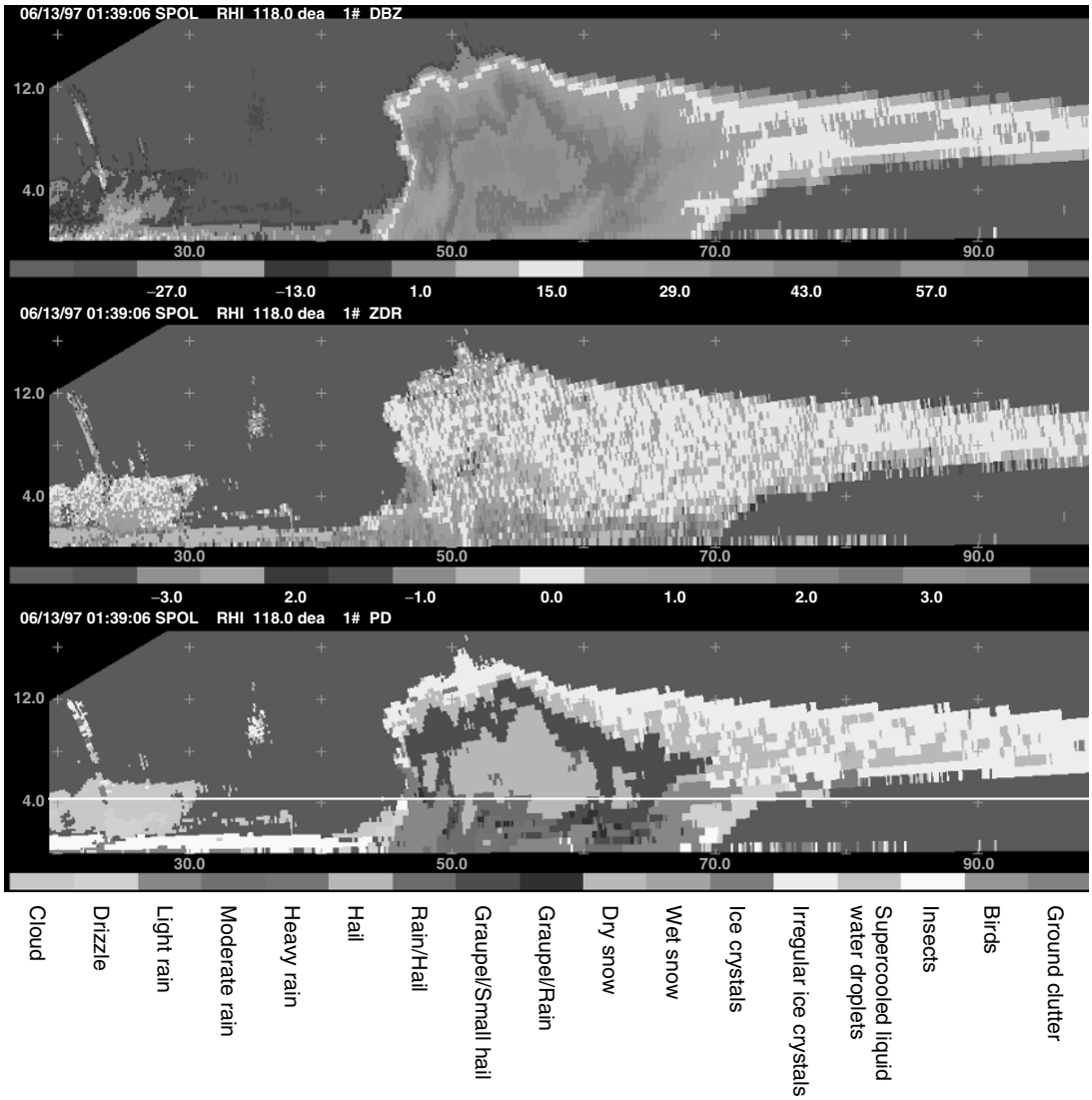


Figure 20. Range-height indicator scans of (top) radar reflectivity factor for horizontal polarization Z; (middle) differential reflectivity Z_{DR}; (bottom) particle classification results based on analysis of all polarimetric parameters. (From J. Vivekanandan, D. S. Zrnic, S. M. Ellis, R. Oye, A. V. Ryzhkov, and J. Straka, Cloud microphysics retrieval using S band dual-polarization radar measurements. *Bulletin of the American Meteorological Society* **80**, 381–388 (1999). Courtesy of J. Vivekanandan and the American Meteorological Society.) See color insert.

by changing the feed cable lengths. Profilers typically use a three-beam pattern that has a vertically pointing beam and beams pointing in orthogonal directions (e.g., north and east). In the absence of precipitation, the radial velocities measured by a profiler along each beam are

$$v_{re} = u \cos e + w \sin e, \tag{18}$$

$$v_{rn} = v \cos e + w \sin e, \tag{19}$$

$$v_{rv} = w, \tag{20}$$

where v_{re}, v_{rn}, and v_{rv} are the radial velocities measured by the beams in the east, north, and vertical pointing positions; u, v, and w are the wind components in the west–east, south–north, and upward directions; and e is the angle of elevation of the east and north beams above the earth's surface. Because the vertical beam measures w directly and e and the radial velocities are measured, these equations can be solved for u and v, which can be easily converted to wind speed and direction. Profilers are pulsed radars, so the round-trip travel time of the pulse measures the height of the wind. Precipitation particles also scatter energy and therefore contribute to the measurement of the radial velocity. When precipitation occurs, w in Eqs. (18)–(20) would be replaced by $w + w_t$, where w_t is the terminal fall velocity of the precipitation particles. Problems arise if all three beams are not filled by the same size precipitation particles. In this case, the winds may not be recoverable. The radars also require a recovery time after the pulse is transmitted before accurate data can be received by the transmitter; so information in the lowest 500 meters of the atmosphere is typically not recoverable.

Wind profiles can be obtained at high time resolution, often at times as short as 6 minutes, and vertical resolution of the order of 250 meters. The 6-minute time interval, compared with the standard 12-hour time interval for standard weather balloon launch wind measurements, represents a dramatic increase in a meteorologist's capability of sampling upper atmospheric winds. Wind profilers are used for a wide range of research applications in meteorology, often in combination with other instruments, such as microwave radiometers and acoustic sounding devices that remotely measure moisture and temperature profiles. The data from profilers are presented in a format essentially identical to Fig. 15.

Currently, the U.S. National Oceanic and Atmospheric Administration operates a network of 30 profilers called the Wind Profiler Demonstration Network. Most of these profilers are located in the central United States and are used for weather monitoring by the National Weather Service and in forecasting applications such as the initialization and verification of numerical weather forecasting models.

MOBILE RADAR SYSTEMS

Mobile radar systems consist of three distinct classes of instruments: rapidly deployable ground based radars, airborne radars, and satellite-borne radars. Each of these radar systems is well-suited to address particular research problems that are either difficult or impossible to carry out from fixed ground-based systems.

Rapidly deployable ground-based radars are used to study tornadoes, land-falling hurricanes, and other atmospheric phenomena that have small space and timescales and that must be sampled at close range. These phenomena are unlikely to occur close to a fixed radar network, particularly during relatively short research field campaigns. At this time, six mobile radars are available to the meteorological research community: two truck-mounted 3-cm wavelength radars called the "Doppler on Wheels," or DOWs, operated jointly by the University of Oklahoma and the National Center for Atmospheric Research; two truck-mounted 5-cm wavelength radars called the Shared Mobile Atmospheric Research and Teaching (SMART) radars operated by the National Severe Storms Laboratory, Texas A&M University, Texas Tech University, and the University of Oklahoma; a 3-mm wavelength truck-mounted Doppler system operated by the University of Massachusetts; and a 3-mm wavelength trailer-mounted Doppler radar operated by the University of Miami. The DOW and SMART radars are used in dual-Doppler arrangements to measure wind fields near and within tornadoes and hurricanes. The millimeter wavelength radars have been used to study tornadoes and cloud and precipitation processes.

Figure 21 shows a DOW image of the radar reflectivity and radial velocity of a tornado near Scottsbluff, Nebraska. The tornado is located at the position of the tight inbound/outbound radial velocity couplet near the center of the image in the left panel. The reflectivity factor in the right panel of the figure shows a donut-shaped reflectivity region that has a minimum in reflectivity at the tornado center. It is thought that this occurs because debris is centrifuged outward from the tornado center. In a tornado, the backscattered energy comes primarily from debris.

Although great progress has been made using ground-based Doppler radars to study storm structure, remote storms such as oceanic cyclones and hurricanes cannot be observed by these systems. In addition, small-scale phenomena such as tornadoes rarely occur close enough to special fixed dual-Doppler networks so that detailed data, aside from new data obtained by the DOWs, are hard to obtain. Airborne meteorological radars provide a means to measure the structure and dynamics of these difficult to observe weather systems. Currently, three research aircraft flown by the meteorological community have scanning Doppler radars, two P-3 aircraft operated by the National Oceanic and Atmospheric Administration and a Lockheed Electra aircraft operated by the National Science Foundation through the National Center for Atmospheric Research (NCAR). The scanning technique used by these radars to map wind fields is shown in Fig. 22. In the P-3 aircraft, a single radar uses an antenna designed to point alternately at angles fore and then aft of the aircraft. In the Electra, two radars are used; the first points fore, and the second points aft. In both cases, the scan pattern consists of an array of beams that cross at an angle sufficient to sample both components of the horizontal wind. Therefore, the data from the fore and aft beams can be used as a dual-Doppler set that permits recovery of wind fields. Three other aircraft, an ER-2 high-altitude aircraft and a DC-8 operated by NASA, and a King Air operated by the University of Wyoming also have radars used for meteorological research.

Airborne radars have significant limitations imposed by weight, antenna size, and electrical power requirements. Aircraft flight capabilities in adverse weather, stability in turbulence, and speed all impact the quality of the measurements. The aircraft's precise location, instantaneous three-dimensional orientation, and beam pointing angle must be known accurately to position each sample in space. At altitudes other than the flight altitude, the measured radial velocity contains a component of the particle fall velocity, which must be accounted for in processing data. This component becomes progressively more significant as the beam rotates toward the ground and zenith. A time lag exists between the forward and aft observations during which the storm evolution will degrade the accuracy of the wind recovery. Short wavelengths (e.g., 3 cm) must be used because of small antenna and narrow beam-width requirements that make attenuation and radial velocity folding concerns.

Despite these limitations, airborne Doppler radars have provided unique data sets and significant new insight into storm structure. For example, Fig. 23 shows an RHI scan of the radar reflectivity factor through a tornadic thunderstorm near Friona, TX, on 2 June 1995 measured by the ELDORA radar on the NCAR Electra aircraft. The data, which was collected during the Verification of the Origins of Rotations in Tornadoes experiment (VORTEX), shows a minimum in the reflectivity of the Friona tornado extending from the ground to the cloud top, the first

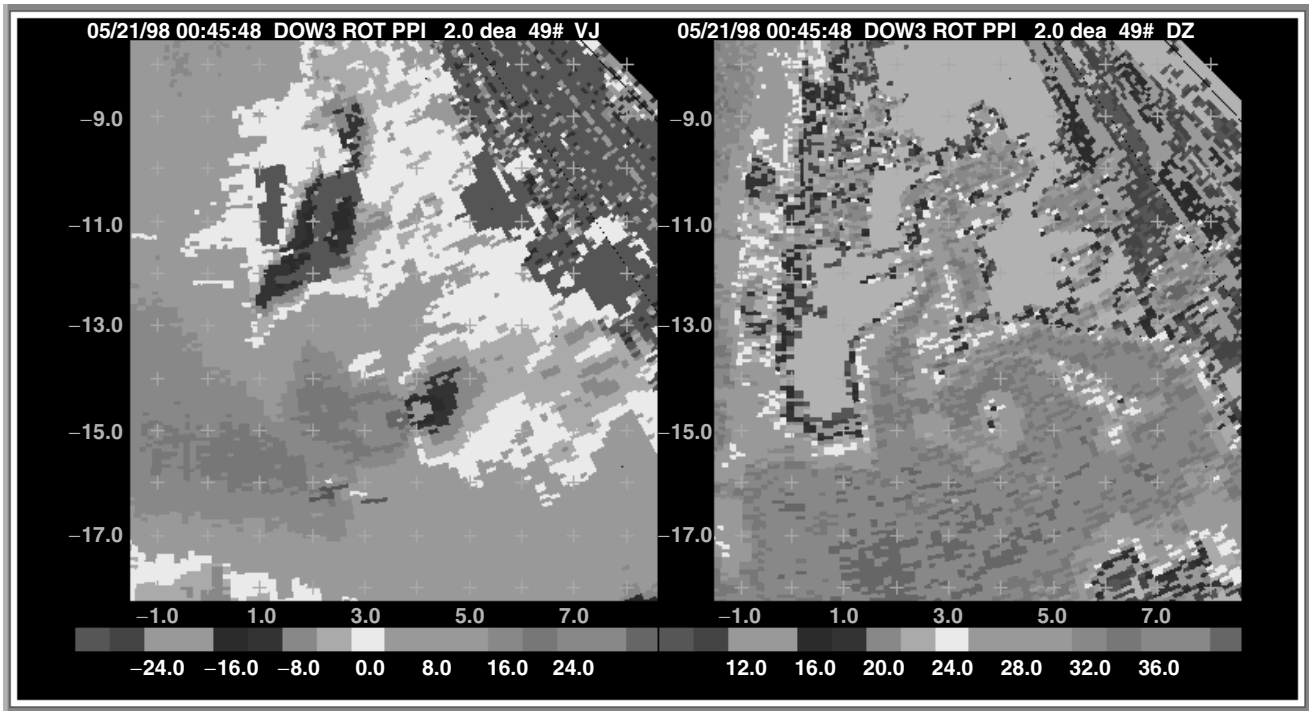


Figure 21. Image of radial velocity ($m\,s^{-1}$; left) and reflectivity factor (*dBZ*; right) taken by a Doppler on Wheels radar located near a tornado in Scottsbluff, Nebraska, on 21 May 1998. Blue colors denote inbound velocities in the radial velocity image. (Courtesy of J. Wurman, University of Oklahoma.) See color insert.

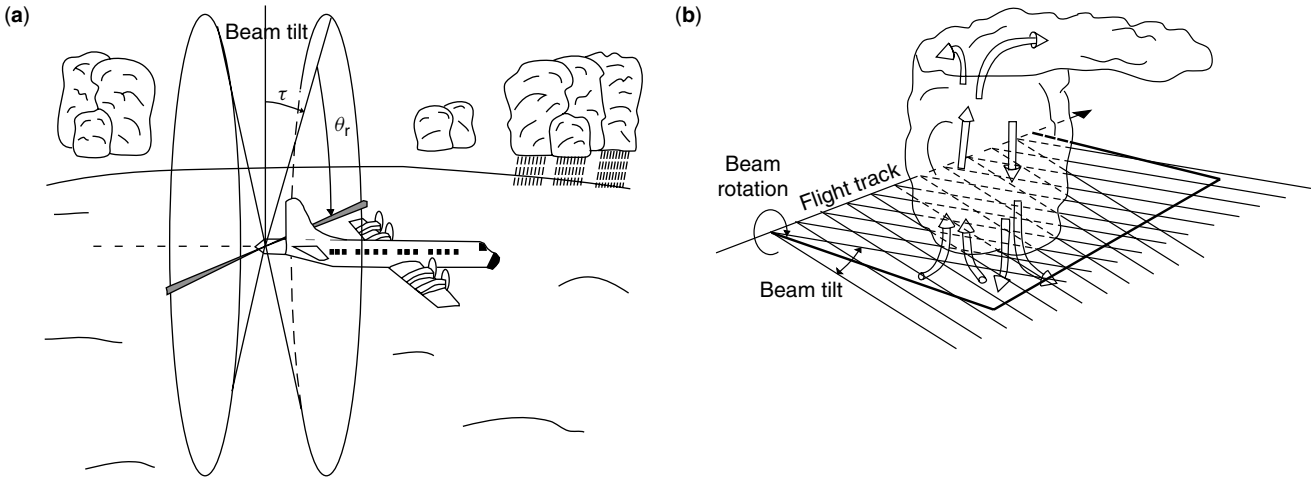


Figure 22. (**a**) ELDORA/ASTRAIA airborne radar scan technique showing the dual-radar beams tilted fore and aft of the plane normal to the fuselage. The antennas and radome protective covering rotate as a unit about an axis parallel to the longitudinal axis of the aircraft. (**b**) Sampling of a storm by the radar. The flight track past a hypothetical storm is shown. Data are taken from the fore and aft beams to form an analysis of the velocity and radar reflectivity field on planes through the storm. The radial velocities at beam intersections are used to derive the two-dimensional wind field on the analysis planes. (From P. H. Hildebrand et al., The ELDORA-ASTRAIA airborne Doppler weather radar: High-resolution observations from TOGA-COARE. *Bulletin of the American Meteorological Society* **77**, 213–232 (1996) Courtesy of American Meteorological Society.)

time such a feature has ever been documented in a tornadic storm.

Space radars include altimeters, scatterometers, imaging radars, and most recently, a precipitation radar whose measurement capabilities are similar to other radars described in this article. Altimeters, scatterometers and imaging radars are used primarily to determine properties of the earth's surface, such as surface wave height,

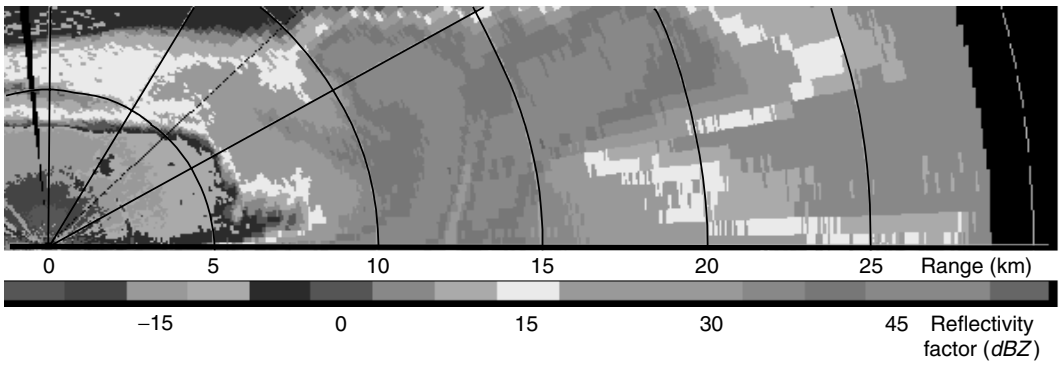


Figure 23. RHI cross section of the radar reflectivity factor through a severe thunderstorm and tornado near Friona, TX, on 2 June 1995 measured by the ELDORA radar on board the National Center for Atmospheric Research Electra aircraft. The data was collected during VORTEX, the Verification of the Origins of Rotations in Tornadoes experiment. (From R. M. Wakimoto, W. -C. Lee, H. B. Bluestein, C. -H. Liu, and P. H. Hildebrand, ELDORA observations during VORTEX 95. *Bulletin of the American Meteorological Society* **77**, 1,465–1,481 (1996). Courtesy of R. Wakimoto and the American Meteorological Society.) See color insert.

the location of ocean currents, eddies and other circulation features, soil moisture, snow cover, and sea ice distribution. These parameters are all important to meteorologists because the information can be used to initialize numerical forecast and climate models. Surface winds at sea can also be deduced because short gravity and capillary waves on the ocean surface respond rapidly to the local near-instantaneous wind and the character of these waves can be deduced from scatterometers.

The first precipitation radar flown in space was launched aboard the Tropical Rainfall Measuring Mission (TRMM) satellite in November 1997. The TRMM radar, jointly funded by Japan and the United States, is designed to obtain data concerning the three-dimensional structure of rainfall over the tropics where ground-based and ocean-based radar measurements of precipitation are almost nonexistent. Figure 24, for example, shows the reflectivity factor measured in Hurricane Mitch over the Caribbean during October 1998. A unique feature of the precipitation radar is its ability to measure rain over land, where passive microwave instruments have more difficulty. The data are being used in conjunction with other instruments on the satellite to examine the atmospheric energy budget of the tropics. The radar uses a phased array antenna that operates at a frequency of 13.8 GHz. It has a horizontal resolution at the ground of about 4 km and a swath width of 220 km. The radar measures vertical profiles of rain and snow from the surface to a height of about 20 kilometers at a vertical resolution of 250 m and can detect rain rates as low as 0.7 millimeters per hour. The radar echo of the precipitation radar consists of three components: echoes due to rain; echoes from the surface; and mirror image echoes, rain echoes received through double reflection at the surface. At intense rain rates, where the attenuation effects can be strong, new methods of data processing using these echoes have been developed to correct for attenuation. The Precipitation Radar is the first spaceborne instrument that provides three-dimensional views of storm structure. The measurements are currently being analyzed and are expected to yield invaluable information on the intensity and distribution of rain, rain type, storm depth, and the height of the melting level across tropical latitudes.

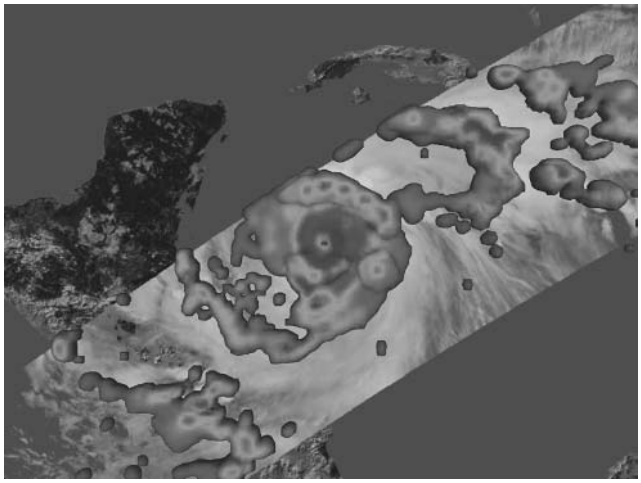


Figure 24. Radar reflectivity factor within Hurricane Mitch in 1998 over the Caribbean Sea measured by the Precipitation Radar on board the Tropical Rainfall Measuring System (TRMM) satellite in October 1998. The viewing swath is 215 kilometers wide. (Courtesy of the National Aeronautics and Space Administration.) See color insert.

FUTURE DEVELOPMENTS

A number of advancements are on the horizon in radar meteorology. Scientists now know that polarimetric radars have the potential to obtain superior estimates of rainfall compared to nonpolarimetric radars, and give far better information concerning the presence of hail. Because of these capabilities, polarization diversity should eventually become incorporated into the suite of radars used by the U.S. National Weather Service for weather monitoring and severe weather warning. Experimental bistatic radars, systems that have one transmitter but

many receivers distributed over a wide geographic area, were first developed for meteorological radars in the 1990s and are currently being tested as a less expensive way to retrieve wind fields in storms. Bistatic radar receivers provide a means to obtain wind fields by using a single Doppler radar. Each of the additional receivers measures the pulse-to-pulse phase change of the Doppler shift, from which they determine the wind component toward or away from the receiver. Because these receivers view a storm from different directions, all wind components are measured, making it possible to retrieve wind fields within a storm similarly to that currently done by using two or more Doppler radars. Future networks of bistatic radars may make it possible to create images of detailed wind fields within storms in near-real time, providing forecasters with a powerful tool to determine storm structure and severity. At present, the operational network of wind profilers in the United States is limited primarily to the central United States. Eventual expansion of this network will provide very high temporal monitoring of the winds, leading to more accurate initialization of numerical weather prediction models and ultimately, better forecasts. Mobile radars are being developed that operate at different wavelengths and may eventually have polarization capability. One of the biggest limitations in storm research is the relatively slow speed at which storms must be scanned. A complete volume scan by a radar, for example, typically takes about 6 minutes. Techniques developed for military applications are currently being examined to reduce this time by using phased array antennas that can scan a number of beams simultaneously at different elevations. Using these new techniques, future radars will enhance scientists' capability to understand and predict a wide variety of weather phenomena.

ABBREVIATIONS AND ACRONYMS

COHO	coherent local oscillator
DPW	doppler on wheels
ELDORA	electra doppler radar
LDR	linear depolarization ratio
NASA	national aeronautics and space administration
NCAR	national center for atmospheric research
PPI	plan position indicator
RHI	range-height indicator
SDD	synthetic dual doppler
SMART	shared mobile atmospheric research and teaching
STALO	stable local oscillator
TREC	objective tracking of radar echo with correlations
TRMM	tropical rainfall measuring mission
UHF	ultra high frequency
VAD	velocity azimuth display
VHF	very high frequency
VORTEX	verification of the origins of rotations in storms experiment
WRS-57	weather surveillance radar 1957
WSR-88D	weather surveillance radar 1988-doppler

BIBLIOGRAPHY

1. D. Atlas, ed., *Radar in Meteorology*, American Meteorological Society, Boston, 1990.
2. L. J. Battan, *Radar Observation of the Atmosphere*, University of Chicago Press, Chicago, 1973.
3. T. D. Crum, R. E. Saffle, and J. W. Wilson, *Weather and Forecasting* **13**, 253–262 (1998).
4. R. J. Doviak and D. S. Zrnić, *Doppler Radar and Weather Observations*, 2nd ed., Academic Press, San Diego, CA, 1993.
5. T. Matejka and R. C. Srivastava, *J. Atmos. Oceanic Technol.* **8**, 453–466 (1991).
6. R. Reinhart, *Radar for Meteorologists*, Reinhart, Grand Forks, ND, 1997.
7. F. Roux, *Mon. Weather Rev.* **113**, 2,142–2,157 (1985).
8. M. A. Shapiro, T. Hample, and D. W. Van De Kamp, *Mon. Weather Rev.* **112**, 1,263–1,266 (1984).
9. M. Skolnik, *Introduction to Radar Systems*, 2nd ed., McGraw-Hill, NY, 1980.
10. J. Tuttle and R. Gall, *Bull. Am. Meteorol. Soc.* **79**, 653–668 (1998).
11. R. M. Wakimoto et al., *Bull. Am. Meteorol. Soc.* **77**, 1,465–1,481 (1996).
12. B. L. Weber et al., *J. Atmos. Oceanic Technol.* **7**, 909–918 (1990).
13. J. Vivekanandan et al., *Bull. Am. Meteorol. Soc.* **80**, 381–388 (1999).
14. D. S. Zrnic and A. V. Ryzhkov, *Bull. Am. Meteorol. Soc.* **80**, 389–406 (1999).

X

X-RAY FLUORESCENCE IMAGING

SHIH-LIN CHANG
National Tsing Hua University
Hsinchu, Taiwan
Synchrotron Radiation
Research Center
Hsinchu, Taiwan

INTRODUCTION

In Wilhelm Conrad Röntgen's cathode-ray-tube experiments, the lighting up of a screen made from barium platinocyanide crystals due to fluorescence led to his discovery of X rays in 1895. Subsequent investigations by Röntgen and by Straubel and Winkelmann showed that the radiation that emanated from a fluorspar, or fluorite crystal CaF_2 excited by X rays is more absorbing than the incident X rays. This so-called fluorspar radiation is X-ray fluorescence (1). Nowadays, it is known that X-ray fluorescence is one of the by-products of the interaction of electromagnetic (EM) waves with the atoms in matter in the X-ray regime (0.1–500 Å). During the interaction, the incident X rays have sufficient energy to eject an inner electron from an atom. The energy of the atom is raised by an amount equal to the work done in ejecting the electron. Therefore, the atom is excited to states of higher energies by an incident EM wave. The tendency to have lower energy brings the excited atom back to its stable initial states (of lower energies) by recapturing an electron. The excess energy, the energy difference between the excited and the stable state, is then released from the atom in the form of EM radiation, or photons — this is fluorescence (2–4). The elapsed time between excitation and emission is only about 10^{-16} second. The energy of X-ray fluorescence is lower than that of the incident radiation. In general, the distribution of the X-ray fluorescence generated is directionally isotropic; the fluorescence is radiated in all directions at energies characteristic of the emitting atoms. Thus, X-ray fluorescence is element-specific.

X-ray fluorescence is a two-step process; X-ray absorption (the ejection of an inner electron) and X-ray emission (the recapture of an electron) (4). Both involve the electron configuration of atoms in the irradiated material. To eject an electron, the atom needs to absorb sufficient energy from the incident X rays of energy E and $E \geq W_\gamma$, where W_γ is the energy required to remove an electron from the γ shell. The γ shell corresponds to K, L, M electrons. An atom is said to be in the K quantum state if a K electron is ejected. It is similar for the L and M quantum states. Because a K electron is more tightly bound to the nucleus than the L and M electrons, the energy of the K quantum state is higher than that of the L quantum state. The energies corresponding to the work W_K, W_L, W_M, done by the EM wave to remove K, L, M electrons are called the K, L, ... absorption edges of the atom, or in terms of wavelength, $\lambda_K, \lambda_L, \ldots$, respectively (2–4). The absorption at/near $E \cong W_\gamma$ is different from the common phenomena of absorption, because the absorptive cross section and the linear absorption coefficient increase abruptly at absorption edges. This phenomenon is the so-called anomalous dispersion due to its resonant nature (2,3). In normal absorption, the intensity of a transmitted X-ray beam through a material is attenuated exponentially as the material increases in thickness according to a normal linear absorption coefficient μ. Usually, the normal linear absorption coefficient μ is related to the photoelectric absorption cross section of the electron σ_e by $\mu = \sigma_e n_0 \rho$, where n_0 is the number of electrons per unit volume in the material and ρ is the mass density. The atomic absorption coefficient is $\mu^a = (A/N_O)(\mu/\rho)$, where A is the atomic weight of the element in question and N_O is Avogadro's number. For $\lambda \neq \lambda_\gamma$ (i.e., $E \neq W_\gamma$), μ^a is approximately proportional to λ^3 and Z^4, according to quantum mechanical considerations (5,6), Z is the atomic number. When E increases (i.e., λ decreases), μ^a decreases according to λ^3. When $E = W_\gamma$ (i.e., $\lambda = \lambda_\gamma$), μ^a increases abruptly because X rays are absorbed in the process of ejecting γ electrons. For $E > W_\gamma$, the absorption resumes a decreasing trend as λ^3 (2,3).

According to the electron configuration of atoms, the K, L, M electrons, etc. are specified by the principal quantum number n, orbital angular quantum number ℓ, magnetic quantum number m_l, and spin quantum number m_s, or by n, ℓ, j, m, where $j = \ell \pm m_s$, $m = \pm j$ (3–6). Following the common notations used, the s, p, d, f, ... subshells and the quantum numbers n, ℓ, j, m are also employed to indicate the absorption edges or to relate to the ejection of electrons in the subshell. For example, there are three absorption edges for L electrons, L_I,L_{II}, L_{III}, corresponding to the three energy levels (states) specified by n, ℓ, and j: two electrons in (2, 0, 1/2), two electrons in (2, 1, 1/2), and four electrons in (2, 1, 3/2). The L_I state involves 2s electrons, and L_{II} and L_{III} involve six 2p electrons. The difference between L_{II} and L_{III} is that $j < \ell$ for the former and $j > \ell$ for the latter.

X-ray emission, the other process, involves the allowed transition of an atom from a high-energy state to a lower one. Usually, a high-energy state has a vacancy in the tightly bound inner shell. During the transition, the vacancy in the inner shell is filled with an electron coming from the outer shell. The energy of the fluorescence emitted is the energy difference between the two states involved. According to quantum mechanics (5–7), the allowed transition, the transition whose probability is high, is governed by selection rules. In the electric dipole approximation (5–7), the changes in quantum numbers in going from one state to the other follow the conditions $\Delta\ell = \pm 1$ and $\Delta j = 0$ or ± 1. For example, in the transition from a K state to an L state, the atom can only go from the K state to either the L_{II} or L_{III} state. The X-rays emitted are the familiar $K\alpha_2$ and $K\alpha_1$ radiation, respectively. The

transition from the K state to the L_I state is forbidden by selection rules. Moreover, an initial double ionization of an atom can also result in the emission of X rays. For example, the $K\alpha_{3,4}$ emission is the satellite spectrum produced by a transition from a KL state to an LL state, where the KL state means that the absorbed X-ray energy is sufficient to eject a K and an L electron (3,4).

The absorption and emission of X-ray photons by atoms in fluorescent processes can be rigorously described by using the quantum theory of radiation (7). The vector potential $\vec{A}(\vec{r},t)$ that represents the electromagnetic radiation (8) is usually expressed as a function of the position vector $\vec{r}$ and time t (5–7):

$$\vec{A}(\vec{r},t) = \left(\frac{c^2 N\hbar}{\omega V}\right)^{1/2} \hat{\varepsilon} e^{i(\vec{k}*\vec{r}-\omega t)} + \left[\frac{c^2(N+1)\hbar}{\omega V}\right]^{1/2} \hat{\varepsilon} e^{-i(\vec{k}*\vec{r}-\omega t)}, \quad (1)$$

where the first term is related to the absorption of a photon quantum and the second term to the emission of a photon quantum by an electron. N is the photon occupation number in the initial state, $\hat{\varepsilon}$ is the unit vector of polarization, and $\vec{k}$ and ω are the wave vector and angular frequency of the EM wave. c, $\hbar$, and V are the speed of light in vacuum, Planck's constant $h/2\pi$, and the volume irradiated, respectively. The absorption probability, the transition probability, and the differential cross section can be calculated quantum mechanically by considering the fluorescence as the scattering of photons by atomic electrons (4–7). Before scattering, the photon, whose wavevector is $\vec{k}$, angular frequency is ω, and polarization vector is $\hat{\varepsilon}$, is incident on the atom at its initial state A. After scattering, the atom is left in its final state B, and the scattered photon of the polarization vector $\hat{\varepsilon}'$ and angular frequency ω' propagates along the $\vec{k}'$ vector. In between are the excited intermediate states. The Kramers–Heisenberg formula, modified to include the effects of radiation damping, gives the differential cross section, the derivative of the cross section σ with respect to the solid angle Ω, as follows (7):

$$\frac{d\sigma}{d\Omega} = r_O^2 \left(\frac{\omega'}{\omega}\right) \left| \hat{\varepsilon} * \hat{\varepsilon}' \delta_{AB} - \frac{1}{m} \sum_I \left[\frac{(\vec{p} * \hat{\varepsilon}')_{BI} (\vec{p} * \hat{\varepsilon})_{IA}}{E_I - E_A - \hbar\omega - i\frac{\Gamma_I}{2}} + \frac{(\vec{p} * \hat{\varepsilon})_{BI} (\vec{p} * \hat{\varepsilon'})_{IA}}{E_I - E_A + \hbar\omega} \right] \right|^2, \quad (2)$$

where r_O is the classic radius of the electron, $r_O = (e^2/mc^2)$, and the subscript I stands for the excited intermediate state. $\vec{p}$ is the electric dipole moment, and φ is the wave function of a corresponding state. δ_{AB} is a Kronecker delta that involves the wave functions φ_A and φ_B of states A and B. The matrix elements $(\vec{p} * \hat{\varepsilon})_{BI}$ and $(\vec{p} * \hat{\varepsilon})_{IA}$ involve the wave functions, φ_A, φ_B, and φ_I (7). The probability of finding the intermediate state I is proportional to $\exp(-\Gamma_I t/\hbar)$, where $\Gamma_I = \hbar/\tau_I$ and τ_I is the lifetime of state I. The first term in Eq. (2) represents the nonresonant amplitude of scattering ($\omega \neq \omega_{IA} = (E_I - E_A/\hbar)$). The second and third terms have appreciable amplitude in the resonant condition, that is $\omega = \omega_{IA}$. This phenomenon is known as resonant fluorescence.

The sum of the nonresonant amplitudes is usually of the order of r_O, which is much smaller than the resonant amplitude (of the order of c/ω). By ignoring the nonresonant amplitudes, the differential cross section of a single-level resonant fluorescent process in the vicinity of a nondegenerate resonance state R takes the form (7)

$$\frac{d\sigma}{d\Omega} = \frac{r_O^2}{m^2} \left(\frac{\omega'}{\omega}\right) \left[\frac{|(\vec{p} * \hat{\varepsilon})_{AR}|^2}{(E_R - E_A - \hbar\omega)^2 + \Gamma_R^2/4} \right] |(\vec{p} * \hat{\varepsilon'})_{RB}|^2. \quad (3)$$

This expression is related to the product of the probability of finding state R formed by the absorption of a photon characterized by $(\vec{k}, \hat{\varepsilon})$ and the spontaneous emission probability per solid angle for the transition from state R to state B characterized by the emitted photon $(\vec{k}', \hat{\varepsilon}')$. The term in the square bracket is proportional to the absorption probability, and the matrix element after the bracket is connected to the emission probability.

For a conventional X-ray and a synchrotron radiation (SR) source, the energy resolution ΔE of an incident X-ray beam is around 10^{-2} to 10 eV (2,3,9,10), except that for some special experiments, such as those involving Mössbauer effects, 10^{-9} eV is needed (11,12). Under normal circumstances, $\Delta E \gg \Gamma_R$ ($\sim 10^{-9}$ eV) according to the uncertainty principle. This means that the temporal duration of the incident photon (X-ray) beam is shorter than the lifetime of the resonance. In other words, the formation of the metastable resonance state R via absorption and the subsequent X-ray emission can be treated as two independent quantum mechanical processes. In such a case, the emission can be thought of as a spontaneous emission, that is, the atom undergoes a radioactive transition from state R to state B as if there were no incident electromagnetic wave. Then, the corresponding $d\sigma/d\Omega$ takes the following simple form under the electric dipole approximation (7):

$$\frac{d\sigma}{d\Omega} \cong P_{AB} \frac{\alpha\omega^3}{2\pi c^3} V |\overrightarrow{X_{AB}}|^2 \sin^2\theta, \quad (4)$$

where P_{AB} is the absorption probability equal to the term in the square bracket given in Eq. (3), and $|X_{AB}|^2$ is the transmission probability equal to the term after the bracket of Eq. (3). α is the fine-structure constant ($\sim 1/137$). θ is the angle between the electric dipole moment $\vec{p}$ and the direction of the X-ray fluorescence emitted from the atom. As is well known, this type of fluorescence has the following distinct properties: (1) In general, fluorescence is directionally isotropic, so that its distribution is uniform in space. (2) The intensity of fluorescence is maximum when the emission direction is perpendicular to $\vec{p}$ and is zero along $\vec{p}$. (3) X-ray fluorescence emitted from different atoms in a sample is incoherent, and the emitted photons will not interfere with each other. (4) Whenever the energy of an incident photon is greater than an absorption edge, the energy of the emitted fluorescence is determined by the energy difference between the two states involved in the transition, which is independent of the incident energy.

X-ray fluorescence has diverse applications, including identifying elements in materials and quantitative and

qualitative trace-element analyses from X-ray fluorescent spectra. Together with other imaging techniques, X-ray fluorescent imaging can, in principle, provide both spectral and spatial distributions of elements in matter in a corresponding photon-energy range. Unlike other fluorescent imaging techniques, such as optical fluorescence and laser-induced fluorescent imaging, the development and application of X-ray fluorescent imaging is still in a state of infancy, though X-ray fluorescence and related spectroscopy have been known for quite sometime. Yet, due to the advent of synchrotron radiation, new methods and applications using this particular imaging technique have been developed very recently.

An X-ray fluorescent image is usually obtained by following these steps: The incident X-ray beam, after proper monochromatization and collimation, impinges on a sample of interest. An absorption spectrum is acquired by scanning the monochromator to identify the locations of the absorption edges of the element investigated. The incident beam is tuned to photon energies higher than the absorption edges so that the constituent elements of the sample are excited. Then, X-ray fluorescence emanates. The spatial distribution of the fluorescent intensity, the fluorescent image, is recorded on a two-dimensional detector or on a point detector using a proper scan scheme for the sample. The following items need to be considered for good quality images of high spatial and spectral resolution: (1) appropriate X-ray optics for shaping the incident beam and for analyzing the fluoresced X rays and (2) a suitable scanning scheme for the sample to cross the incident beam. In addition, image reconstruction of intensity data versus position is also important for direct mapping of the trace elements in a sample.

TECHNICAL ASPECTS

An X-ray fluorescent imaging setup consists of an X-ray source, an X-ray optics arrangement, a sample holder, an imaging detector, and an image processing system. These components are briefly described here:

X-ray Sources

Conventional X-ray sources are sealed X-ray tubes and rotating-anode X-ray generators. Synchrotron radiation emitted from accelerated relativistic charged particles is another X-ray source. The spectra of conventional sources consist of characteristic lines according to specific atomic transitions and white radiation background due to the energy loss from the electrons bombarding the metal target used in X-ray generators (2,3). The energy resolution of the characteristic lines is about $\Delta E/E \approx 10^{-4}$. In contrast, the spectrum of synchrotron radiation is continuous and has a cutoff on high energies, resulting from the emission of a collection of relativistic charged particles (9). High directionality (parallelism), high brilliance, well-defined polarization (linear polarization in the orbital plane and elliptical polarization off the orbital plane), pulsed time structure (usually the pulse width is in picoseconds and the period in nanoseconds), and a clean environment (in vacuum of 10^{-9} torr) are the advantages of synchrotron X-ray sources (9). Other sources such as plasma-generated X rays can also be used.

Beam Conditioning

Proper X-ray monochromators need to be employed for monochromatic incident X-rays. Different incident beam sizes for different imaging purposes, can be obtained by using a suitable collimation scheme. Beam conditioning for X-ray fluorescent imaging is similar to that for X-ray diffraction and scattering. Single- or double-crystal monochromators (10) are usually employed for beam monochromatization. For synchrotron radiation, grazing incidence mirrors before or after the monochromator for focusing or refocusing, respectively, are used (13,14). Pinhole or double-slit systems and an evacuated beam path have been found useful for beam collimation. To obtain a microbeam, two double-crystal monochromators — one vertical and the other horizontal, or multilayered mirrors such as the Kirkpatrick–Baez (K–B) configuration (15) are reasonable choices for the collimation system.

Sample Holder

Usually a sample holder is designed to facilitate the positioning of the sample with respect to the incident X-ray beam. In general, the sample holder should provide the degree of freedom needed to translate the sample left and right and also up and down. In addition, in some cases, the sample needs to be rotated during the imaging process. Therefore, the azimuthal rotation around the sample-surface normal and rocking about the axis perpendicular to both the surface normal and the incident beam are indispensable. Accuracy in translation is of prime concern in high spatial resolution fluorescent imaging. If soft X-ray fluorescent imaging is pursued, a sample holder in a high vacuum (HV) or ultrahigh vacuum (UHV) condition is required to reduce air absorption.

Detector

A suitable detector is essential in studies/investigations for recording X-ray fluorescent images. Area detectors or two-dimensional array detectors are usually employed to obtain two-dimensional images. They include imaging plates (IP) (a resolution of 100 μm for a pixel), charge coupled device (CCD) cameras (a resolution of about 70–100 μm), microchannel plates (MCP) (25 μm per pixel), and solid-state detectors (SSD) (an energy resolution of about 100–200 eV). Traditionally, point detectors such as the gas proportional counter and NaI(Li) scintillation counter have been used frequently. However, to form images, proper scan schemes for the sample as well as the detector are required. For hard X-ray fluorescent imaging, an imaging plate, CCD camera, and point detectors serve the purpose. For soft X-ray fluorescence imaging, a vacuum CCD camera, microchannel plate, and semiconductor pin-diode array may be necessary.

Computer-Aided Image Processor

The fluorescent signals collected by a detector need to be stored according to the sample position. Usually analog signals, are converted into digital signals, and

this digital information is then stored in a computer. Image processing, which includes background subtraction, image-frame addition, and cross-sectional display, can be carried out on a computer. In addition, black-and-white contrast as well as color images can be displayed on a computer monitor and on a hard copy. In some cases when the rotating-sample technique is employed to collect a fluorescent image, corrections to avoid spatial distortion need to be considered (16,17).

Image reconstruction depends on the scan mode chosen. There are three different modes scan: the point scan, line scan, and area scan. Usually the data from a point scan give a one to one spatial correspondence. Therefore, there is no need to reconstruct the image using mathematical transformation techniques, provided that high-resolution data are collected. Because the areas covered by the incident beam during a line scan (involving a line beam and sample translation and rotation) and an area scan are larger than the sample area, these two types of scans require image reconstruction from the measured data. The estimated relative errors σ_{rel} in the reconstructed images as a function of the percentage of the sample area a % that provide fluorescent intensity are shown in Fig. 1 for the three different scan types without (Fig. 1a) and with (Fig. 1b) a constant background. Except for very small fluorescent areas, the line scan provides the lowest error (18,19). Images from both the line scan and area scan can be reconstructed by employing appropriate filter functions and convolution techniques (16,20).

The following representative cases illustrate in detail the general basic experimental components necessary for X-ray fluorescent imaging. The required experimental conditions specific for the investigation are given, and the actual images obtained are shown.

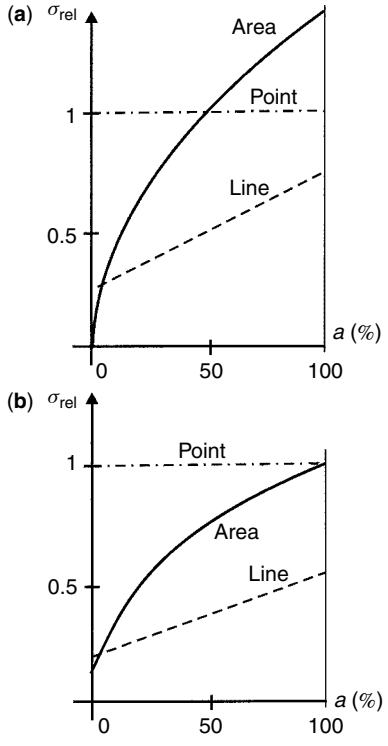


Figure 1. Estimated error σ_{rel} in reconstructed image vs. signal area a% for point, line, and area scans: (**a**) with, (**b**) without constant background (19) (courtesy of M. Bavdaz et al.; reprinted from *Nucl. Instrum. Methods* **A266**, M. Bavdaz, A. Knochel, P. Ketelsen, W. Peterson, N. Gurker, M. H. Salehi, and T. Dietrich, Imaging Multi-Element Analysis with Synchrotron Radiation Excited X-ray Fluorescence Radiation, pp. 308–312, copyright 1988, Elsevier Science with permission).

CASE STUDIES

Nondestructive X-ray Fluorescent Imaging of Trace Elements by Point Scanning

The two-dimensional spatial distribution of multiple elements in samples is usually observed by X-ray fluorescent imaging using synchrotron radiation, mainly because of the tunability of synchrotron photon energy (21). Proper photon energy can be chosen for a specific element investigated. In addition, the brightness and directionality of synchrotron radiation provide superb capability for microbeam analysis, which can improve the spatial resolution of imaging approximately 100-fold. The minimal radiation damage of synchrotron radiation to specimens is another advantage over charged particles (22) such as ions and electrons. A typical experimental setup (23) for this purpose is shown schematically in Fig. 2. Polychromatic white X-rays from synchrotron storage ring are monochromatized for a selected wavelength using either a Si (111) cut double-crystal monochromator (DCM) or a Si–W synthetic multilayer monochromator (SMM). Using the DCM, the incident photon energy can be varied continuously by rotating the crystal around the horizontal axis perpendicular to the incident beam. Using the SMM, the photon energy can be changed by tilting the multilayer assembly about the vertical axis perpendicular to the incident beam. However, the entire system of the SMM may need rearrangement. Cooling the SMM may be necessary, and this should be taken into consideration. A multilayer substrate is usually attached to a copper block, which can be water-cooled. The energy resolution $\Delta E/E$ is about 10^{-4} for the

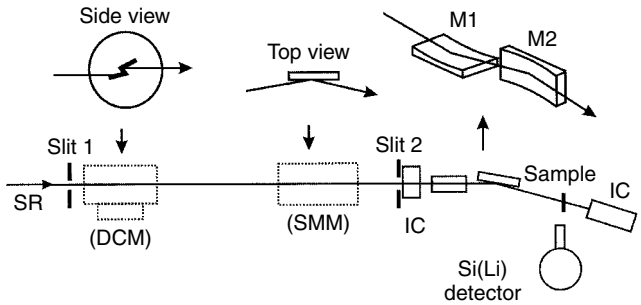


Figure 2. Schematic representation of the experimental setup for synchrotron radiation (SR) fluorescent imaging: The incident SR is monochromatized by either a double-crystal monochromator (DCM) or a synthetic multilayer (SMM) and is then focused by the K–B mirror system (M1 and M2). An ionization chamber (IC) and a Si(Li) SSD are used to monitor the incident beam and the fluorescence (23) (courtesy of A. Iida et al.; reprinted from *Nucl. Instrum. Methods* **B82** A. Lida and T. Norma, Synchrotron X-ray Microprobe and its Application to Human Hair Analysis, pp. 129–138, copyright 1993, Elsevier Science with permission).

DCM and 10^{-2} for the SMM. A double-slit system is usually used to trim the incident beam to a desired size, say from 750×750 μm to as small as 3×4 μm for point scans. The intensity of the incident beam is monitored by an ionization chamber (IC) as usual for synchrotron experiments. A Kirkpatrick–Baez mirror system that consists of a vertical mirror M1 and a horizontal mirror M2 is placed after the IC to modulate the beam size for spatial resolution and the photon flux for sensitivity. Usually, elliptical mirrors are used to reduce spherical aberration (14,24). A sample is located on a computer-controlled X,Y translation stage. The translation step varies from 5 to 750 μm. Smaller steps can be achieved by special design of the translation stage. The fluorescence emitted from the sample at a given position is detected by a Si (Li) solid-state detector. The fluorescent intensity as a function of the sample position (x, y) is recorded and displayed on a personal computer. For illustration, the following examples are X-ray fluorescent images obtained by using this specific setup.

X-ray Fluorescent Images of Biological Tissues

Two-dimensional Distribution of Trace Elements in a Rat Hair Cross Section (25). Trace element analysis of hair samples has been widely employed for biological monitoring of health conditions and for environmental investigations of heavy metal exposure and contamination (26,27). Consider a heavy toxic metal such as Hg, which has already been taken up from the blood to the hair of animals and humans due to exposure to methylmercury (MeHg) — a commonly encountered and widely used form of environmental mercury. It is known that MeHg damages the central nervous system (28) by penetrating the blood–brain barrier (29). The effects on and alteration of biological systems and the dynamics of the distribution of Hg in the organs of animals exposed to MeHg have long been investigated (21,30). Because hair specimens usually provide a historical record of trace elements and are easily accessed, hair is a good bioindicator of metal contamination in mammals. The following example involves Hg distribution in rat hair.

A hair sample of rats endogenously exposed to methylmercury (MeHg) (internal administration) was subjected to synchrotron X-ray fluorescent measurement at the incident photon energy of 14.28 keV, using an experimental setup similar to that shown in Fig. 2. The beam size was 5×6 μm^2. Figure 3 shows the fluorescent spectrum of the hair that was collected in 50 minutes. The presence (contamination) of Hg, Zn, and S is clearly seen from the presence of Hg Lα, Hg Lβ, Zn Kα, and S Kα lines in the spectrum.

The elemental images of Hg Lα, Zn Kα, and S Kα from a hair cross section of the MMC (methylmercury chloride)-treated rat 13.9 days after the first administration are shown in Fig. 4. The sample was obtained at 1 mm

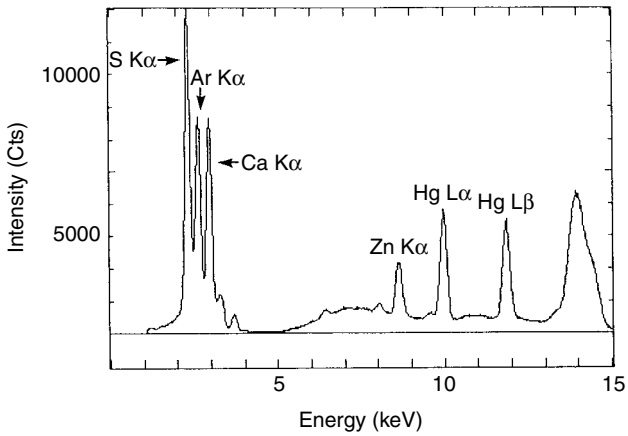


Figure 3. X-ray fluorescence spectrum obtained from the hair cross section of a rat given MMC (25) (courtesy of N. Shimojo et al.; reprinted from *Life Sci.* **60**, N. Shimojo, S. Homma-Takeda, K. Ohuchi, M. Shinyashiki, G.F. Sun, and Y. Kumagi, Mercury Dynamics in Hair of Rats Exposed to Methylmercury by Synchrotron Radiation X-ray Fluorescence Imaging, pp. 2,129–2,137, copyright 1997, Elsevier Science with permission).

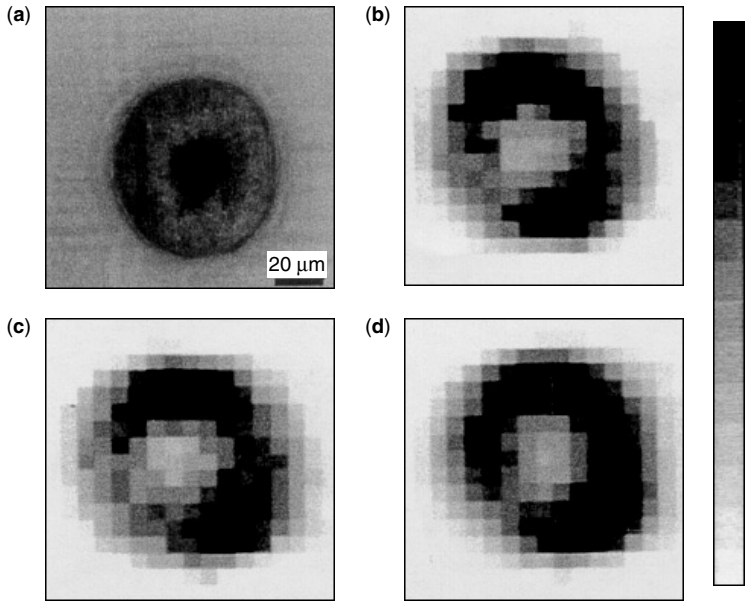


Figure 4. (**a**) Optical micrograph and (**b**) X-ray fluorescent images of Hg Lα, (**c**) Zn Kα, and (**d**) S Kα from a hair of a MMC-treated rat (25) (courtesy of N. Shimojo et al.; reprinted from *Life Sci.* **60**, N. Shimojo, S. Hommo-Takeda, K. Ohuchi, M. Shinyashiki, G. K. Sun, and Y. Kumagi, Mercury Dynamics in Hair of Rats Exposed to Methylmercury by Sychrotron Radiation X-ray Fluorescence Imaging, pp. 2,129–2,137, copyright 1997, Elsevier Science with permission).

from the root end of the hair. The sample thickness was about 25 μm. The scanning condition was 16×16 steps at 5 μm/step, and the counting time was 10 s per step. An optical micrograph of the hair cross section is also shown for comparison (Fig. 4a). The black-and-white contrast, classified into 14 degrees from maximum to minimum, represents linearly the 14 levels of the element concentrations. Figure 4c clearly shows that the element Zn is localized in the hair cortex rather than the medulla after MMC administration. In addition, endogenous exposure to MMC results in preferential accumulation of Hg in the hair cortex than in the medulla or cuticle (Fig. 4b). However, for hair under exogenous exposure to MMC (external administration), detailed analysis of the fluorescent images, together with the fluorescent spectra across the hair cross section, indicates that Hg is found distributed on the surface of the hair cuticle rather than the cortex (25). This result is consistent with other element characterization using flameless atomic absorption spectrometry (FAAS). Similarly, the Hg concentration profile along the length of a single hair can also be determined by using this imaging technique.

Human hair can also be analyzed by using the same X-ray fluorescent imaging technique (23). The distribution of the elements S, Ca, Zn, Fe, and Cu in a hair cross section, similar to Fig. 3, can be determined similarly. The concentrations of these elements measured by the quantitative analysis of fluorescent intensity are comparable with the values obtained by other techniques (31–33).

This imaging technique can be also applied to dynamic studies of metal contamination in hair and blood because the samples prepared can be repeatedly used due to negligibly small radiation damage. For example, fluorescent images of hair cross sections cut at every 1 mm from the root end of MMC-treated rats show that the Hg concentration, distributed in the center of cortex, increases first to about 1,000 μg/g for about 10 days after MMC administration. Then, the concentration decreases to about 600 μg/g for another 10 days (25). Thus, the history of Hg accumulation in animals can be clearly revealed.

In summary, the X-ray fluorescent imaging technique is useful in revealing the distribution and concentrations of metal elements in hair and also in providing dynamic information about the pathway of metal exposure in animals and the human body.

Distribution of Cu, Se, and Zn in Human Kidney Tumors. Copper (Cu), selenium (Se), and zinc (Zn) are important metal cofactors in metalloenzymes and metalloproteins. Their presence in these enzymes and proteins directly influences many biochemical and physiological functions. As is well known, ceruloplasmin (34) and dopamine-β-hydroxylase (35) are involved in iron metabolism and neurotransmitter biosynthesis, respectively. Both contain Cu. Se is an essential component of glutathione peroxidase, which plays an important role in protecting organisms from oxidative damage via the reduction of lipoperoxides (R–O–OR) and hydrogen peroxide (36). Zn is also an important metal element in DNA polymerase and appears in many enzymes such as carbonic anhydrase, alcohol dehydrogenase, and alkaline phosphatase (37,38).

Recent studies of the physiological roles of these essential trace elements have been emphasized in connection with possible causes of cancer. For example, investigations of the blood serum levels of these trace elements in cancer patients show the possible involvement of these elements in many cancerous conditions (39–44). In particular, serum levels of Cu and Zn and Cu/Zn ratios of patients who have malignant neoplasms have been used as an indicator for assessing disease activities and prognoses. Increased serum Cu levels and decreased serum Zn levels have also been found in patients who have sarcomas (39), lung cancer (40), gynecologic tumors (41), and carcinoma of the digestive system (stomach) (42). However, only a limited number of works in the literature, are concerned with the distribution of Cu, Se, and Zn in malignant neoplasms. In the following, two-dimensional distributions of Cu, Se, and Zn in human kidney tumors were determined and visualized using nondestructive synchrotron radiation X-ray fluorescent imaging (45).

The distribution of Cu, Se, and Zn in cancerous and normal renal tissues and the correlation among these distributions were studied. The experimental setup shown in Fig. 2 and an incident photon energy of 16 keV were used. The experimental conditions were 750×750 μm for the beam size, 750 μm/step for the sample translation, and 10 s counting time for each position.

The fluorescence of Cu Kα, Zn Kα, and Se Kα were identified in X-ray fluorescent spectra. Compare the optical micrograph shown in Fig. 5a to the spatial distributions of fluorescent intensities of Cu Kα, Zn Kα, and Se Kα for normal (N) and cancerous (C) renal tissues shown in Fig. 5b, c, and d, respectively. The frame of

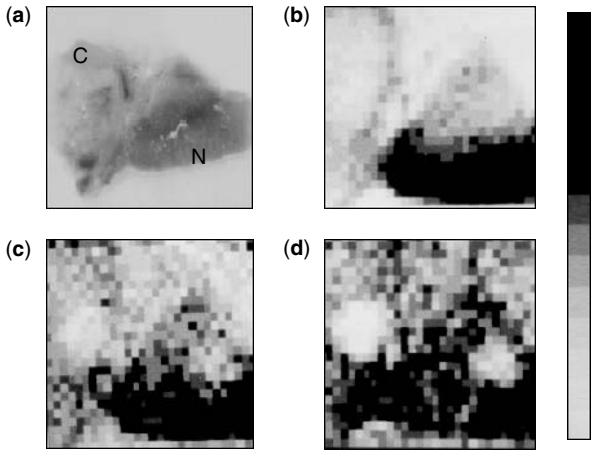


Figure 5. Chemical imaging of trace elements in normal (N) and cancerous (C) renal tissue from an aged female: (**a**) Optical micrograph of a sliced sample and the distributions of (**b**) Zn, (**c**) Cu, and (**d**) Se (45) (courtesy of S. Homma-Takeda et al.; translated from *J. Trace Elements in Exp. Med.* **6**, S. Homma, A. Sasaki, I. Nakai, M. Sagal, K. Koiso, and N. Shimojo, Distribution of Copper, Selenium, and Zinc in Human Kidney Tumours by Nondestructive Synchrotron X-ray Fluorescence Imaging, pp. 163–170, copyright 1993, Wiley-Liss Inc. with permission of John Wiley & Sons, Inc. All rights reserved).

the fluorescent images covers the entire tissue sample. The upper left portion of the frame corresponds to the cancerous tissue, and the low right portion corresponds to the normal tissue. Darker pixels indicate positions of higher metal concentration. The images reveal that Cu, Zn, and Se accumulate more densely in normal tissue than in cancerous tissue. The concentrations of these metals in the samples determined by ICP-AES (inductively coupled plasma atomic emission spectrometry) agree qualitatively with those estimated from the intensities of the fluorescent images. The average Zn concentration in the cancerous tissues is 12.30 ± 5.05 μg/g compared to 19.10 ± 10.19 μg/g for the normal tissues. The average concentration of Cu is 0.991 ± 0.503 μg/g in the cancerous tissues, compared to 17.200 ± 0.461 μg/g in the normal tissues. The decreases in Zn and Cu concentrations in the cancerous tissues are statistically significant across hundreds of measured data sets. Moreover, the correlation coefficients among the distributions of trace elements can be calculated from the X-ray intensity data at each analytical point. In general, it is found that the correlation coefficients among the metal elements Cu, Se, and Zn in cancerous tissues are qualitatively lower than those in normal tissues. However, the correlation between Cu and Zn in the cancerous tissue investigated is significantly decreased, more than 30%, compared with that in the normal tissues.

It should be noted that the change in the trace element level in cancerous tissues is not the same for all types of tumors. Moreover, Zn levels in cancerous tissues vary in different organs. According to (46), tumors located in organs that normally exhibit a low Zn concentration have a Zn accumulation that is similar to or greater than that in the tissue around the tumor. Tumors located in organs that normally have a high Zn concentration also, exhibit a lower uptake of Zn than tissues not involved in the growth of the tumor. Kidney is an example of an organ that has a high Zn concentration. Therefore, the observation of decreased Zn concentration in the kidney tumor is expected.

The age dependence of the distribution of these metal elements in human kidney can also be observed in X-ray fluorescent images (47). Figures 6a–d and 6e–h show the X-ray fluorescent images of Zn, Cu, and Se in adult human kidneys from a 22-year-old and a 61-year-old man, respectively. The same experimental setup and conditions as those used for imaging the kidney tumor were employed. The exposure time was 6 s/point for the sample from the 22-year-old man and 10 s/point from the 61-year-old man. Clearly, Cu, Zn, and Se are more concentrated in the renal cortex than in the medulla. This result agrees with that reported in (48–50) and is consistent with the functions of each tissue that is, the kidney cortex contains glomeruli, responsible for cleaning waste materials from blood, and the proximal tubule is responsible for reabsorbing metal ions.

The elemental concentrations of Cu and Zn in the kidney tissue determined by both ICP-AES and X-ray fluorescent analysis are 1.67 ± 0.22 μg/g (Cu) and 13.3 ± 2.6 μg/g (Zn) for the 22-year-old man and 1.06 ± 0.34 μg/g (Cu) and 4.42 ± 1.52 μg/g (Zn) for the 61-year-old man. The correlation coefficients between the two trace elements calculated from the X-ray intensity data indicate that the correlation between Zn and Cu is higher for the 22-year-old man than for the 61-year-old man. Moreover, the correlation between Cu and Se is the lowest among the Zn–Cu, Zn–Se, and Cu–Se correlations (47).

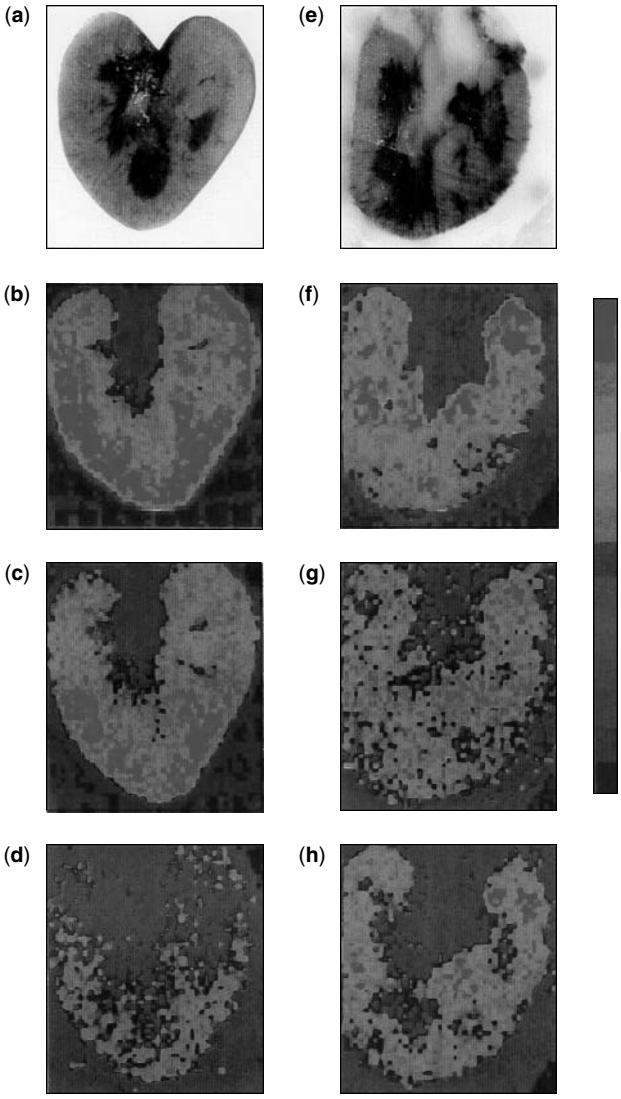


Figure 6. X-ray fluorescent images of trace elements in adult human kidney: (**a**) photograph of a sliced sample and the distributions of (**b**) Zn, (**c**) Cu, (**d**) Se for a 22-year-old man, and (**e**) photograph and the distributions of (**f**) Zn, (**g**) Cu, and (**h**) Se for a 61-year-old man (field width 750×750 μm^2; medulla: the upper central region; cortex: the periphery) (47) (courtesy of S. Homma-Takeda et al.; reprinted from *Nucl. Instrum Methods* **B103**, S. Homma, I. Nakai, S. Misawa, and N. Shimojo, Site-Specific Distribution of Copper, Selenium, and Zinc in Human Kidney by Synchrotron Radiation Induced X-ray fluorescence, pp. 229–232, copyright 1995, Elsevier Science with permission).

In summary, the X-ray fluorescent imaging technique is well suited for evaluating levels of trace elements in human organs that may influence smoking, cancer, or other diseases. Consequently, the physiological roles of trace elements could be better understood. Furthermore, using synchrotron radiation, this X-ray fluorescent

imaging technique causes negligible damage to samples. As a result, samples can be examined histologically after analysis. This fact makes X-ray fluorescent imaging a useful technique for histochemical analysis of biological specimens.

The X-ray fluorescent imaging technique can be used in combination with other techniques, such as isoelectric focusing agarose gel electrophoresis (IEF-AGE), histochemical staining, and X-ray microprobes, to obtain functional and structural information about biological substances. For example, using IEF-AGE, direct detection of structural or functional alteration of protein attributed to the binding of mercury (51) and efficient monitoring of changes in mRNA levels, protein contents, and enzyme activities for brain Cu, Zn-, and Mn-SOD (superoxide dismutase) by MMC administration (52) are feasible. Furthermore, detailed distribution of metal elements and morphological changes in biological specimens that are histochemically stained (53), and high-resolution fluorescent images without tedious sample preparation and treatment using X-ray microprobes (54) can also be obtained.

X-ray Fluorescent Imaging by Line Scanning

This example is included to demonstrate the X-ray fluorescence imaging of a Ni-wire cross by synchrotron radiation using line scans. Images different from those obtained by point scans are expected. The schematic of the experimental setup for line scanning is shown in Fig. 7 (19). The incident beam is collimated by a beam aperture (1) and focused by a mirror (2). An additional guard aperture (3) is then placed between the linear slit (4) and the mirror to cut down the parasitic scattering from the first aperture and the mirror. A sample aperture (5) in front of the sample holder (6) prevents X-rays scattered by the linear slit from participating in the sample excitation. The sample holder can be rotated around the incident beam and translated across the line-shaped incident beam. A solid-state detector (7) is used to monitor the fluorescence that emanates from the sample. During excitation, the sample, a Ni-wire cross, is translated vertically and then rotated around the incident beam. Because a monochromator is not used, both Ni Kα and

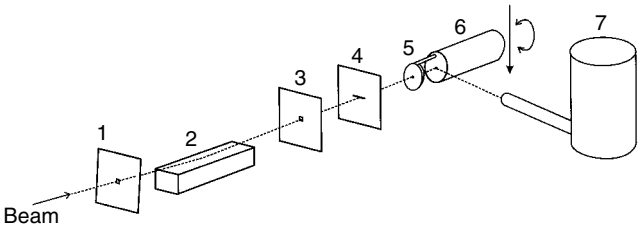


Figure 7. Experimental setup for X-ray fluorescent imaging using line scan: (1) aperture, (2) mirror, (3) antiscattering stop, (4) slit-aperture, (5) aperture-stop, (6) sample holder, (7) solid-state detector (19) (courtesy of M. Bavdaz et al.; reprinted from *Nucl. Instrum. Methods* **A266**, M. Bavdaz, A. Knochel, P. Ketelsen, W. Petersen, N. Gurker, M. H. Salehi, and T. Dietrich, Imaging Multi-Element Analysis with Sychronotron Radiation Excited X-ray Fluorescence Radiation, pp. 308–312, copyright 1998, Elsevier Science with permission).

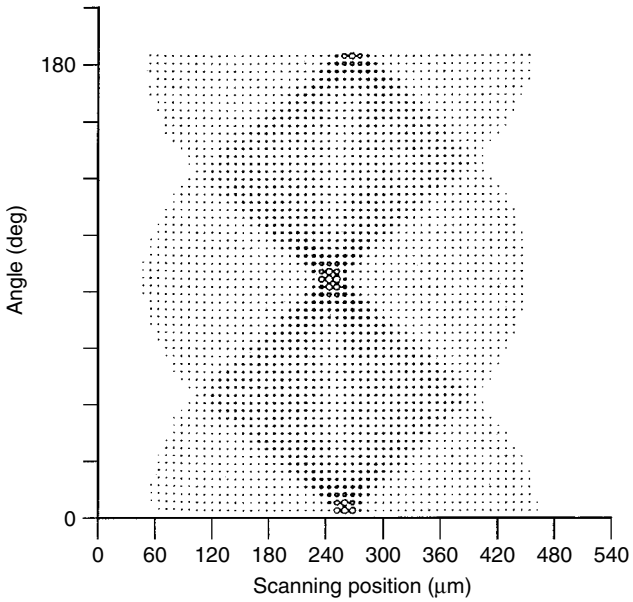


Figure 8. Measured Ni K fluorescent intensity distribution of a Ni-wire cross (19) (courtesy of M. Bavdaz et al.; reprinted from *Nucl. Instrum. Methods* **A266**, M. Bavdaz, A. Knochel, P. Ketelsen, W. Petersen, N. Gurker, M. H. Salehi, and T. Dietrich, Imaging Multi-Element Analysis with Sychronotron Radiation Excited X-ray Fluorescence Radiation, pp. 308–312, copyright 1998, Elsevier Science with permission).

Ni Kβ are excited. Figure 8 shows the two-dimensional Ni Kα and Ni Kβ fluorescent distribution measured at a translation of 9 μm per step and 3° per step in rotation. Clearly, the measured distribution is far from the real image of the cross, which can usually be obtained by point scans. The spatial resolution of Fig. 8 depends on the sizes of the apertures and slits used and the distances between the sample, the apertures, and the detector. The accuracy in rotating the sample is also an important factor. Using this kind of setup, 10-μm resolution can easily be achieved. By applying the filtered back-projection technique (20) for image reconstruction from the line scanned distribution, the final reconstructed image of the Ni cross shown in Fig. 9 is obtained. This fluorescent image from linear scanning is equally applicable to imaging trace elements for various types of specimens (16).

Soft X-ray Fluorescent Imaging at Submicron Scale Resolution

Spatial resolution of soft X-ray fluorescent imaging depends in some cases on the fluorescent material used in making the screen for image recording and display. Among the various materials available for making fluorescent screens, polycrystalline phosphor is most frequently used to detect X rays and electrons. The required features of a material such as phosphor for imaging are high conversion efficiency and high spatial resolution. However, polycrystalline phosphor (powder) usually suffers from particle size and the scattering of fluorescent light among powder particles, which limit the spatial resolution ($\geq$1 μm) and the resultant conversion efficiency. To improve the optical quality of fluorescent

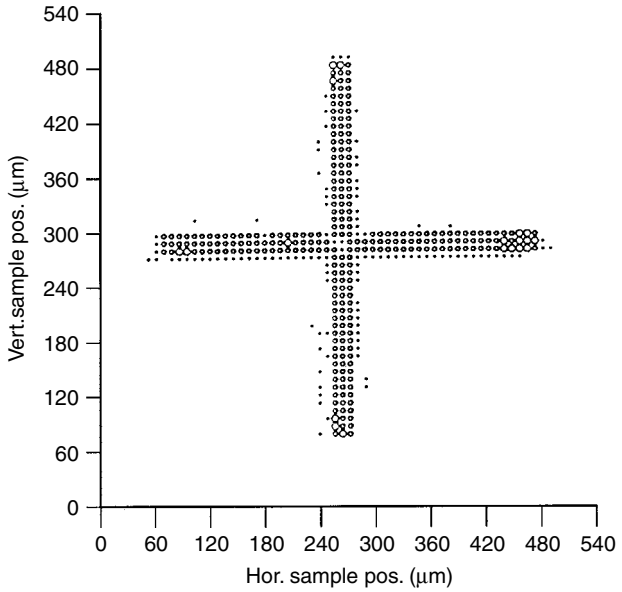


Figure 9. Reconstructed image of a Ni-wire cross (19) (courtesy of M. Bavdaz et al.; reprinted from *Nucl. Instrum. Methods* **A266**, M. Bavdaz, A. Knochel, P. Ketelsen, W. Petersen, N. Gurker, M. H. Salehi, and T. Dietrich, Imaging Multi-Element Analysis with Sychronotron Radiation Excited X-ray Fluorescence Radiation, pp. 308–312, copyright 1998, Elsevier Science with permission).

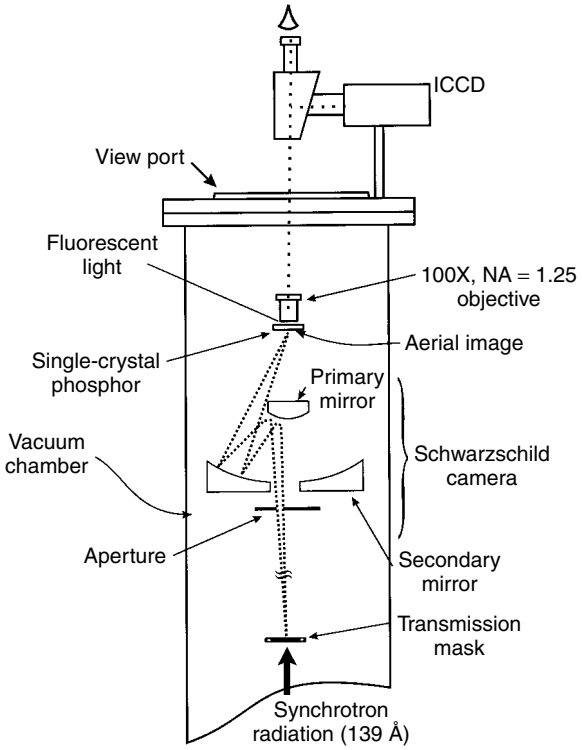


Figure 10. Experimental setup for submicron soft X-ray fluorescent imaging (58) (courtesy of B. LaFontaine et al.; reprinted from *Appl. Phys. Lett.* **63**, B. La Fontaine, A. A. MacDowell, Z. Tan, D. L. Taylor, O. R. Wood II, J. E. Bjorkholm, D. M. Tennant, and S. L. Hulbert, Submicron, Soft X-ray Fluorescence Imaging, pp. 282–284, copyright 1995, American Institute of Physics with permission).

screen materials, the use of a plastic scintillator, or doped polymer films, has been suggested to provide better spatial uniformity. However, the efficiency of this material is lower than that of phosphor powder (55). Recently, phosphor single crystals have been grown by liquid-phase epitaxy (LPE) (56,57). The resulting monocrystalline phosphor possesses excellent optical quality and high conversion efficiency. The characteristics of a short absorption depth of soft X-rays in the phosphor (200 Å for $\lambda = 139$ Å) provide a well-defined and localized fluorescent image without any smearing. The high efficiency of conversion and the high sensitivity of the phosphor crystal that covering the emission spectrum in the soft X-ray regime afford low light level detection. Moreover, because the single-crystal phosphor is transparent to emitted light and possesses a superpolished crystal surface, deleterious scattering effects are eliminated, and a well-defined perfect X-ray image plane is achieved. Collectively, these superb optical properties of the single-crystal phosphor lead to soft X-ray fluorescent imaging at submicron resolution, which is described here (58).

The imaging experiment was performed by using soft X rays from a synchrotron storage ring. The schematic representation of the experimental setup is given in Fig. 10. It comprises a 20X reduction Schwarzschild camera working at 139 Å, a fluorescent crystal, an optical microscope, and an intensified charge-coupled device (ICCD) camera. The Schwarzschild camera contains two highly reflecting mirrors that have multilayer coatings. The main component of the optical microscope is an oil immersion 100X objective that has a numerical aperture of 1.25. Referring to Fig. 10, an X-ray image of a transmission mask is projected onto the image plane of the Schwarzschild soft X-ray camera, at which a commercially available phosphor crystal [a STI-F10G crystal, manufactured by Star Tech Instruments Inc. (58)] is positioned to convert the soft X-ray image into the visible. This visible image is well localized and easily magnified by the optical microscope. The magnified image, located at about 11 cm from the crystal outside the vacuum chamber, is then magnified with a 5X long-working-distance microscope and viewed by the ICCD camera.

The X-ray fluorescent spectrum of the fluorescent crystal, STI-F10 G, excited by soft X rays from 30 to 400 Å is shown in Fig. 11. The broad fluorescence in the emission spectrum is useful for uniform image mapping. Figure 12 illustrates the submicron-resolution fluorescent image (Fig. 12a) of a mask, which can be compared with the actual mask shown in Fig. 12b. As can be seen, all of the lines whose widths range from 0.5 to 0.1 μm after the reduction due to the Schwarzschild camera can be clearly identified in the fluorescent image.

To improve the resolution of the fluorescent imaging system further, the diffraction limit ($0.1 \sim 0.2$ μm, the Rayleigh criterion) of the microscope used needs to be considered. According to the Rayleigh criterion, the smallest size which can be resolved by a lens is proportional to the ratio between the wavelength used and the numerical aperture. The latter is related to the

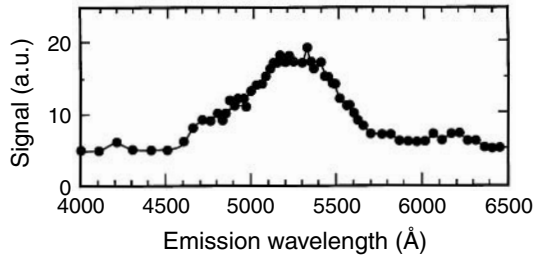


Figure 11. X-ray fluorescent spectrum of the STI-F10 G crystal (58) (courtesy of B. LaFontaine et al.; reprinted from *Appl. Phys. Lett.* **63**, B. La Fontaine, A. A. MacDowell, Z. Tan, D. L. Taylor, O. R. Wood II, J. E. Bjorkholm, D. M. Tennant, and S. L. Hulbert, Submicron, Soft X-ray Fluorescence Imaging, pp. 282–284, copyright 1995, American Institute of Physics with permission).

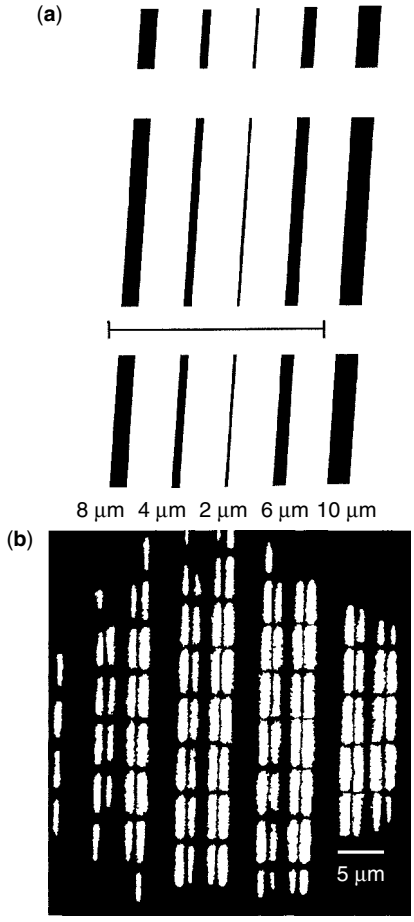


Figure 12. Soft X-ray fluorescent mask (**a**) and image (**b**) (58) (courtesy of B. LaFontaine et al.; reprinted from *Appl. Phys. Lett.* **63**, B. La Fontaine, A. A. MacDowell, Z. Tan, D. L. Taylor, O. R. Wood II, J. E. Bjorkholm, D. M. Tennant, and S. L. Hulbert, Submicron, Soft X-ray Fluorescence Imaging, pp. 282–284, copyright 1995, American Institute of Physics with permission).

refractive index of the lens. Therefore, using fluorescent crystals that have a high index of refraction and a shorter wavelength emitter may put the resolution below a tenth of a micron.

This high-resolution X-ray fluorescent imaging technique has potential in a variety of applications, especially in optimizing designs for deep ultraviolet and extreme ultraviolet lithographic exposure tools.

X-ray Fluorescent Holographic Images at Atomic Resolution

The relative positions of atoms in a crystal unit cell — the so-called single-crystal structure — are valuable structural information for a crystalline material. Diffraction methods are usually applied to determine crystal structure. However, due to the fact that a diffraction pattern provides intensity but not phase information for diffracted beams — the so-called phase problem (59) — unique determination of crystal structure cannot usually be achieved. Several methods have been developed to solve this phase problem: the Patterson method (60), molecular replacement (60) and isomorphous replacement (61), direct methods (62,63), multiwavelength anomalous dispersion (MAD) (64), and multiple diffraction (65,66). It is, most desirable however, to image the crystallographic structure and to visualize the atomic arrangement in a crystal directly. The holographic approach is certainly a promising candidate that is well suited to obtain phase information for a diffracted beam. The idea of using X-ray fluorescen holographic imaging has been proposed to attain atomic resolution and element-specific requirements (67–79). Very recently, effective methods for mapping the structure of single crystals and crystal surfaces at atomic resolution have been developed, and they are described here.

Normal (single-energy) X-ray Fluorescent Holographic Imaging (NXFH). The formation of a holograph involves an objective beam (wave) and a reference beam (wave) (80). The phase information (or contrast) can be retained (or revealed) by the interference between an objective (scattered) beam and a reference beam. For X-ray fluorescence, the internal X-ray source inside a crystal is the key to holographic formation. This internal X-ray source is the emitted fluorescence from atoms inside the crystal excited by incident X radiation. Figure 13a shows the schematic of single-energy X-ray fluorescent holography, where externally excited fluorescence from a target atom A is treated as a holographic reference beam. Fluorescence scattered from neighboring atoms serves as the objective beam. The interference between the two modulates the scattered fluorescent intensity, which is monitored by a detector located at a large distance from the crystal (73). By moving the detector around the crystal, an intensity pattern $I(\vec{k})$ is recorded, where $\vec{k}$ is the wave vector of the scattered beam. The Fourier transform of $I(\vec{k})$ yields an atomic resolution hologram.

The basic formula that describes the scattered intensity $I(\vec{k})$ is given by (72).

$$I(\vec{k}) = I_o/R^2[1 + |\sum a_j|^2 + 2\mathrm{Re}(\sum a_j)], \qquad (5)$$

where I_o is the intensity of the source atom A, a_j is the amplitude of the objective wave scattered by the jth atom inside the crystal, and R is the distance between the

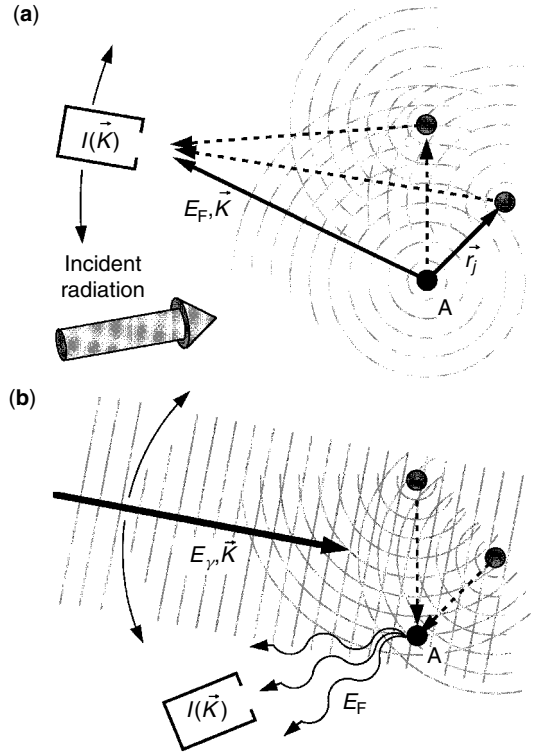


Figure 13. Schematic of (**a**) normal X-ray fluorescent holographic (NXFH) imaging and (**b**) inverse X-ray fluorescent holographic (IXFH) imaging (73) (courtesy of T. Gog et al.; reprinted from *Phys. Rev. Lett.* **76** T. Gog, P. M. Len, G. Materlik, D. Bahr, C. S. Fadley, and C. Sanchez-Hank, Multiple-Energy X-Ray Holography: Atomic Images of Hematite (Fe_2O_3), pp. 3,132–3,135, copyright 1996, American Physical Society with permission).

sample and the detector. Re means real part. The first term in the square bracket represents the intensity of the reference wave without interaction with neighboring atoms. The second term corresponds to the intensity of the scattered objective waves. The last term results from the interference between the reference and objective waves. Equation (5) also holds for multiple scattering. Because the scattering cross section of X rays is relatively small compared with that of electrons, the main contribution to interference is from the holographic process. Usually, for weak scattering of photons, the second term is of an order of 10^{-3} smaller than that of the interference term. For a crystal that has multiple atoms, the amplitudes of the objective and the interference waves are much more complicated to estimate, and separating the two is difficult. In addition, because many atoms radiate independently, many holograms thus formed complicate the image. The basic requirements for obtaining a single hologram are the following: (1) The crystal size needs to be small. (2) All of the irradiating atoms must have the same environment, and isotropy of the irradiation centers distributed in the sample is essential. For the former, the crystal size illuminated by the incident X rays has to be smaller than $R\lambda/r$, where λ is the wavelength of X-ray fluorescence and r is the atomic resolution expected. (3) Due to the translational symmetry imposed by the crystal, the value of the second term of Eq. (5) can be substantial. As pointed out in (68), the spatial frequency of this term is much larger than that of the holographic information provided by the nearest neighbor radiating atoms. Therefore, a low-pass filter (68) can be used to remove the contribution of the second term. (4) The effect of absorption by the crystal on the scattered intensity, depending on the crystal shape and the scattering geometry, needs to be corrected to see the expected 0.3% oscillatory signals in the hologram. (5) Correcting the dead time of the detector is necessary. (6) Reducing the noise by using a high-pass filter is required (78).

In addition to these requirements, a large scanning angular range must be covered to obtain a three-dimensional image that has isotropic atomic resolution. Some diffraction effects, such as the formation of Kossel lines (65,81), generated by the widely distributed fluorescence in the crystal, need to be suppressed.

A typical experimental setup for normal X-ray fluorescent holographic (NXFH) imaging is shown in Fig. 14 (72). The incident X-rays from a conventional sealed X-ray tube or synchrotron radiation facility are monochromatized by a graphite crystal and impinge at 45° on the crystal sample mounted on a goniometer, which provides θ-angle tilting and azimuthal ϕ rotation normal to the crystal surface. A Ge solid-state detector cooled by liquid nitrogen is used to monitor the scattered fluorescence. The two-dimensional hologram is measured by turning the sample around the ϕ axis and moving the detector by rotating the θ axis in the plane of incidence defined by the incident beam and the ϕ axis.

The upper panel of Fig. 15 shows the 2,402-pixel hologram of a $SrTiO_3$ single crystal after absorption correction and removing the incident-beam contribution. The crystal is plate-like, 30 mm in diameter, and 0.5 mm thick. It is crystallized in the perovskite-type $SrTiO_3$ structure whose lattice constant $a = 3.9$ Å. The Sr atoms form a simple cubic lattice. The large surface is parallel to the (110) plane. Mo Kα radiation ($E \sim 17.425$ keV) is used to excite Sr Kα, where the Sr K edge is 16.105 keV. The

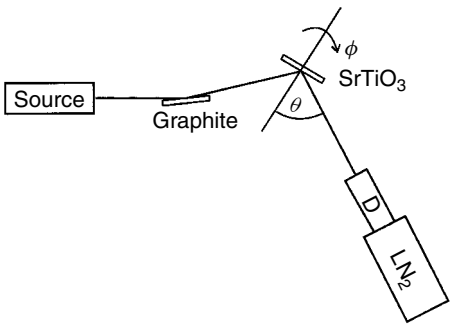


Figure 14. Experimental setup for NXFH imaging using a conventional X-ray source: the monochromator (graphite), the sample ($SrTiO_3$), and the detector (liquid nitrogen LN_2 cooled SSD) (72) (courtesy of M. Tegze et al.; reprinted with permission from *Nature* **380**, M. Tegze and G. Faigel X-Ray Holography with Atomic Resolution, pp. 49–51, copyright 1996 Macmillan Magazines Limited).

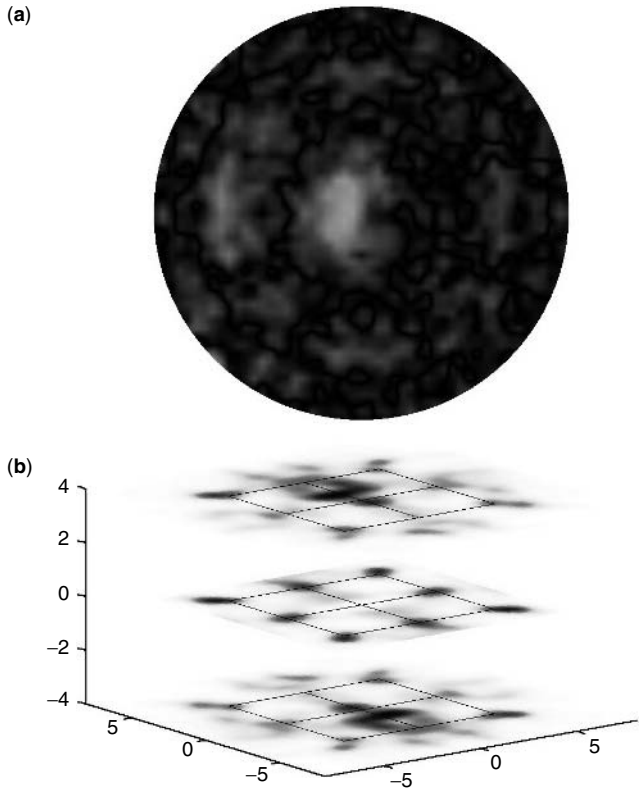


Figure 15. (**a**) The normal X-ray fluorescent hologram of $SrTiO_3$, and (**b**) three-dimensional reconstructed image of $SrTiO_3$ structure (only Sr atoms are revealed) (72) (courtesy of G. Faigel et al.; reprinted with permission from *Nature* **380**, M. Tegze and G. Faigel X-Ray Holography with Atomic Resolution, pp. 49–51, copyright 1996 Macmillan Magazines Limited).

photon-counting statistic for each pixel is about 0.05%, and the overall anisotropy, that is, the spatial distribution of interference signals, in the measured intensities is 0.3%, in agreement with the theoretical expectations described in Eq. (5). The atomic arrangement is reconstructed by using the Helmholtz–Kirchoff formula (78,82) and the proper absorption and dead-time corrections mentioned. Three atomic planes parallel to the crystal surface are depicted in the lower panel of Fig. 15. Only the Sr atom that has a large atomic scattering factor can be observed. Twin images are known to occur in traditional holography. However, due to the centrosymmetry of the crystal lattice, the twin and real images of different atoms occur at the same position. Because these two waves (one of the twin and the other of the real image) may have different phases, the interference between the two can cause an intensity modulation close to the atomic positions. This effect can lead to an appreciable shift in the positions of the atoms in question or to cancellation of a given atomic image if the two waves are out of phase. The different sizes of the atoms in Fig. 15 are due mainly to the flat plate geometry of the crystal because of the different resolutions in the in-plane (parallel to the crystal surface) and the out-of-plane (perpendicular direction).

Inverse (multiple-energy) X-ray Fluorescent Holographic Imaging (IXFH)

Multiple-energy holographic imaging (83–85) is well established for photoelectrons, Auger electrons, backscattered Kikuchi lines, and diffuse low-energy electrons and positrons. The idea of using multiple energy for X-ray fluorescence has been adopted in inverse X-ray fluorescent holographic imaging (IXFH). A schematic representation of IXFH is shown in Fig. 13b (73). The radiation source and detector are interchanged compared to their positions in normal X-ray fluorescent holography (NXFH) (Fig. 13a). The detector used in NXFH is now replaced by the incident monochromatic radiation of energy E_K, which produces an incident plane wave at the sample that propagates along the wave vector $\vec{k}$. As the plane wave moves toward atom A, a holographic reference wave is formed. The scattered radiation from the neighboring atoms serves as the holographic objective wave. The overlap and interaction of the reference and the objective waves at atom A excite it, which results in the emission of fluorescence of energy E_F. The intensity of the fluorescence is proportional to the strength of the interference. Thus, atom A, formerly the source of radiation in NXFH, now serves as a detector. By scanning the incident beam direction $\vec{k}$ and changing its energy E_K, the fluorescent intensity distribution $I(\vec{k})$ that is emitted at atom A as a function of $\vec{k}$ is collected. The intensity distribution $I(\vec{k})$ is then Fourier transformed to produce an atomic holographic image.

The two holographic schemes, NXFH and IXFH, are, in principle, equivalent according to the reciprocity theorem in optics (80). The main difference between these two schemes is that NXFH uses monochromatic fluorescence from internal excited atoms to produce a holographic scattering field, which is measured in the far field (8). The IXFH scheme, on the other hand, utilizes energy-tuned external radiation to generate a holographic scattering field, which is detected in the near field (8). Therefore, the fluorescence of the latter plays no role in establishing the scattering field for the holographic process.

Because an energy-tunable source is required for IXFH, the use of synchrotron radiation is essential. Figure 16 shows the experimental configuration for IXFH measurements using synchrotron radiation. Monochromatic radiation, after the double-crystal monochromator, is focused onto the sample, which is mounted on a six-circle diffractometer. The sample can be tilted (θ-angle) in the plane of incidence and rotated via the azimuthal angle Φ around an axis normal to the crystal surface. Fluorescence emitted from the atoms in the sample excited by the incident radiation is collected by a cylindrical graphite analyzer and detected by a proportional counter.

The inset of Fig. 16 shows a typical scan at $E = 9.65$ keV and $\theta = 55^\circ$ for a (001) cut hematite (Fe_2O_3) natural crystal slab (73). The effect of the detector dead time has been corrected. The mosaic spread of the crystal is 0.01°. The angular acceptance of the curved analyzer is approximately 14° in the direction of curvature and 0.5° along the straight width. As can be seen, the signal modulation is about 0.5% of the averaged background. The local structural images, such as Kossel lines, are

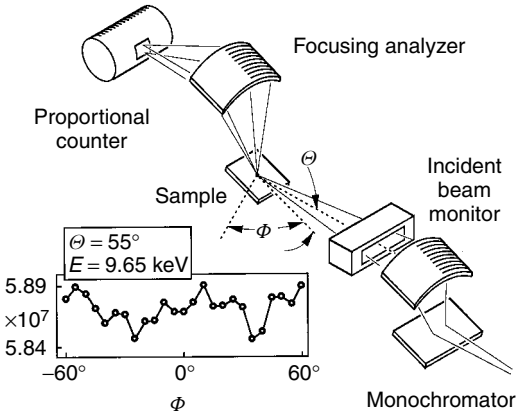


Figure 16. Experimental setup for multiple-energy IXFH imaging. The inset shows the intensity of fluorescence, corrected for detector dead time, versus the azimuthal angle Φ (73) (courtesy of T. Gog et al.; reprinted from *Phys. Rev. Lett.* **76**, T. Gog, P. M. Len, G. Materlik, D. Bahr, C. S. Fadley, and C. Sanchez-Hank, Multiple-Energy X-Ray Holography: Atomic Images of Hematite (Fe_2O_3), pp. 3,132–3,135, copyright 1996, American Physical Society with permission).

suppressed by the large angular acceptance of the curved analyzer.

Hematite forms a hexagonal crystal that has lattice constants $a = 5.038$Å and $c = 13.772$Å. Figure 17a is the projection of Fe atoms onto the basal plane (001) perpendicular to the c axis, where the Fe positions in two different kinds of stacking order are shown. For the first kind, the black circles represent iron atoms in an upper plane, and the gray circles are atoms in the plane 0.6Å below the black circles. For the second kind, white circles denote upper plane atoms, and black circles are also 0.6Å under the white ones. Figure 17b shows the calculated fluorescent image of the Fe sites. Using the experimental arrangement mentioned, the two different Fe atom stackings are not distinguishable in the IXFH images. A superposition of Fe atom images in a holographic reconstruction of these layers occurs, as described in the previous section. The reconstructed image (Fig. 17c) of the (001) Fe layer from the experimental data indeed shows the effect of superposition (see Fig. 17a for comparison). As can be seen, six nearest neighbored Fe atoms, three for each kind of stacking order, are located 2.9Å from the center atom. The diagonal distance of this figure corresponds to 8.7Å.

The experimental data consist of the fluorescent intensities for three incident energies, $E = 9.00$, 9.65, and 10.30 keV, measured in 150 hours within the ranges of $45° \leq \theta \leq 85°$ and $-60° \leq \Phi \leq 60°$ by the scan window $\Delta\theta = 5°$ and $\Delta\Phi = 5°$. The measured intensity $I(\vec{k})$ is mapped onto the entire 2π hemisphere above the sample by considering the threefold symmetry of the c axis. The background $I_0(k)$ derived from a Gaussian low-pass convolution (86) is subtracted from the measured intensity. The theoretically reconstructed image (Fig. 17b) calculated for the clusters of Fe atoms shows good agreement with the experiment.

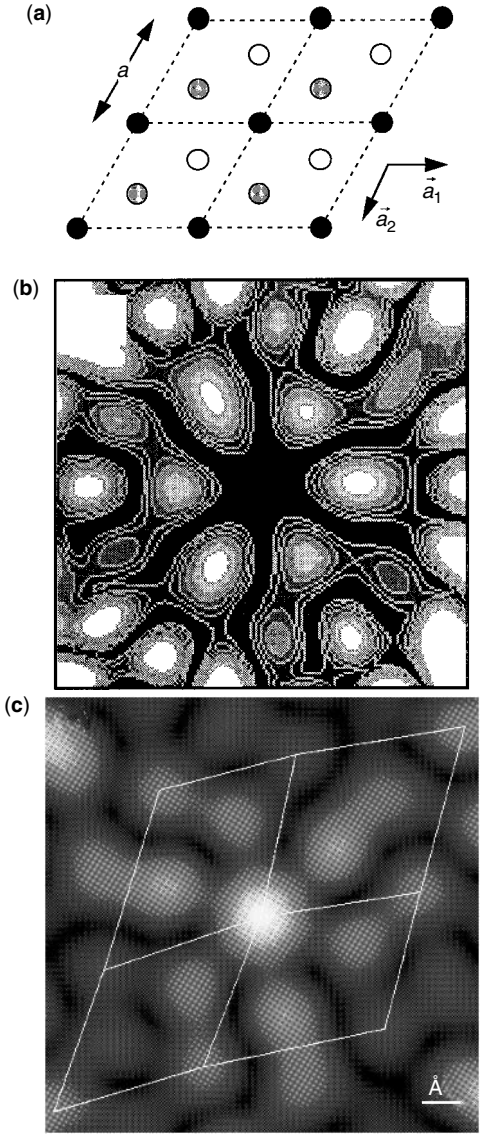


Figure 17. (**a**) The projection of Fe atom positions on the (001) plane of hematite. (**b**) Image calculated for the Fe sites. (**c**) Holographic reconstruction of the (001) Fe layer (73) (courtesy of T. Gog et al.; reprinted from *Phys. Rev. Lett.* **76**, T. Gog, P. M. Len, G. Materlik, D. Bahr, C. S. Fadley, and C. Sanchez-Hank, Multiple-Energy X-Ray Holography: Atomic Images of Hematite (Fe_2O_3), pp. 3,132–3,135, copyright 1996, American Physical Society with permission).

Isotropic resolution in every direction for three-dimensional reconstruction can be improved by combining the NXFH and IXFH techniques (79). Figure 18 is the experimental setup suitable for this combined NXFH and IXFH experiment. The arrangement is similar to that shown in Figs. 14 and 16, except that the three-circle diffractometer comprises two coaxial vertical rotations θ and θ' goniometers. The upper stacked one (θ') is on top of the other θ circle. A sample can be mounted on the horizontal ϕ axis situated on the lower θ circle. The sample can be rotated as usual by varying the incident angle θ (tilting angle) and the azimuthal ϕ angle. A detector, for example, a high purity germanium solid-state detector

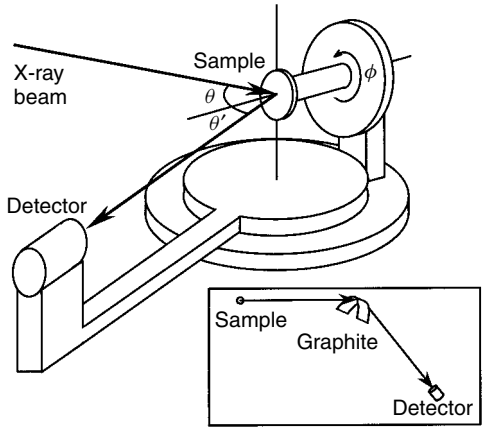


Figure 18. Experimental setup for combined NXFH and IXFH imaging. The inset shows a curved graphite analyzer used in the synchrotron experiments in place of the detector in the main picture (79) (courtesy of G. Faigel et al.; reprinted from *Phys. Rev. Lett.* **82**, M. Tegze, G. Faigel, S. Marchesini, M. Belakhovsky, and A. I. Chumakov, Three Dimensional Imaging of Atoms with Isotropic 0.5Å Resolution, pp. 4,847–4,850, copyright 1999, American Physical Society with permission).

(SSD), is mounted on the θ' circle. Again a doubly focused curved graphite monochromator is used for synchrotron sources (see the inset in Fig. 18), and an avalanche photodiode (APD) is employed to handle the high counting rate (87). The θ and θ' circles facilitate both the NXFH and IXFH experiments without changing the diffractometer. Special care must be taken with respect to the detector and sample motions to avoid losing any information between pixels and to maintain uniform resolution throughout the holographic image.

Figures 19, 20, and 21 represent the holograms and reconstructed atomic images of a circular platelet of CoO crystal (79). The large face of the crystal is parallel to the (111) plane, and the mosaic spread of the crystal is 0.3°. CoO forms a face-centered cubic (fcc) crystal whose lattice constant $a = 4.26$Å. The sample was subjected to both NXFH and IXFH imaging experiments at several photon energies. The time for data collection ranged from three hours for synchrotron sources to some 20 days for conventional sources. The measured Co Kα fluorescent intensity distribution, normalized to the incident intensity shown in Fig. 19a, is the projection of the hologram mapped onto the surface of a sphere defined by the coordinates θ and ϕ, where $\theta \leq 70°$. The picture is obtained by a synchrotron measurement in the IXFH mode at the photon energy $E = 13.861$ keV sufficient to excite Co Kα. The dominant feature in this picture is due to the strong θ-dependent absorption of the emitted fluorescence by the sample, according to its shape (78). This absorption effect can be easily corrected by a theoretical calculation (78). Figure 19b is the hologram after this correction. The wide radial stripes and the narrow conic lines (the Kossel lines) are the two main features. The former originates from the change of the crystal orientation relative to the detector during the crystal scan, so as to keep the incident radiation source on the hemisphere above the crystal. This relative change in crystal orientation

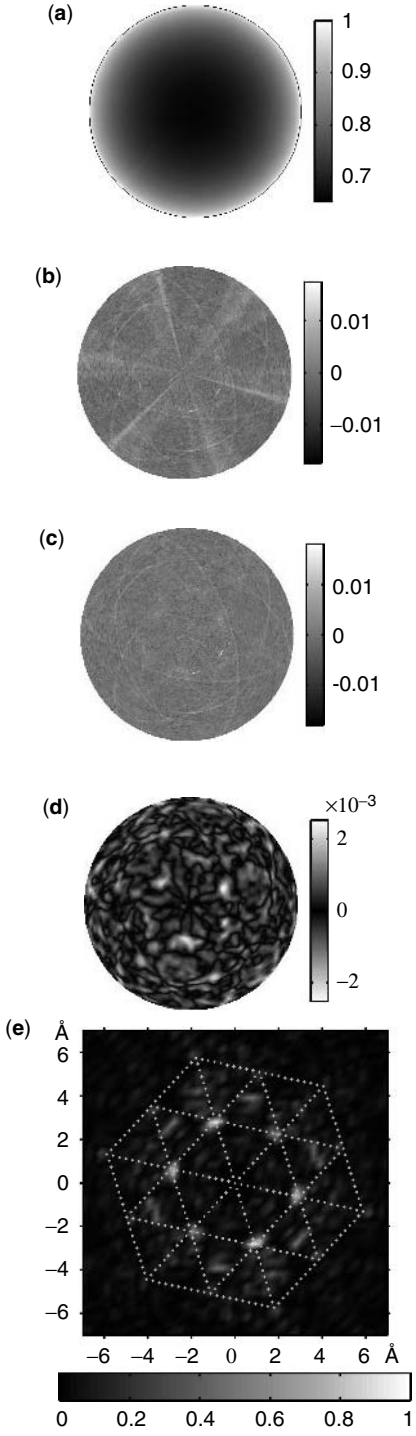


Figure 19. Holograms obtained at various stages: (**a**) the normalized intensity distribution of Co Kα fluorescence measured; (**b**) intensity distribution after correction for sample absorption; (**c**) image after the correction for the detector position and crystal orientation; (**d**) image after removal of Kossel lines, using filter techniques; (**e**) reconstructed atomic images in the plane containing the source atom and parallel to the crystal surface. The dotted lines indicate the crystal lattice (79) (courtesy of G. Faigel et al.; reprinted from *Phys. Rev. Lett.* **82**, M. Tegze, G. Faigel, S. Marchesini, M. Belakhovsky, and A. I. Chumakov, Three Dimensional Imaging of Atoms with Isotropic 0.5Å Resolution, pp. 4,847–4,850, copyright 1999, American Physical Society with permission).

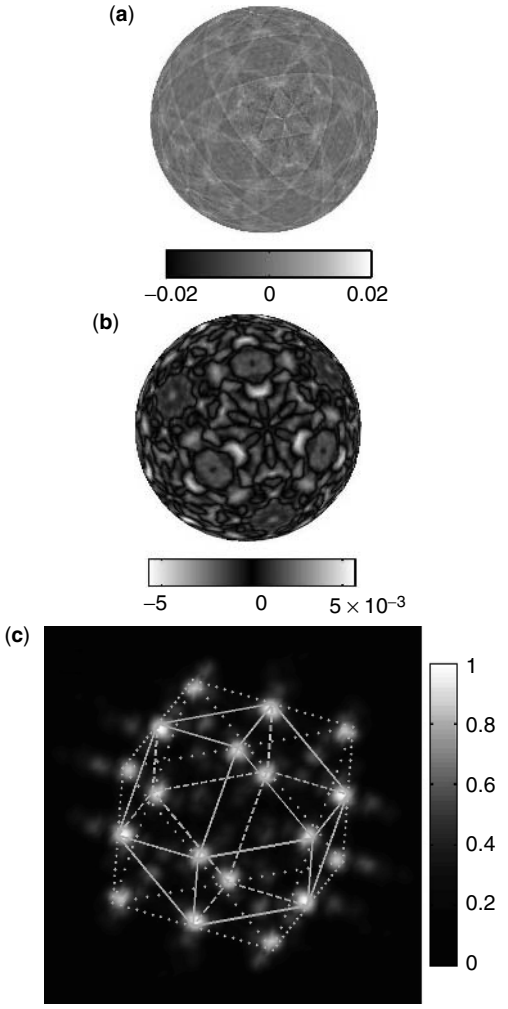


Figure 20. (**a**) Extended hologram before and (**b**) after employing a low-pass filter; (**c**) reconstructed atomic image of Co atoms. The first nearest neighbors are shown in dashed lines, and the next nearest neighbors in dotted lines (79) (courtesy of G. Faigel et al.; reprinted from *Phys. Rev. Lett.* **82**, M. Tegze, G. Faigel, S. Marchesini, M. Belakhovsky, and A. I. Chumakov, Three Dimensional Imaging of Atoms with Isotropic 0.5Å Resolution, pp. 4,847–4,850, copyright 1996, American Physical Society with permission).

introduces modulation of the fluorescent intensity, which can be precisely measured and subtracted. Figure 19c is the corrected hologram. The Kossel, or, sometimes, called XSW (X-ray standing wave) lines, can be filtered out by a low-pass spatial filter, as already mentioned. This filtering is carried out by calculating the convolution of the hologram that has a Gaussian on the surface of the sphere. The width of the Gaussian is $\sigma \approx \lambda/2\pi r_{\max}$, where the maximum radius $r_{\max} \approx 5$Å of the region for imaging and λ is the X-ray wavelength. The hologram after filtering is shown in Fig. 19d. Threefold symmetry due to the (111) surface is clearly seen in the figure. The atomic image in the plane that contains the source atom, parallel to the crystal surface, was reconstructed by using the Helmholtz–Kirchhoff formula (78,82) and is shown in Fig. 19e. The six nearest neighbored Co atoms appear at approximately the proper positions relative to the central fluorescent atom (not shown). Clearly, Fig. 19e is merely a two-dimensional reconstructed atomic image. For three-dimensional imaging, the sampling angular range should be increased up to 4π in solid angle, and the structural information provided by the Kossel lines (or XSW lines) needs to be included. Figure 20a shows the extended hologram after accounting for these two factors. The images of conic Kossel lines become much clearer. The low-pass filtered image is shown in Fig. 20b for comparison. The corresponding three-dimensional image reconstructed from Fig. 20b and the outline of the crystal lattice are given in Fig. 20c. Both the nearest and next nearest neighbor Co atoms are clearly seen. However, the nearest neighbor sites are slightly shifted toward the central atom due to the twin-image interference and the angular dependence of the atomic scattering factor (78). This distortion can be corrected by summing holograms taken at different energies (82), that is, the properly phased summing of the reconstructed wave amplitudes suppresses the twin images and reduces the unwanted intensity oscillation. Therefore, the combination of NXFH and IXFH (multiple-energy) could provide better image quality. Figures 21a–d show the holograms of CoO measured at $E = 6.925$, 13.861, 17.444, and 18.915 keV, respectively. The first and the third pictures were taken in the normal mode (NXFH) using a conventional X-ray source, and the second and the fourth were taken in the inverse mode (IXFH) using a synchrotron source. The combination of these four measurements leads to the three-dimensional atomic structure of the Co atoms shown in Fig. 21e. Slight improvement of the image quality in Fig. 21e, compared with the single-energy image in Fig. 20c can be detected. The atomic positions of Co atoms are closer to their real positions in the known CoO crystal, and the background noise is substantially reduced. The spatial resolution of the atomic positions is estimated from the full width at half-maxima (FWHM) of the fluorescent intensity at approximately 0.5Å. The deviation of the intensity maxima from the expected real positions of the nearest Co neighbors is less than 0.1Å.

In summary, X-ray fluorescent holographic imaging can be used to map three-dimensional local atomic structures at isotropic resolution and without any prior knowledge of the structures. The combination of the experimental techniques of NXFH and IXFH, single-energy and multiple-energy, together with mathematical evaluation methods, reinforces the three-dimensional capabilities of X-ray fluorescent holographic imaging. Very recently, improved experimental conditions and a data-handling scheme have made possible the imaging of light atoms (88), as well as atoms in quasi-crystals (89). It is hoped that this X-ray fluorescent imaging method can be applied in the future to more complex crystal systems, such as macromolecules.

X-ray Fluorescence Images from Solids in Electric Fields

X-ray fluorescence emitted from atoms usually exhibits a homogeneous spatial intensity distribution, except for the direction along a polarization axis. The familiar doughnut-shaped distribution that has zero intensity along the

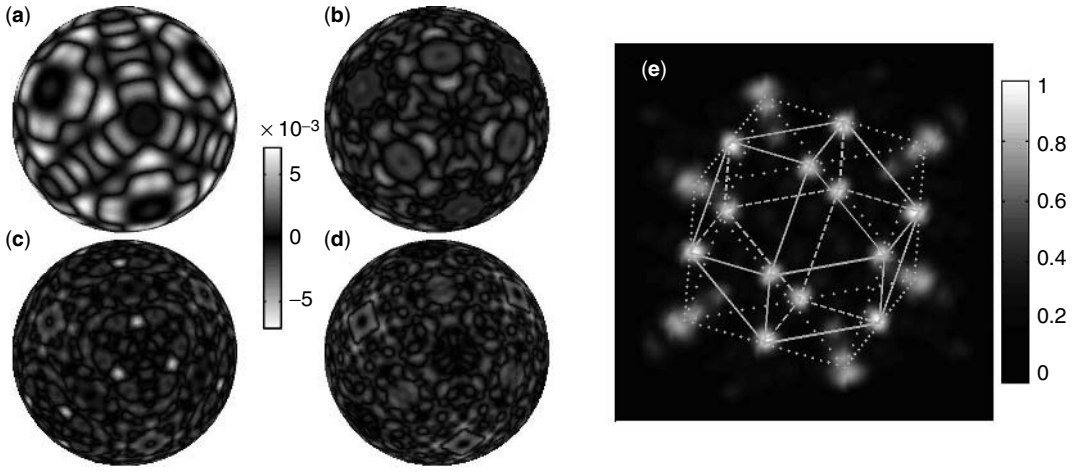


Figure 21. Holograms of CoO measured at $E =$ (**a**) 6.925, (**b**) 13.861, (**c**) 17.444, and (**d**) 18.915 keV. (**e**) Three-dimensional images obtained from the combination of the four measurements (79) (courtesy of G. Faigel et al.; reprinted from *Phys. Rev. Lett.* **82**, M. Tegze, G. Faigel, S. Marchesini, M. Belakhovsky, and A. I. Chumakov, Three Dimensional Imaging of Atoms with Isotropic 0.5Å Resolution, pp. 4,847–4,850, copyright 1996, American Physical Society with permission).

polarization axis, as expected from Eq. (4), is commonly observed in many experiments (8, 90–96). In contrast to this common feature, different X-ray fluorescent images could be formed if a solid sample is placed in an external electric field during photon excitation. In recent experiments, discrete ring-like fluorescent images have been observed from solids excited by synchrotron radiation (97). The semi angles of the rings found are less than 20° and are closely related to the atomic number Z and to the principal and the orbital angular momentum quantum numbers n and ℓ. The details of this unusual observation are described here.

Synchrotron radiation of energies ranging from 15 eV to 9 keV is used as the X-ray source to cover the excitation energies of the elements of $4 < Z < 27$. The formation of X-ray fluorescent images in amorphous, single- and polycrystalline materials that contain these elements was investigated. The samples were placed on a holder at the center of a UHV chamber (10^{-9} Torr) (see Fig. 22). Incident radiation whose beam size was 1.5×1.5 mm and photon energy E_0 hit the sample and fluorescent radiation generated was recorded on a five-plate MCP (microchannel plate) detector placed 6 to 10 cm from the sample (Fig. 22). The MCP detector is operated at −4,050 volts during fluorescent detection. The diameter of the MCP was 30 mm, and the detecting area contained $1{,}022 \times 1{,}022$ pixels. The spatial resolution was 25 μm, the width of a pixel. A copper grid with a +150 volts bias was mounted on the front face of the detector to stop the positive ions. The −4,050 volt bias of the MCP is sufficient to prevent negatively charged particles from entering the detector. The fluorescent signals amplified by the electron gain of 5×10^7 were digitized and displayed on a color monitor. The sample current I_s can be measured and a variable bias V_s can be applied to the sample. φ and θ are the angles of the sample surface and the detection direction with respect to the incident beam, respectively. The detector is placed at $\theta = 45°$ or $\theta = 90°$. Total fluorescence yield (TFY) and total

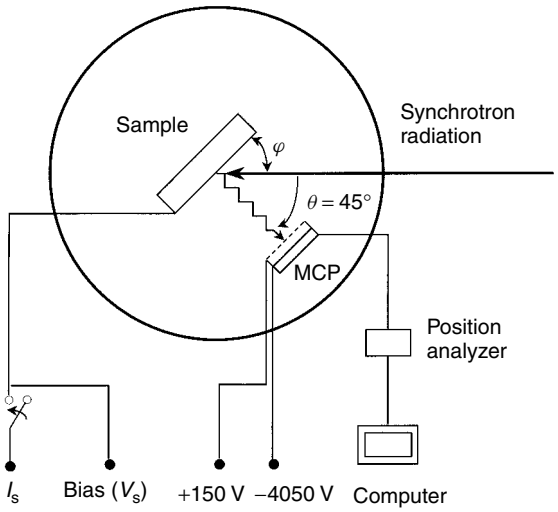


Figure 22. Schematic of the experimental setup for X-ray fluorescence in an electric field. The monochromatic SR excites the sample and generates fluorescence measured by the MCP area detector. The electric field results from the working bias of the MCP (97).

electron yield (TEY) versus E_0, measured from the MCP and the sample current I_s, respectively, gave absorption spectra that indicate the energy positions of absorption edges. Figures 23a and b show the fluorescent images obtained at $\theta = 45°$ for $\varphi = 45°$ from a sapphire sample whose photon energy was slightly above ($E_0 = 545$ eV) and below ($E_0 = 500$ eV) the oxygen K edge ($E_K = 543$ eV). A bright ring and a dark hole are observed for $E_0 > E_K$ (Fig. 23a) and $E_0 < E_K$ (Fig. 23b), respectively. The ring images remain the same for various φ angles, as long as $E_0 \geq E_K$ (the energies of absorption edges), irrespective of the crystalline form of the samples. For example, the fluorescent rings of oxygen 1s from a single-crystal Al_2O_3,

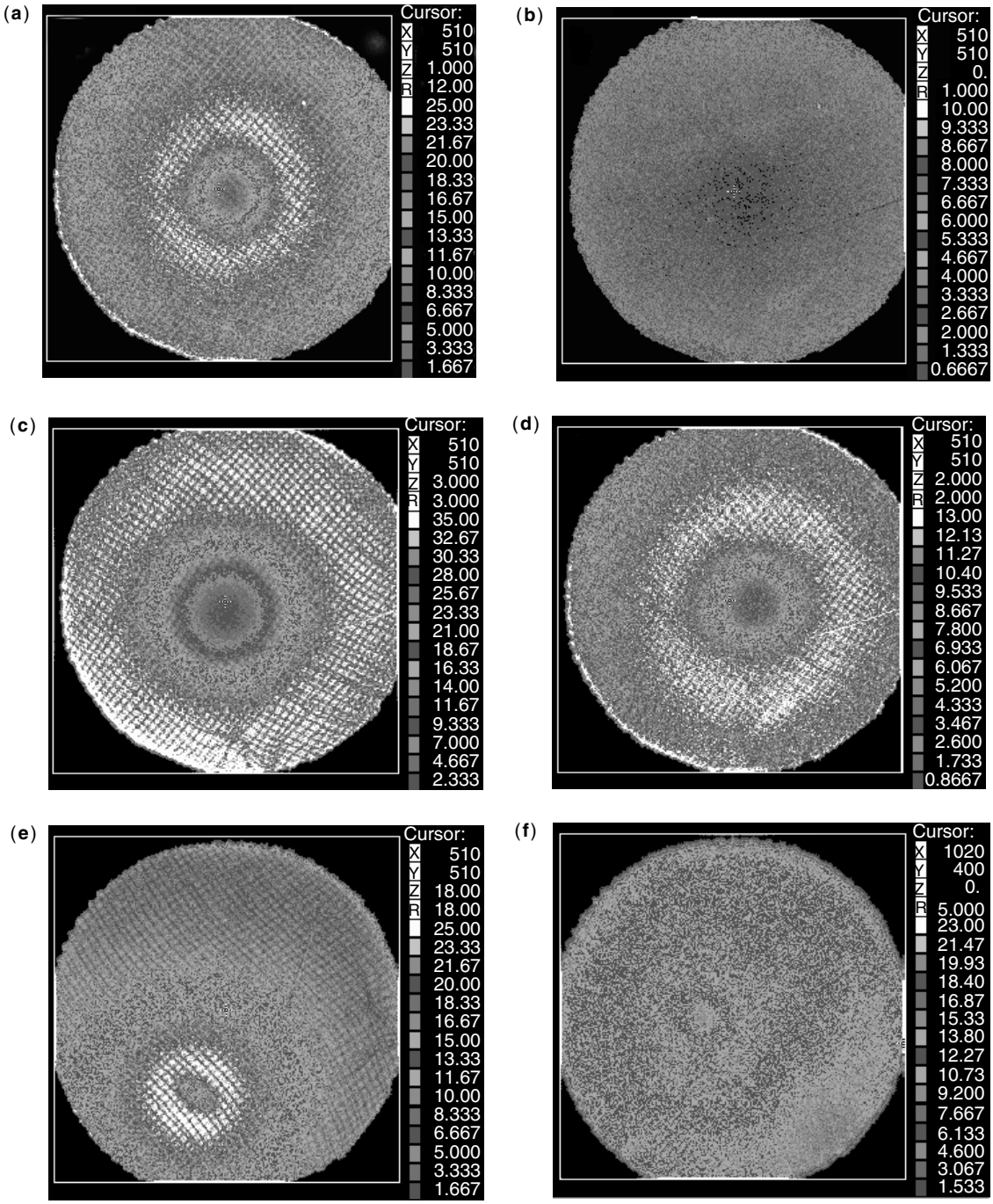


Figure 23. X-ray fluorescent ring images at $\theta = 45^\circ$ and $\varphi = 45^\circ$ of (**a**) O_{1s} of sapphire excited at $E_0 = 545$ eV, (**b**) sapphire at $E_0 = 500$ eV, (**c**) B_{1s} of boron nitride (BN) powder excited at $E_0 = 194$ eV $> E_K$ (B_{1s}), (**d**) C_{1s} of graphite at $E_0 = 285$ eV $> E_K$ (C_{1s}), (**e**) Ti_{1s} excited at $E_0 = 5{,}000$ eV $> E_K$ (Ti_{1s}), and (**f**) the photoelectron image of Mn_{1s} excited at $E_0 = 6{,}565$ eV $> E_K$ (Mn_{1s}), where $E_K(O_{1s}) = 543$ eV, $E_K(B_{1s}) = 192$ eV, $E_K(C_{1s}) = 284$ eV, $E_K(Ti_{1s}) = 4{,}966$ eV, and $E_K(Mn_{1s}) = 6{,}539$ eV. (A scratch is always observed on the MCP screen (97).

a piece of glass, and amorphous $LaNiO_3$ and SiO_2 samples are nearly the same.

Fluorescent rings are repeatedly observed for various elements of atomic numbers $4 < Z < 27$ at $\theta = 45^\circ$ and 90°. For illustration, Figs. 23c–e display the images of the excitations of B_{1s}, C_{1s}, and Ti_{1s} in color. Compared with fluorescence, photoelectron ring images (98) can also be clearly observed by applying a $-2{,}500$ volt sample bias. Figure 23f shows such a ring image of Mn_{1s} photoelectrons [$E_K(Mn) = 6{,}539$ eV]. The experimental conditions for recording Fig. 23 are the following: The energy resolutions are $\Delta E = 100$ meV for Figs. 23a and b, 50 meV for Fig. 23c, and 800 meV for Fig. 23e and f. The sample-detector distances are 9.4 cm for Figs. 23a–d and 7.9 cm for

Figs. 23e and f. The angle φ is set at 45° and the counting time is 300 s for each image. Fluorescence involving 2p and 3d electrons that show ring images is also observed (not shown here). In addition, the excitation of the samples by photons of higher harmonics can also give ring images (not shown). The emitted photon spectra, measured by using a seven-element solid-state detector ($\Delta E = 100$ eV), proved to be the fluorescent radiation of energies $E < E_K$.

The semiangles δ corresponding to the ring radii follow the $E^{-1/2}$ relationship (Fig. 24a), where E is the emission energy between the energy levels of atoms. The δ value remains the same for the same element involved, irrespective of its crystalline form.

The origin of the formation of fluorescent rings is the following: The linearly polarized monochromatic radiation polarizes the atoms of the samples by the excitation of a given state (n, ℓ). The excited atoms then emit fluorescence in all directions when returning to their stable states. Because fluorescent rings are observed at $\theta = 45°$ and 90°, irrespective of the rotation of the sample, this implies that wherever the detector is placed, the fluorescent ring is observed along that direction. The high working bias of the MCP detector may also affect the polarized atoms of the sample during photon excitation. The influence of the electric field between the detector and the sample on the fluorescent image from changing the sample bias V_s is shown in Figs. 25a, b, and c for Cr_{1s} at $V_s = 0$, 2,000 and −2,500 volts, respectively. As can be seen, the intensity of the fluorescence (Fig. 25b) increases when the electric potential difference between the detector and the sample increases. The image becomes blurred when the electric potential difference decreases (Fig. 25c). The latter is also due partly to the presence of photoelectrons. This result indicates that the presence of the electric field $\vec{E}_R$ between the detector and the sample clearly causes the polarized atoms, namely, the emitting dipoles $\vec{p}$, to align to some extent with the $\vec{E}_R$. Thus, the spherically distributed dipoles originally in random fashion are now redistributed as shown in the inset of Fig. 24a. The preferred orientation of the dipoles is along $\vec{E}_R$, and there are no dipoles in the area defined by the opening angle 2β. Because the $\vec{E}_R$ field is very weak, angle β is very small.

The emitted fluorescent intensity distribution $I_F(\theta)$ versus the angle θ with respect to $\vec{E}_R$ for the dipole

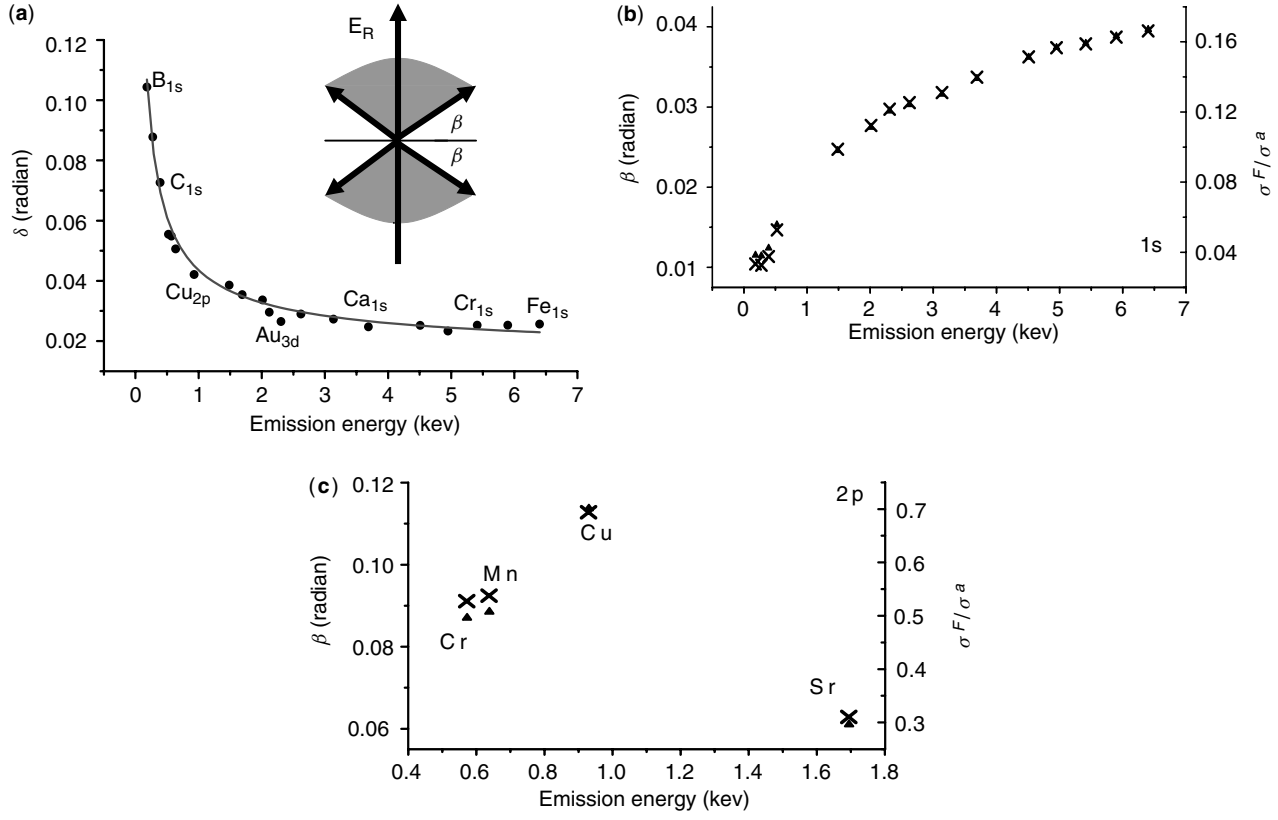


Figure 24. (**a**) The semiangles δ of the fluorescent cones versus the emission energy E of the various elements investigated; B_{1s}, C_{1s}, N_{1s}, O_{1s}, Cr_{2p}, Mn_{2p}, Cu_{2p}, Al_{1s}, Sr_{2p}, P_{1s}, Au_{3d}, S_{1s}, Cl_{1s}, K_{1s}, Ca_{1s}, Ti_{1s}, V_{1s}, Cr_{1s}, Mn_{1s}, Fe_{1s} in the order of emission energy. The fitting curve is the function $0.058(\pm 0.0016) \times \sqrt{1/E} + 0.012(\pm 0.0017)$. The standard deviation in δ is about 5%. (**b**) and (**c**) The corresponding opening angles β (the cross) and σ^F/σ^a (the triangle) for the elements investigated involving 1s and 2p fluorescence, respectively (see the text). The inset: Schematic of the redistributed dipoles (97).

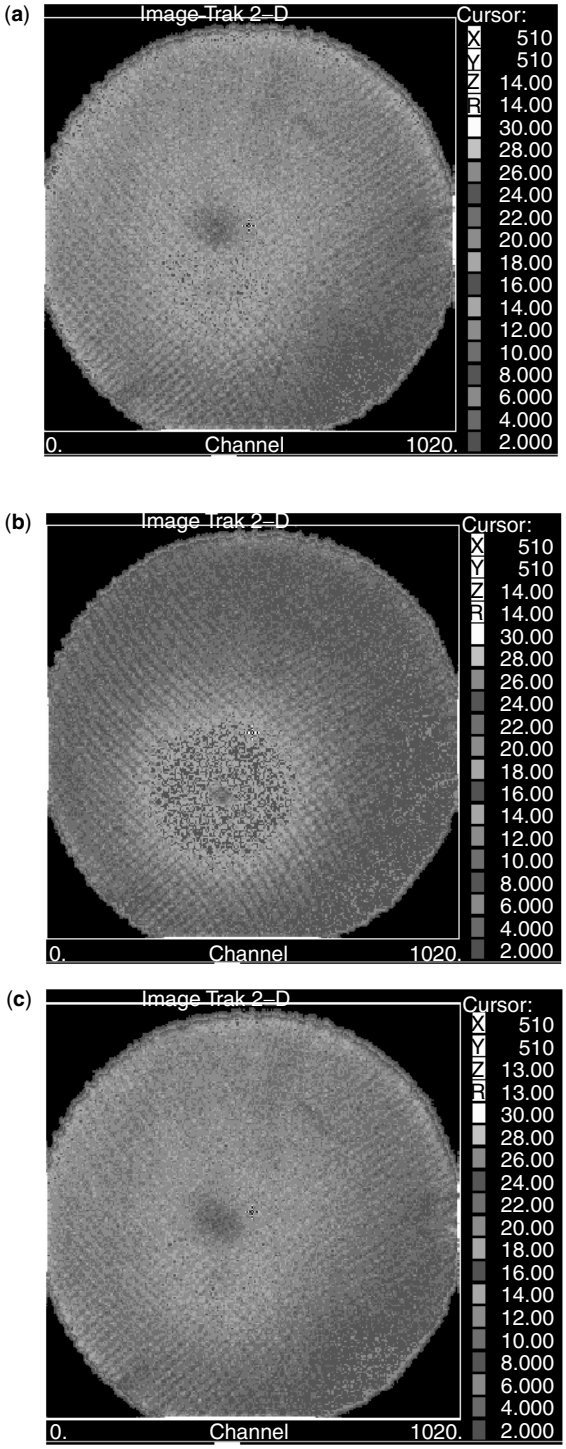


Figure 25. The fluorescent images of Cr_{1s} where the sample bias V_s is equal to (**a**) 0, (**b**) 2,000, and (**c**) −2,500 volts in a 200 s-exposure. The photoelectron ring appears near the center of the screen in (**a**) and (**c**), and the fluorescent ring is slightly below the center (97).

distribution shown in the inset of Fig. 24a can be calculated by considering the absorption of the incident and the fluorescent beams by the sample and the usual doughnut-shaped distribution $\sin^2(\theta)$ of the dipoles (see, also Eq. 4):

$$I_F(\theta) = \int_0^\infty dz \int_{\gamma=-a}^{a} d\gamma \frac{Bn}{\cos\varphi} \sin^2(\theta+\gamma) e^{-\frac{\mu_a z}{\cos\varphi}} e^{-\frac{\mu_F z}{\cos\theta}}, \quad (6)$$

where $B = I_0 A P_{AB} \frac{\alpha\omega^3}{2\pi c^3} |\vec{X}_{AB}|^2 V$ and $a = \frac{\pi}{2} - \beta$. μ_a and μ_F are the linear absorption coefficients of the incident and the fluoresced beams, respectively, where $\mu_a(\omega) = n_0\sigma^a$ and $\mu_F(\omega) = n_0\sigma^F$; σ's are the total absorption/fluorescence cross sections, and n_0 is the total number of the dipoles per unit volume. γ covers all of the angular range, except for the β range. z is along the inward surface normal. I_0, A, V, and P_{AB} are the intensity and the cross section of the incident beam, the irradiated volume, and the probability of absorption of the incident photons by the atoms in the transition from state A to state B, respectively. X_{AB} is the corresponding transition matrix element (see Eq. 4). Integration of Eq. (6) leads to

$$I_F(\theta) = B \times \frac{\sec\varphi(\pi/2 - \beta - 1/2 \sin 2\beta \cos 2\theta)}{(\sigma^F \sec\theta - \sigma^a \sec\varphi)}, \quad (7)$$

which describes a ring-like distribution. From the values of σ's given in Refs. 99 and 100, the β angles for various emission energies are determined by fitting the measured fluorescent profile $I_F(\theta)$ where β is the only adjustable variable. Figures 24b and c are the β angles determined and the ratios σ^F/σ^a versus emission energy for the 1s and 2p fluorescence. Clearly, both the β and σ^F/σ^a behave very similarly. Because the mean free path of X rays in the sample equals $1/\mu$, therefore, the β angle is closely related to the mean-free-path ratio ℓ_a/ℓ_F.

The ring centers of Figs. 23a, c–e and the dark hole in Fig. 23b indicate the direction of the $\vec{E}_R$ field. The strength of the $\vec{E}_R$ affects only the fluorescence intensity not the ring size. The formation of a bright ring or a dark hole on the detector depends only on whether the incident photon energy is greater or smaller than the absorption edges. This is analogous to the photoelectric effect.

Based on the experimental data, the semiangle of the fluorescent ring for a given Z, n, and ℓ can be predicted by using the curve-fitting relationship: $\delta = 0.058(\pm 0.0016) \times \sqrt{1/E} + 0.012(\pm 0.0017)$, where E is the emission energy between the energy levels of atoms.

In summary, the observed ring-like discrete spatial distributions of X-ray fluorescence result from the collective alignment effect of the dipoles in solids exposed to an external electric field. The induced small opening of the dipole distribution due to this alignment effect and the self-absorption by the samples related to X-ray mean free paths are responsible for the formation of a fluorescent ring, which differs drastically from our common knowledge about emitted radiation. An empirical formula predicts the dimension of the fluorescent rings for given Z, n, and ℓ. This interesting feature, in turn, provides an alternative for characterizing materials according to their fluorescent images.

X-ray fluorescence is a useful imaging technique for trace element analysis in biomedical and environmental applications and material characterization. It has the

potential for crystallographic structural studies such as crystal-structure determination of surfaces, interfaces, and bulk materials, using X-ray fluorescent holographic imaging. The availability of synchrotron radiation, the excellent soft and hard X-rays obtained from this source, and the increasing number of synchrotron facilities in the world will certainly enlarge the applicability of this fluorescent imaging technique. It is anticipated that in the near future this technique will develop into a major imaging tool for investigating various kinds of materials at submicron or even smaller scales.

Acknowledgments

The author is indebted to G. Faigel, T. Gog, N. Gurker, S. Homma-Takeda, A. Iida, Y. Kumagai, N. Shimojo, the American Institute of Physics, the American Physical Society, and the publisher of *Nature*, Elsevier Science B. V., and John Wiley and Sons for permission to reproduce figures and photographs from their published materials.

ABBREVIATIONS AND ACRONYMS

APD	avalanche photodiode
CCD	charge-coupled device
DCM	double-crystal monochromator
EM wave	electromagnetic wave
FAAS	flameless atomic absorption spectrometry
FWHM	full width at half-maxima
HV	high vacuum
IC	ionization chamber
ICCD	intensified charge-coupled device
ICP-AES	inductively coupled plasma atomic emission spectrometry
IEF-AGE	isoelectric focusing agarose gel electrophoresis
IP	imaging plate
IXFH	inverse (multiple-energy) X-ray fluorescent holography
K-B	Kirkpatrick-Baez
LN_2	liquid nitrogen
LPE	liquid-phase epitaxy
MAD	multiwavelength anomalous dispersion
MCP	microchannel plate
MeHg	methylmercury
MMC	methylmercury chloride
NXFH	normal (single-energy) X-ray fluorescent holography
SMM	synthetic multilayer monochromator
SOD	superoxide dismutase
SR	synchrotron radiation
SSD	solid-state detector
TEY	total electron yield
TFY	total fluorescence yield
UHV	ultrahigh vacuum
XSW	X-ray standing wave

BIBLIOGRAPHY

1. P. P. Ewald, *FiftyYears of X-Ray Diffraction*, N.V.A. Oosthoek's Uitgeversmaatschappij, Utrecht, The Netherlands, 1962.
2. A. H. Compton and S. K. Allison, *X-Rays in Theory and Experiment*, 2nd ed., Van Nostrand, Princeton, 1935.
3. L. V. Azaroff, *Elements of X-Ray Crystallography*, McGraw-Hill, NY, 1968.
4. T. Åberg and J. Tulkki, in B. Crasemann, ed., *Atomic Inner-Shell Physics*, Plenum press, New York, 1985, pp. 419–463 and the references therein.
5. E. Merzbacher, *Quantum Mechanics*, Wiley, NY, 1961.
6. A. Messiah, *Quantum Mechanics*, North-Holland, Amsterdam, 1962.
7. J. J. Sakurai, *Advanced Quantum Mechanics*, Addison-Wesley, NY, 1967.
8. J. D. Jackson, *Classical Electrodynamics*, Wiley, NY, 1967.
9. E. -E. Koch, D. E. Eastman, and Y. Farge, in E. -E. Koch, ed., *Handbook on Synchrotron Radiation*, 1a ed., North-Holland, Amsterdam, 1983, pp. 1–63.
10. T. Matsushita and H. Hashizume, in E. -E. Koch, ed., *Handbook on Synchrotron Radiation*, 1a ed., North-Holland, Amsterdam, 1983, pp. 261–314.
11. R. L. Mössbauer, *Z. Phys.* **151**, 124–143 (1958).
12. D. P. Siddons et al., *Rev. Sci. Instrum.* **60**, 1,649–1,654 (1989).
13. S. Hayakawa et al., *Nucl. Instrum. Methods.* **B49**, 555–560 (1990).
14. Y. Suzuki and F. Uchida, *Rev. Sci. Instrum.* **63**, 578–581 (1992).
15. P. Kirkpatrick and A. V. Baez, *J. Opt. Soc. Am.* **38**, 766–774 (1948).
16. A. Iida, M. Takahashi, K. Sakurai, and Y. Gohshi, *Rev. Sci. Instrum.* **60**, 2,458–2,461 (1989).
17. N. Gurker, *X-Ray Spectrom.* **14**, 74–83 (1985).
18. N. Gurker, *Adv. X-Ray Analysis* **30**, 53–60 (1987).
19. M. Bavdaz et al., *Nucl. Instrum. Methods* **A266**, 308–312 (1988).
20. G. T. Herman, *Image Reconstruction from Projections*, Academic Press, NY, 1980.
21. N. Shimojo, S. Homma, I. Nakgi, and A. Iida, *Anal. Lett.* **24**, 1,767–1,777 (1991).
22. W. J. M. Lenglet et al., *Histochemistry* **81**, 305–309 (1984).
23. A. Iida and T. Norma, *Nucl. Instrum. Methods* **B82**, 129–138 (1993).
24. Y. Suzuki, F. Uchida, and Y. Hirai, *Jpn. J. Appl. Phys.* **28**, L1,660–(1989).
25. N. Shimojo et al., *Life Sci.* **60**, 2,129–2,137 (1997).
26. S. A. Katz and R. B. Katz, *J. Appl. Toxicol.* **12**, 79–84 (1992).
27. W. M. A. Burgess, L. Diberardinis, and F. E. Speizer, *Am Ind. Hyg. Assoc. J.* **38**, 184–191 (1977).
28. W. E. Atchison and M. F. Hare, *FASEB J.* **8**, 622–629 (1994).
29. M. Aschner and J. L. Aschner, *Neurosci. Biobehaviour Rev.* **14**, 169–176 (1990).
30. Y. Kumagai, S. Homma-Takeda, M. Shinyashiki, and N. Shimojo, *Appl. Organomet. Chem.* **11**, 635–643 (1997).
31. A. J. J. Bos et al., *Nucl. Instrum. Methods* **B3**, 654–659 (1984).
32. I. Orlic, J. Makjanic, and V. Valkovic, *Nucl. Instrum. Methods* **B3**, 250–252 (1984).
33. K. Okmoto et al., *Clin. Chem.* **31**, 1,592–1,597 (1985).
34. S. Osaki, D. A. Johnson, and E. Freiden, *J. Biol. Chem.* **241**, 2,746–2,751 (1966).
35. T. L. Sourkes, *Pharmacol. Rev.* **24**, 349–359 (1972).
36. S. H. Oh, H.E. Ganther, and W. G. Hoekstra, *Biochemistry* **13**, 1,825–1,829 (1974).
37. D. Keilin and T. Mann, *Biochem. J.* **34**, 1,163–1,176 (1940).

38. C. G. Elinder, in L. Friberg, G. F. Nordberg, and V. B. Vouk, eds., *Handbook on Toxicology of Metals*, vol. 2, Oxford, Amsterdam, 1986, p. 664.
39. G. L. Fisher, V. S. Byers, M. Shifrine, and A. S. Levin, *Cancer* **37**, 356–363 (1976).
40. B. F. Issel et al., *Cancer* **47**, 1,845–1,848 (1981).
41. N. Cetinkaya, D. Cetinkaya, and M. Tuce, *Biol. Trace Element Res.* **18**, 29–38 (1988).
42. S. Inutsuka and S. Araki, *Cancer* **42**, 626–631 (1978).
43. M. Hrgovcic et al., *Cancer* **31**, 1,337–1,345 (1973).
44. E. Huhti, A. Poukkula, and E. Uksila, *Respiration* **40**, 112–116 (1980).
45. S. Homma et al., *J. Trace Elements Exp. Med.* **6**, 163–170 (1993).
46. B. Rosoff and H. Spence, *Nature* **207**, 652–654 (1965).
47. S. Homma, I. Nakai, S. Misawa, and N. Shimojo, *Nucl. Instrum. Methods* **B103**, 229–232 (1995).
48. K. Julshamn et al., *Sci. Total Environ.* **84**, 25–33 (1989).
49. R. Scott et al., *Urol. Res.* **11**, 285–290 (1983).
50. N. Koizumi et al., *Environ. Res.* **49**, 104–114 (1989).
51. S. Homma-Takeda et al., *Anal. Lett.* **29**, 601–611 (1996).
52. M. Shinyashiki et al., *Environ. Toxicol. Phamacol.* **2**, 359–366 (1996).
53. S. Homma-Takeda, Y. Kumagai, M. Shinyashiki, and N. Shimojo, *J. Synchrotron Radiat.* **5**, 57–59 (1998).
54. N. Shimojo et al., *J. Occup. Health* **39**, 64–65 (1997).
55. J. A. R. Samson, *Techniques of Vacuum Ultraviolet Spectroscopy*, Wiley & Sons, NY, 1967.
56. G. W. Berkstresser, J. Shmulovich, D. T. C. Huo, and G. Matulis, *J. Electrochem. Soc.* **134**, 2,624–2,628 (1987).
57. G. W. Berkstresser et al., *J. Electrochem. Soc.* **135**, 1,302–1,305 (1988).
58. B. La Fontaine et al., *Appl. Phys. Lett.* **63**, 282–284 (1995).
59. H. A. Hauptman, *Phys. Today* **42**, 24–30 (1989).
60. G. H. Stout and L. H. Jensen, *X-ray Structure Determination*, 2nd ed., Wiley, NY, 1989.
61. R. G. Rossmann, ed., *The Molecular Replacement Method*, Gordon and Breach, NY, 1972.
62. H. Schenk, ed., *Direct Methods for Solving Crystal Structures*, Plenum Press, NY, 1991.
63. M. M. Woolfson and H. -F. Han, *Physical and Non-Physical Methods of Solving Crystal Structures*, Cambridge University Press, Cambridge, 1995.
64. W. A. Hendrickson, *Science* **254**, 51–58 (1991).
65. S. -L. Chang, *Multiple Diffraction of X-Rays in Crystals*, Springer-Verlag, Berlin, 1984.
66. S. -L. Chang, *Acta Crystallogr.* **A54**, 886–894 (1998); also in H. Schenk, ed., *Crystallography Across the Sciences*, International Union of Crystallography, Chester, 1998, pp. 886–894.
67. A. Szoke, in D. T. Attwood and J. Boker, eds., *Short Wavelength Coherent Radiation: Generation and Applications*, AIP Conf. Proc. No. 147, American Institute of Physics, NY, 1986, pp. 361–467.
68. M. Tegze and G. Faigel, *Europhys. Lett.* **16**, 41–46 (1991).
69. G. J. Maalouf et al., *Acta Crystallogr.* **A49**, 866–871 (1993).
70. A. Szoke, *Acta Crystallogr.* **A49**, 853–866 (1993).
71. P. M. Len, S. Thevuthasan, and C. S. Fadley, *Phys. Rev.* **B50**, 11 275–11 278 (1994).
72. M. Tegze and G. Faigel, *Nature* **380**, 49–51 (1996).
73. T. Gog et al., *Phys. Rev. Lett.* **76**, 3,132–3,135 (1996).
74. P. M. Len, T. Gog, C. S. Fadley, and G. Materlik, *Phys. Rev.* **B55**, 3,323–3,327 (1997).
75. P. M. Len et al., *Phys. Rev.* **B56**, 1,529–1,539 (1997).
76. B. Adams et al., *Phys. Rev.* **B57**, 7,526–7,534 (1998).
77. S. Y. Tong, C. W. Mok, H. Wu, and L. Z. Xin, *Phys. Rev.* **B58**, 10 815–10 822 (1998).
78. G. Faigel and M. Tegze, *Rep. Prog. Phys.* **62**, 355–393 (1999).
79. M. Tegze et al., *Phys. Rev. Lett.* **82**, 4,847–4,850 (1999).
80. R. J. Collier, C. B. Burckhardt, and L. H. Lin, *Optical Holography*, Academic Press, NY, 1971.
81. W. Kossel, *Ann. Phys. (Leipzig)* **26**, 533–553 (1936).
82. J. J. Barton, *Phys. Rev. Lett.* **61**, 1,356–1,359 (1988).
83. J. J. Barton, *Phys. Rev. Lett.* **67**, 3,106–3,109 (1991).
84. S. Y. Tong, H. Huang, and C. M. Wei, *Phys. Rev.* **B46**, 2,452–2,459 (1992).
85. S. Thevuthasan et al., *Phys. Rev. Lett.* **70**, 595–598 (1993).
86. G. P. Harp et al., *J. Electron Spectrosc. Relat. Phenomena* **70**, 331–337 (1991).
87. A. Q. R. Baron, *Nucl. Instrum. Methods* **A352**, 665–667 (1995).
88. M. Tegze et al., *Nature* **407**, 38–40 (2000).
89. S. Marchesini et al., *Phys. Rev. Lett.* **85**, 4,723–4,726 (2000).
90. J. Muller et al., *Phys. Lett.* **44A**, 263–264 (1973).
91. U. Fano and J. H.Macek, *Rev. Mod. Phys.* **45**, 553–573 (1973).
92. C. H.Greene and R. N.Zare, *Ann. Rev. Phys. Chem.* **33**, 119–150 (1982).
93. D. W. Lindle et al., *Phys. Rev. Lett.* **60**, 1,010–1,013 (1988).
94. S. H. Southworth et al., *Phys. Rev. Lett.* **67**, 1,098–1,101 (1991).
95. Y. Ma et al., *Phys. Rev. Lett.* **74**, 478–481 (1995).
96. J. A. Carlisle et al., *Phys. Rev. Lett.* **74**, 1,234–1,237 (1995)
97. C. K. Chen et al., Paper 14P9, *Third Conf. Asian Crystallogr. Assoc.*, Kuala Lumpur, 1998.
98. H. Helm et al., *Phys. Rev. Lett.* **70**, 3,221–3,224 (1993).
99. J. J. Yeh, *Atomic Calculation of Photoionization Cross-sections and Asymmetry Parameters*, Gordon & Breach, NY, 1993.
100. E. B. Saloman, J. H. Hubbel, and J. H. Scofield, *Atom. Data Nucl. Data Tables* **38**, 1–51 (1988).

X-RAY TELESCOPE

WEBSTER CASH
University of Colorado
Boulder, CO

An X-ray telescope is an optic that is used to focus and image x-rays like a conventional telescope for visible light astronomy. X-ray telescopes are exclusively the domain of X-ray astronomy, the discipline that studies high-energy emissions from objects in space. A telescope, by its very nature, is for studying objects at large distances, concentrating their radiation, and magnifying their angular extent. Because X rays cannot travel large distances through the earth's atmosphere, they can be

used only in spacecraft, observing targets through the vacuum of space.

X rays penetrate matter and are used to image the interior of the human body. Unfortunately, this means that X rays also tend to penetrate telescope mirrors, making the mirror worthless for astronomy. But, by using of a specialized technique called grazing incidence, telescopes can be made to reflect, and astronomy can be performed. In this article, we describe the basic techniques used to build X-ray telescopes for X-ray astronomy.

Because X-ray astronomy must be performed above the atmosphere, it is a child of the space program. Telescopes are carried above the atmosphere by rockets and used to study the sun, the planets, and objects in the depths of space. These telescopes have provided a different view of the hot and energetic constituents of the universe and have produced the key observations to establish the existence of black holes.

X-RAY ASTRONOMY

Spectra of Hot Objects

X rays are created by the interactions of energetic charged particles. A sufficiently fast moving electron that impacts an atom or ion or is accelerated in a magnetic or electric field can create high-frequency radiation in the X-ray band. Thus X rays tend to be associated with objects that involve high-energy phenomena and high temperatures.

In general, if radiation is generated by thermal processes, the characteristic frequency emitted will be given by

$$h\nu \approx kT \tag{1}$$

The X-ray band stretches from 10^{16}–10^{18} Hz, which indicates that the characteristic temperatures of objects range from 10^6 K up to 10^8 K. At these extreme temperatures, matter does not exist in the forms we see in everyday life. The particles are moving so fast that atoms become ionized to form plasma. The plasma can have very low density as in the blast wave of a supernova, or very high density under the surface gravity of a neutron star. Many stars, including the sun, have a hot, X-ray emitting gas around them, called a corona. For example, the X-ray spectrum of the corona of the star HR1,099 is shown in Fig. 1.

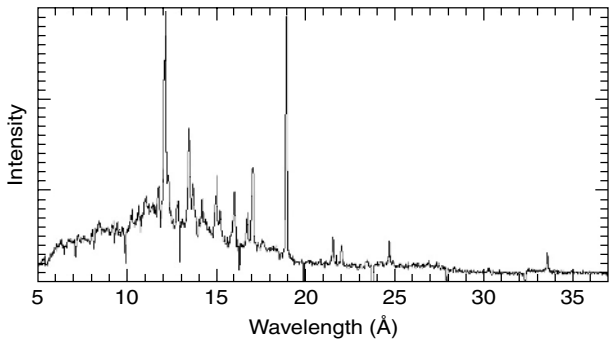


Figure 1. X-ray spectrum of the star HR1,099.

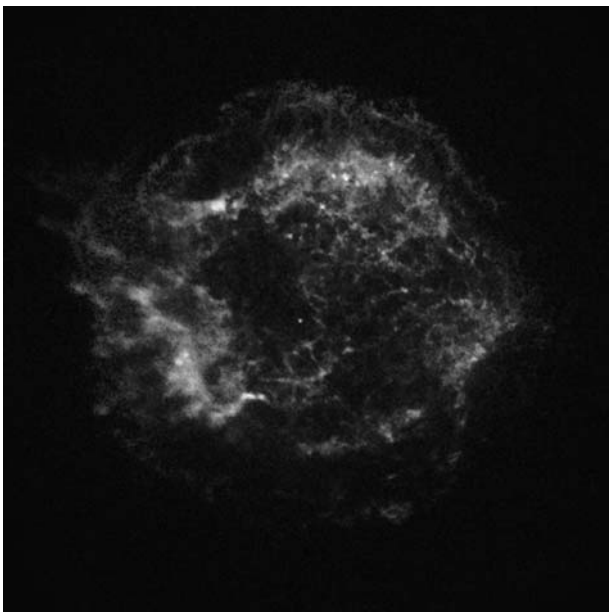


Figure 2. Image of the X rays emitted from the 300-year-old supernova explosion remnant named Cas-A. The picture shows an expanding shock front of interstellar gas and the collapsed remnant star in the middle. This image was acquired using the 1.2-m diameter X-ray telescope on the Chandra Satellite. See color insert.

Dramatic events in the universe, such as supernovae, also generate X rays. The shock waves from the exploding star heat as they pass through interstellar space to create supernova remnants, an expanding shell of hot plasma, as shown in Fig. 2.

Another way by which X rays are created is through the acceleration of matter near the surfaces of collapsed stars, including white dwarfs, neutron stars, and black holes. As matter spirals into these extreme objects, it heats to high temperatures and emits X rays. X-ray telescopes, through imaging, timing, and spectroscopy, have played a central role in proving the existence of black holes and in studying the physics of the most extreme objects in the universe.

Interstellar Absorption

The space between stars is a vacuum, better than the best vacuum ever created in a laboratory on earth. But in the immense stretches of interstellar space, there is so much volume that the quantities of gas become large enough to absorb X rays. In our part of the galaxy, there is an average of about one atom of gas for every 10 cc of space. The composition of this gas is similar to that of the sun, mostly hydrogen and some helium mixed in. There is also a small, but significant quantity of other heavier elements such as oxygen and carbon.

Across the distances between the stars, the absorption caused by these stars can become significant in the soft X-ray band. In Fig. 3, we show a graph of the transmission of interstellar gas as a function of X-ray energy. It shows that low-energy X rays have higher absorption, so that one cannot see as far through the galaxy at 0.1 keV, as at 1 keV. It is for this reason that x-ray telescopes are usually

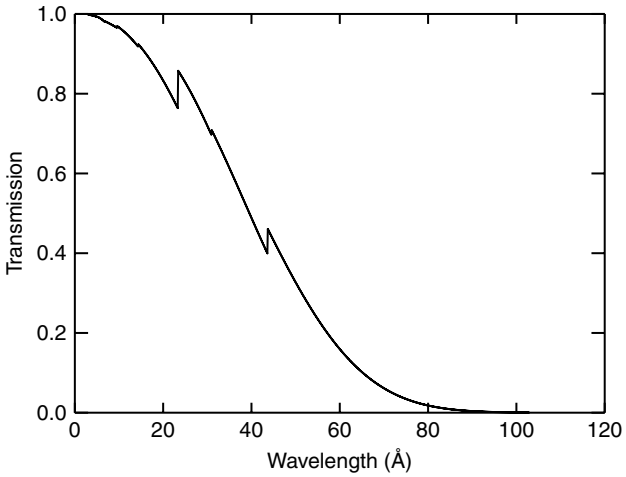


Figure 3. The thin gas between the stars can absorb X rays, particularly at low energies. This graph shows the transmission of the interstellar medium as a function of X-ray wavelength for the amount of interstellar gas expected at 2,000 parsecs.

designed to function above 0.5 keV and up to 10 keV, if possible.

Atmospheric Absorption

Because of absorption, X rays cannot penetrate the earth's atmosphere. The primary interaction is by the photoelectric effect in the oxygen and nitrogen that are the main constituents of our atmosphere. Thus, to use an X-ray telescope, we must be above most of these gases.

In Fig. 4, we show the transmission of the atmosphere at an altitude of 110 km, as it becomes partially transparent. A soft X-ray telescope requires a rocket to gain sufficient altitude. Whether by suborbital rocket to view above the atmosphere quickly for five minutes, by a larger missile carrying a satellite to orbit, or by a larger launcher yet carrying the telescope to interplanetary space, the rocket is essential.

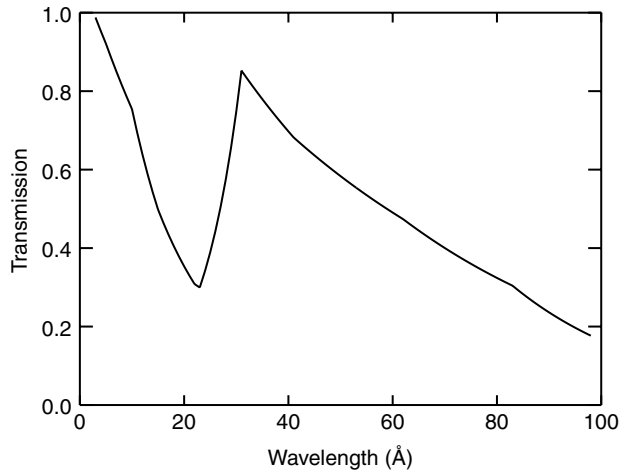


Figure 4. This shows the fraction of X rays transmitted from overhead down to an altitude of 110 km above sea level as a function of X-ray wavelength.

For hard X rays, however, there is a more modest means of achieving altitude —the balloon. Large balloons can carry telescopes to altitudes above 100,000 feet, which is sufficient to observe the harder X rays. Balloons can stay up for hours, days, and even weeks, compensating for the lower flux of observable signal.

Signal-to-Noise Issues

Telescopes were not used in the early days of X-ray astronomy. The first observations of the sky were performed using proportional counters. These are modified Geiger counters that have large open areas and no sensitivity to the direction from which the photon came. Grid and slat collimators were used to restrict the solid angle of sky to which the detector was sensitive, so that individual sources could be observed.

Collimators were built that achieved spatial resolution as fine as 20 arcminutes, which was adequate for the first surveys of the sky and for studying the few hundred brightest X-ray sources in the sky. However, the detectors had to be large to collect much signal, and that led to high detector background. The large angular extent caused signal from the sky to add to the background as well. Additionally, many faint sources could be unresolved in the field of view at the same time, leading to confusion. If X-ray astronomy were to advance, it needed telescopes to concentrate the signal and resolve weak sources near each other in the sky.

GRAZING INCIDENCE

X rays are often referred to as "penetrating radiation" because they pass easily through matter. The medical applications of this property revolutionized medicine. However, when the task is to build a telescope, it becomes necessary to reflect the radiation rather than transmit or absorb it. The fraction of 1-keV X rays reflected from a mirror at normal incidence can be as low as 10^{-10}, effectively killing its utility as an optic.

Another problem of conventional mirrors is their roughness. If a mirror is to reflect radiation specularly, it needs to a have surface roughness substantially lower than the wavelength of the radiation. In the X ray, this can be difficult, considering that the wavelength of a 10-keV X ray is comparable to the diameter of a single atom.

These problems would have made X-ray telescopes impossible, except for the phenomenon of grazing incidence, in which an X ray reflects off a mirror surface at a very low angle, like a stone skipping off a pond (Fig. 5).

Grazing Incidence

The *graze angle* is the angle between the direction of the incident photon and the plane of the mirror surface. In



Figure 5. An X ray approaches a mirror at a very low graze angle, reflecting by the property of total external reflection.

most optical notation, the *incidence angle* is the angle between the direction of the ray and the normal to the plane of the mirror. Thus, the graze angle is the complement of the incidence angle.

At any given X-ray wavelength, there is a *critical angle*, below which the X rays reflect. As the graze angle drops below the critical angle, the efficiency of the reflection rises. As the energy of the X ray rises, the critical angle drops, so hard X-ray optics feature very low graze angles. In general the critical angle θ_c is given by

$$\sin\theta_c = \lambda\left(\frac{e^2N}{mc^2\pi}\right)^{1/2}, \tag{2}$$

where λ is the wavelength of the X ray and N is the number of electrons per unit volume (1).

This behavior comes about because the index of refraction inside a metal is less than one, to an X ray. The X ray interacts with the electrons in the reflecting metal as if it had encountered a plasma of free electrons. The wave is dispersed and absorbed as it passes through the metal, allowing the index of refraction to fall below one. This process can be described by assigning a complex index of refraction to the material (2).

The index of refraction n of a metal is given by

$$n = 1 - \delta - i\beta, \tag{3}$$

where the complex term β is related to the absorption coefficient of the metal. If β is zero, then δ cannot be positive.

A well-known optical effect in the visible is *total internal reflection*, which leads to the mirror-like reflection of the glass in a fish tank. If the radiation approaches the glass at an angle for which there is no solution to Snell's law, it is reflected instead of transmitted. This happens when the radiation attempts to pass from a medium of higher index of refraction to one of lower index. In the X ray, where the index of refraction in metals is less than one, *total external reflection* is experienced when the ray tries to pass into a metal from air or vacuum where the index of refraction is closer to unity.

Fresnel Equations

The equations for reflection of X rays are the same as those for longer wavelength radiation and are known as the Fresnel equations (2). In the X ray, evaluation of the equations differs from the usual because the index of refraction has an imaginary component, so the algebra becomes complex. If radiation approaches at an angle φ_i with respect to the normal (φ, the incidence angle, is the complement of θ, the graze angle), then some of the radiation will reflect at an angle φ_r, which is equal to φ_i. Some of the power will be transmitted into the material at an angle φ_t with respect to the normal, where φ_t is given by Snell's law:

$$\sin\varphi_t = \frac{\sin\varphi_i}{n} \tag{4}$$

but now φ_t is complex, and evaluation involves complex arithmetic.

The reflectance of the transverse electric wave (one of the two plane polarizations) is given by

$$R_E = \left(\frac{\cos\varphi_i - n\cos\varphi_t}{\cos\varphi_i + n\cos\varphi_t}\right)\left(\frac{\cos\varphi_i - n\cos\varphi_t}{\cos\varphi_i + n\cos\varphi_t}\right)^*, \tag{5}$$

and the transverse magnetic wave (the other polarization) is given by

$$R_M = \left(\frac{\cos\varphi_i - \cos\varphi_t/n}{\cos\varphi_i + \cos\varphi_t/n}\right)\left(\frac{\cos\varphi_i - \cos\varphi_t/n}{\cos\varphi_i + \cos\varphi_t/n}\right)^*, \tag{6}$$

where now the star on the second term represents the complex conjugate.

Extensive work has been done over the years to tabulate the index of refraction of X rays in a wide variety of optical materials and elements. Tabulations can be found in the literature (3).

Certainly, the most common material used as a coating in the X ray is gold, which is stable and easy to deposit. A few other materials can be better, including platinum and osmium at high energies and nickel and uranium below 1 keV. In Fig. 6, we show the reflectivity of gold as a function of X-ray energy. From perusal of the chart, one can see that in the quarter kilovolt band (0.1–0.25 keV), graze angles as high as 5° are possible. At 1 keV, angles close to 1° are required, and at 10 keV, angles below half a degree are necessary.

Mirror Quality

The principles of grazing incidence provide a technique for the efficient reflection of X rays, but the mirrors must be of adequate quality. Analogous to all telescope mirrors, they must have good figure, good polish, and adequate size to suppress diffraction (4).

In Fig. 7, we show the reflection of the ray in three dimensions. The direction of the ray can be deflected to the side by a slope error in the "off-plane" direction.

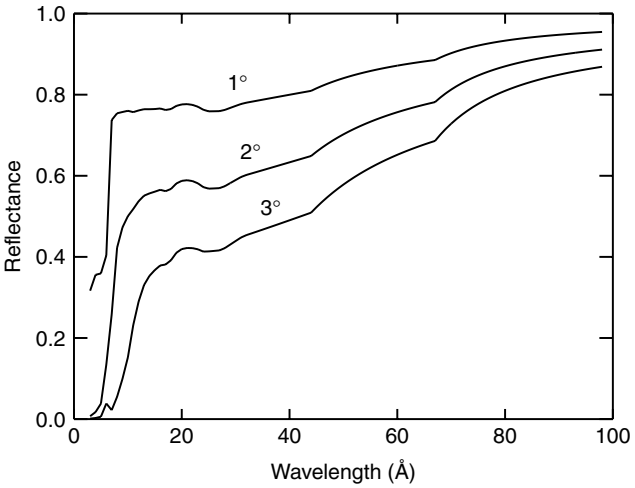


Figure 6. Plot of the fraction of X rays reflected off a polished gold surface as a function of the wavelength of the X ray. Three graze angles are shown. One can see the drop in efficiency as the angle rises.

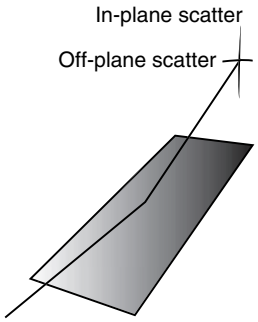


Figure 7. Symmetry is broken by a grazing incidence reflection. Scatter is much worse in the plane of the incident and reflected ray. There is scatter off-plane, but it is much lower.

However, as one can see from the diagram, to change the direction of the ray that reflects at angle θ by an angle close to θ requires a slope error of the order of 30°. The effective angle through which the ray is thrown is reduced by a factor of $\sin\theta$ in the off-plane direction. This means that errors in mirror quality lead to greater error in the in-plane direction. The resulting image is blurred anisotropically and creates a long, narrow image in the vertical direction (4). The height of the image blur is roughly $1/\sin\theta$ times the width.

Microscopic roughness can also degrade the image by scattering the X rays. Scatter is a problem at all wavelengths, but it is particularly severe in the X ray region where the radiation has a short wavelength. Scatter is caused by small deviations in the wave fronts of the reflected light. A deviation in height on the surface of the mirror of size δ will create a deviation in phase of size $2\delta\sin\theta$, where θ is the graze angle. This means that the amount of scattering drops with the sine of the graze angle, that allows relaxation of polish requirements. If one assumes that the surface roughness can be described as a probability function that has Gaussian distribution, then the total amount of scatter will be given by

$$S = e^{-\left(\frac{4\pi\sigma\sin\theta}{\lambda}\right)^2}, \tag{7}$$

where σ is the standard deviation of the surface roughness (5). This means, for example, that for a surface polish of 10-Å, a mirror at a graze angle of 1° can suppress scatter to less than 5%. The angle through which the X ray is scattered is given by

$$\varphi = \lambda/\rho, \tag{8}$$

where ρ is the "correlation length"—the characteristic distance between the surface errors. This equation is similar to the grating equation that governs diffraction and also leads to scattering preferentially in the plane of reflection.

Because X rays are electromagnetic radiation, they are subject to diffraction just like longer wavelength photons. However, because of their very short wavelengths, the effects are not usually noticeable. Visible light telescopes have a diffraction limit that is given by

$$\varphi = 0.7\frac{\lambda}{D} \tag{9}$$

where λ is the wavelength of the light, D is the diameter of the aperture through which the wave passes, and φ is the width of the diffracted beam in radians. The same equation holds true for the X ray, but now D represents the projected aperture of the optic, as viewed by the incoming beam. A grazing incidence optic of length L would create a projected aperture of size $D\sin\theta$, where θ is the graze angle.

A large grazing incidence optic has a length of 50 cm, which indicates that D will be typically 1 cm in projection. Because an X ray has $\lambda = 1$ nm, φ will have a value of 10^{-7} radians, which is 0.02 arcseconds. This resolution is higher than that of the Hubble Space Telescope, so diffraction has not yet become a limiting factor in X-ray telescopes.

WOLTER TELESCOPES

The previous section describes how one can make a mirror to reflect X rays efficiently, but it does not provide a plan for building a telescope. The geometry of grazing incidence is so different from the geometry of conventional telescopes that the designs become radically different in form.

At the root of the design of a telescope is the parabola of rotation. As shown schematically in Fig. 8, parallel light from infinity that reflects off the surface of a parabola will come to a perfect focus. However, this is a mathematical fact for the entire parabola, not just the normal incident part near the vertex. Figure 9 shows a full parabola. The rays that strike the part of the parabola where the slope is large compared to one, reflect at grazing incidence but also pass through the same focus. The mirror can be a figure of rotation about the axis of symmetry of the parabola,

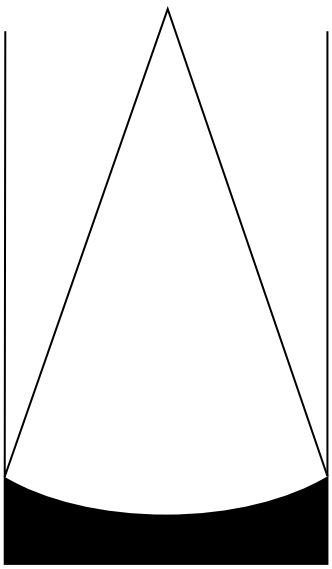


Figure 8. A conventional telescope features a parabolic surface that focuses parallel rays from infinity onto a single point.

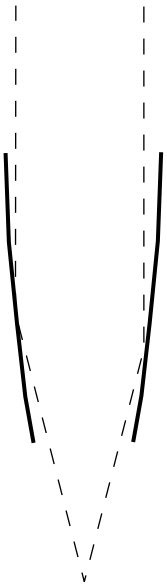


Figure 9. Extending the parabola to a grazing geometry does not interfere with its focusing properties.

just as at normal incidence. Such a figure of revolution is called a paraboloid.

Thus, a paraboloidal mirror can reflect at grazing incidence near the perimeter and at normal incidence near the center. But, because the reflectivity is low near the center, in practice the mirrors are truncated, as shown in Fig. 9. The resulting shape resembles a circular wastepaper basket that has a polished interior. The aperture of the telescope is an annulus.

Because the telescope is a simple parabola, the surface can be described by

$$Z(r) = \frac{r^2}{2\rho} - \frac{\rho}{2}, \tag{10}$$

where $Z(r)$ is the height of the telescope above the focal position at a radius r. The parameter ρ is the radius of curvature at the vertex, which is never part of the fabricated paraboloid.

As such, it is merely a formal parameter of the system. In a normal incidence telescope, the radius of curvature is approximately twice the focal length, but in a grazing parabola, the focal length is the distance from the point of reflection to the focal point.

This causes a major problem for the performance of the telescope. Radiation that enters the telescope on the optic axis is theoretically concentrated into a perfect focus, but that approaching off-axis is not. The effective focal length of the ray is the distance between the point of reflection and the focus, which is approximately equal to Z. Because the paraboloid usually is long in the Z direction, the focal length is a highly variable function of the surface position. This is called comatic aberration. It is so severe that it limits the value of the paraboloid as an imaging telescope. Such paraboloids have been used, but mostly for photometric and spectroscopic work on bright, known targets, where field of view is unimportant.

The severity of this problem was recognized early in the history of X ray optics and was solved by adding a second reflection. In 1952, Wolter (6) showed that two optical surfaces in sequence could remove most of the coma, allowing for the design of a true imaging telescope. It is a mathematical property of the hyperbola that light converging to one focus will be refocused onto the other. Thus, by placing the focal point of the paraboloid at one of the foci of a hyperboloid, the light will be refocused, as shown schematically in Fig. 10. In the process, the distances traveled by the rays after reflection become close to equalized, and the comatic aberration is reduced.

Wolter described three types of these paraboloid-hyperboloid telescopes, as shown in Fig. 11. Each has a paraboloidal surface followed by a hyperboloidal surface. The first reflection focuses parallel light to a point. The second reflection redirects the light to a secondary focus. In X-ray astronomy, we use mostly Type I because the angles at which the rays reflect are additive and create a shorter distance to the focal plane. When the graze angles are low, this can be very important. Sometimes, in the extreme ultraviolet, where graze angles can be in the range of 5–15°, the Wolter Type 2 has been used (7).

The equations for the paraboloid-hyperboloid telescopes are

$$z_1 = \frac{r_1^2}{2\rho} - \frac{\rho}{2} - 2\sqrt{a^2 + b^2} \tag{11}$$

and

$$z_2 = \frac{a}{b}\sqrt{b^2 + r_2^2} - \sqrt{a^2 + b^2}. \tag{12}$$

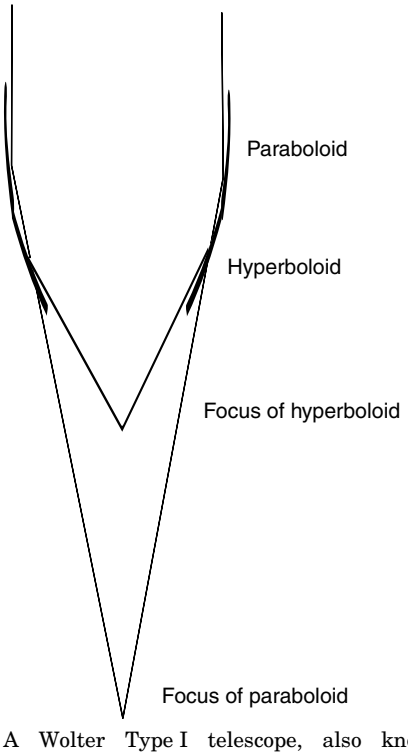


Figure 10. A Wolter Type I telescope, also known as a paraboloid-hyperboloid telescope, features two reflections. The first focuses the light to a distant point, the second reflection, off the hyperboloid, refocuses to a nearer point and provides a wider field of view.

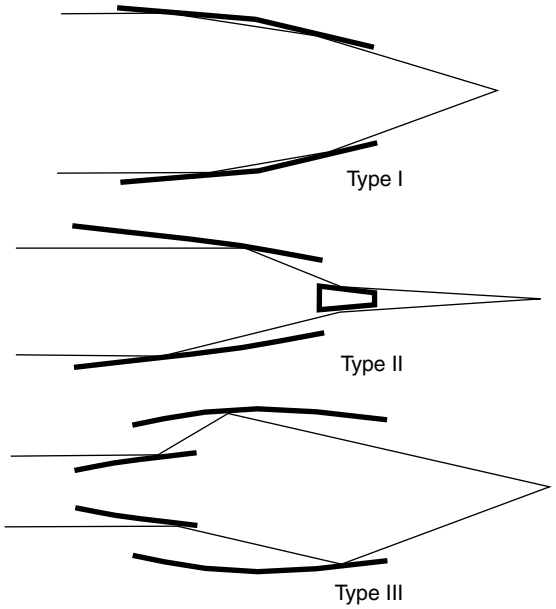


Figure 11. There are three types of Wolter telescopes shown. The Wolter Type I is dominant because it shortens the focal length and allows nesting. Type II is used for spectroscopic experiments at the longest X-ray wavelengths, where graze angles are higher.

Thus, three parameters define the surfaces of the optics, and the designer must additionally define the range of z_1 and z_2 over which the optic will be built. During the design process, care must be taken to ensure that the proper range of graze angles is represented in the optic, so that the spectral response of the telescope is as expected. This is usually done by computer ray tracing for field of view and throughput.

In 1953, Wolter extended the theoretical basis for grazing incidence telescopes by applying the Abbe sine condition in the manner of Schwarzschild to create the Wolter–Schwarzschild optic (8). This optic is a double reflection that has three types analogous to the paraboloid-hyperboloids, but have, theoretically, perfect lack of coma on-axis.

The Wolter–Schwarzschild is described by the parametric equations,

$$r_1 = F \sin\alpha, \tag{13}$$

$$z_1 = \frac{-F}{C} + \frac{FC\sin^2\alpha}{4} + \frac{F}{R}\left[1 - C\sin^2(\alpha/2)\right]^{\left(\frac{2-C}{1-C}\right)} \times \cos^{\frac{-2C}{1-C}}(\alpha/2), \tag{14}$$

$$z_2 = d\cos\alpha, \tag{15}$$

$$r_2 = d\sin\alpha, \tag{16}$$

where

$$\frac{1}{d} = \frac{C}{F}\sin^2(\alpha/2) + \frac{R}{F}\left[1 - C\sin^2(\alpha/2)\right]^{\left(\frac{-C}{1-C}\right)} \times \cos^{\left(\frac{2}{1-C}\right)}(\alpha/2). \tag{17}$$

For this system, a ray that approaches on-axis, strikes the primary at (r_1, z_1), reflects, and strikes the secondary at (r_2, z_2), and then approaches the focus (located at the origin) at an angle of α off-axis. The parameter F is the effective focal length of the system and has units of length. C and R are dimensionless shape parameters.

In practice, as the graze angle becomes small, the performance advantage of the Wolter–Schwarzschild over the paraboloid-hyperboloid becomes small. So often added complexity is avoided, and simple paraboloid-hyperboloids are built.

One advantage of the type I Wolter telescopes is that their geometry allows nesting. Because the effective focal length of each optic is approximately the distance from the plane at the intersection of the paraboloids and hyperboloids to the focus, a series of Wolters of different diameter can be nested, one inside the other, to increase the effective area of the telescope. This implies that the outer nested pairs must reflect X rays at higher graze angles.

The actual design process for X ray telescopes is usually performed by computer ray tracing. The resolution as a function of angle off-axis is difficult to estimate in closed form. Furthermore, the reflectivity of the X rays is a function of angle, and the angle can change substantially across the aperture, meaning that not all parts of the aperture have equal weight and that the response is a function of the energy of the incident radiation. By ray tracing, the resolution and throughput can be evaluated as a function of X-ray energy across the field of view of the telescope.

THIN MIRROR TELESCOPES

For the astronomer, the ability to study faint objects is crucial, meaning that a large collecting area is of primary importance. Wolter telescopes can be nested to enhance the total collecting area, as described in the previous section. However, the thickness of the mirrors employed occults much of the open aperture of a typical Wolter. The ideal Wolter telescope would have mirrors of zero thickness, so that all of the aperture could be used. Thus, mirrors made from thin, densely packed shells can significantly enhance capability. This is also important because thin mirrors allow the maximum collecting area for a given weight. Because these mirrors are to be launched into space aboard rockets, maximizing the ratio of collecting area to weight can be of central concern. This is especially true at high energies, where graze angles are low and the mass of mirror needed to collect X rays rises.

Unfortunately, as mirrors are made thinner, they become less rigid and less able to hold a precise optical form. Thus a telescope made of thin, closely packed mirrors is likely have lower quality. Additionally, paraboloid and hyperboloid mirrors are expensive to fabricate in large numbers on thin backings.

One solution has been to make the mirrors out of thin, polished foils. When a thin (often well below a millimeter) foil is cut to the right shape and rolled up, it can form a conical shape that approximates a paraboloid or hyperboloid, as shown in Figure 12. Two of these, properly

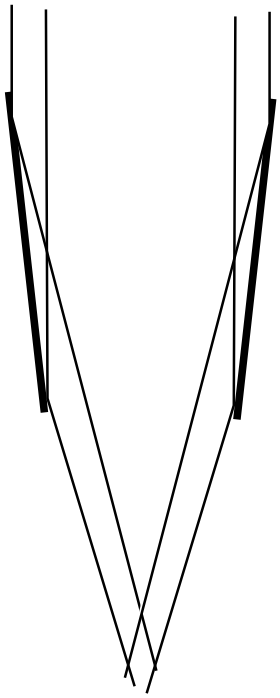


Figure 12. The paraboloid of Fig. 9 can be approximated by a cone. This allows fabricating low-cost, low-quality optics.

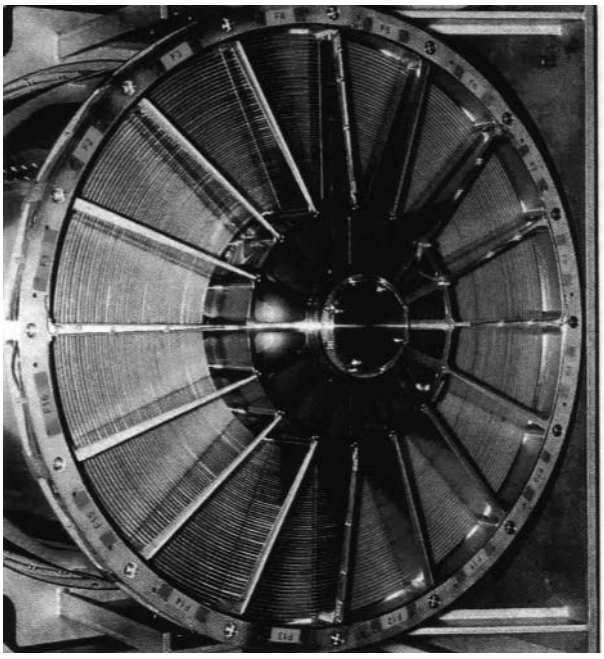


Figure 13. Photograph of an X-ray telescope that has densely nested thin shell mirrors to build a large collecting area.

configured, then approximate a Wolter. Because there is no sag to the mirrors, the rays do not come to a theoretically perfect point focus. Instead, they form a spot of diameter equal to the projected width of the conical mirror annulus.

If the goal of the builder is to create a focus of modest quality that has a large collecting area, especially at high energies, then this is an acceptable trade. A number of such telescopes have been built (Fig. 13), usually for spectroscopy. In X-ray spectroscopy, the targets that are bright enough to be studied tend to be far apart in the sky, so low resolution is not an issue. Also, because the background signal from the sky is very low, there is no loss of signal-to-noise by letting the focal spot size grow.

An alternative to simple cones has been to use electroforming techniques to make thin shell mirrors. In the electroforming process, one builds a "mandrel" which is like a mold, the inverse of the desired optic. The mandrel is machined from a piece of metal and then polished. When placed in an electroforming bath, a shell of metal (usually nickel) is built up around the mandrel by electrochemical processes. When the shell is sufficiently thick, it is removed from the bath and separated from the mandrel. The disadvantage of this process is that many mandrels must be made. The advantage is that many copies are possible at low cost. Because the mandrel is machined, a sag can easily be added to approximate the ideal Wolter shape more closely.

Metals (such as nickel) that are used in electroforming tend to have high density and lead to heavy telescopes. Sometimes, replication is used to reduce weight. A shell of some lightweight material like carbon-fiber epoxy is built, that approximates the desired shape. The mandrel is covered by a thin layer of epoxy, and the shell placed over it. The epoxy dries to the shape and polish of the mandrel. When the replica is pulled away, it provides a mirror that has low weight and good imaging properties.

KIRKPATRICK–BAEZ TELESCOPES

The first imaging optic built using grazing incidence was an X-ray microscope built by Kirkpatrick and Baez in 1948 (9). (At the time, the space program had not really started, and only the sun was known to be a source of X rays. So there was no immediate application for X-ray telescopes.) The optic consisted of two, standard, spherically shaped mirrors in sequence, which, together, formed an imaging optic, as shown in Fig. 14. Parallel light incident onto a mirror that has a concave, spherically shaped surface of curvature R and graze angle θ will come to a line focus a distance $(R/2)\sin\theta$ from the mirror, parallel to the mirror surface. There is very little focusing in the other dimension. Because we usually want a two-dimensional image, a second reflection is added by placing the second mirror oriented orthogonally, beyond the first. If the two mirrors are properly placed, then both dimensions focus at the same point.

Kirkpatrick and Baez used spheres because they are readily available, but spheres have severe on-axis coma, and thus can have poor focus. The usual solution for

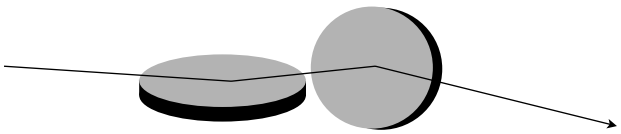


Figure 14. The Kirkpatrick–Baez optic is the first focusing X-ray optic made. It features two flat (or nearly flat) mirrors that reflect X rays in sequence. Each mirror provides focus in a different dimension.

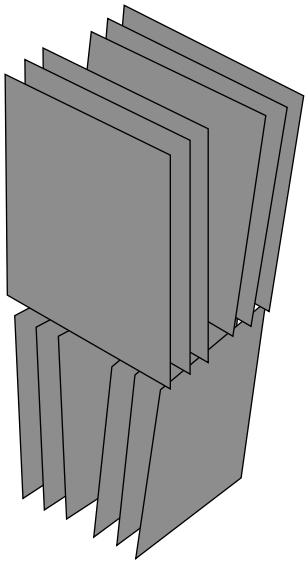


Figure 15. To achieve a large collecting area, the Kirkpatrick–Baez telescopes require many co-aligned mirrors.

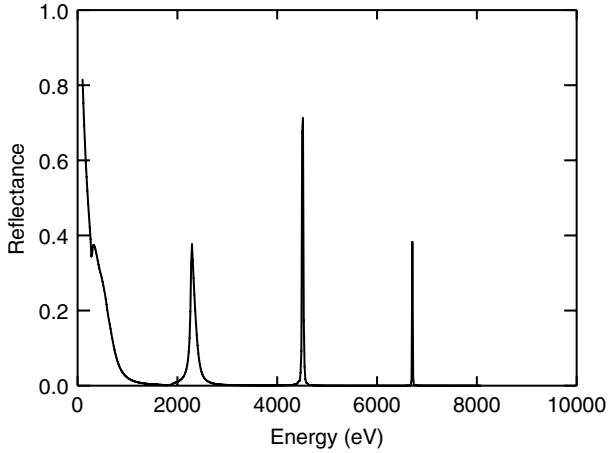


Figure 16. Multilayer-coated mirrors can provide high reflectivity, where otherwise there would be none. In this graph, we show the response of a multilayer-coated mirror at a 3° graze. To the left, at low energy, the X rays reflect as usual. The multilayer provides narrow bands of high reflectivity at higher energies. Three Bragg orders are visible in this plot.

making a telescope is to replace the sphere by a one-dimensional paraboloid. The paraboloid, which can be made by bending a sheet of glass or metal, has the geometric property of focusing the on-axis light to a perfect line focus. The paraboloids can then be nested to add extra signal into the line focus. To achieve a two-dimensional image, another set of co-aligned paraboloids must be placed after the first set, rotated 90° around the optic axis, as shown in Fig. 15. This creates a two-dimensional image.

Using this "Kirkpatrick–Baez" geometry, one can build telescopes that have very large collecting areas, suitable for studying faint sources. The disadvantages of these telescopes are that they have relatively poor resolution, typically no better than 20 arcseconds, and have a fairly small field of view due to comatic aberration.

MULTILAYER TELESCOPES

A multilayer coating can be considered a synthetic version of a Bragg crystal. It consists of alternating layers of two materials deposited on a smooth substrate. Typically, one material has a high density and the other a low density to maximize the change in the index of refraction at the material interface. If the radiation is incident on a multilayer of thickness d for each layer pair at an angle θ, then constructive interference is experienced if the Bragg condition,

$$m\lambda = 2d\cos\theta, \tag{18}$$

is met. This creates a narrow-wavelength band where the reflectance of the surface is much higher than using a metal coating alone. In Fig. 16, we show the reflectance of a multilayer mirror that consists of alternating layers of tungsten and silicon as a function of incident energy. This narrow energy response can be tuned to the strong emission lines of an object like a star but leads to the absorption of most of the flux from a continuum source.

By placing the multilayer on the surface of a conventional, normal incidence mirror, it becomes possible to create a narrow band of good reflectivity at normal incidence, where before there was none. This is usually applied in the extreme ultraviolet, but multilayers now work effectively up to 0.25 keV and even 0.5 keV. This approach has been used to excellent advantage in the study of the sun (10).

The spectral band in which most X-ray telescopes have functioned is between 0.1 and 3.0 keV. Above 3 keV, the required graze angle becomes so small that the aperture annulus becomes small as well. The problem is compounded by the relatively low flux of the sources. Thus, to build a telescope, either we live with low efficiency or find a better way to improve the collecting area.

Multilayers can be used to enhance reflectivity at grazing incidence. Higher energy radiation can be reflected at any given graze angle, but the narrow spectral response is difficult to match to the changing graze angles of the surface of a Wolter telescope. It has now been shown (11) that by varying the thickness of the multilayers as a function of the depth, a broad band response can be created, making the multilayers more useful for spectroscopy and the study of continuum sources. This effect works particularly well for hard X-ray telescopes, where the penetrating power of the X rays is higher, leading to interaction with a larger number of layers. The next generation of hard X-ray telescopes will probably use this effect.

X-RAY INTERFEROMETRIC TELESCOPES

Telescope resolution is limited by diffraction. Diffraction in radio telescopes is so severe that most new major observatories are interferometric. By linking together the signals from the telescopes without losing the phase information, synthetic images may be created to match the resolution of a single giant telescope whose diameter is equal to the baseline of the interferometer. The need for

interferometry comes from engineering realities. At some point, a telescope simply becomes so large that either it cannot be built or the expenses cannot be met.

The largest, practical mirrors of quality sufficient for X rays are about 1 m long. At a graze angle of 2°, the entrance aperture of a 1-m mirror is about 3 cm; this means that a 1-keV signal will be diffraction-limited at about 10 milliarcseconds (0.01"). Thus, conventional X-ray telescopes can achieve very high resolution, ten times that of the Hubble Space Telescope, before the diffraction limit becomes a serious problem. But, there is much exciting science beyond the X-ray diffraction limit. For example, at 10^{-3} arcseconds, it is possible to image the corona of Alpha Centuari, and at 10^{-6} arcseconds, it will be possible to resolve the event horizons of black holes in nearby active galactic nuclei.

To build an X-ray interferometer that can accomplish these goals may, seem impossible at first, but the properties of grazing incidence, coupled with the properties of interferometers, provide a pathway. First, one must achieve diffraction-limited performance in two or more grazing incidence mirrors and then combine the signals in a practical way to achieve a synthetic aperture.

Such a system has now been built and demonstrated in the laboratory (12), but, although there are plans to fly an X-ray interferometer in the 2,010 time frame, none has yet been launched. In theory, paraboloid-hyperboloid telescopes can reach the diffraction limit but in practice are too expensive. So, flat mirrors are used to ease the mirror figure and polish requirements. Figure 17 shows a schematic of a practical X-ray interferometer that has produced fringes in the laboratory and is being scaled up for flight.

The idea is provide two flat, grazing incidence mirrors set at an arbitrary separation that direct the X rays into a beam combiner. The separation of these two mirrors sets the resolution of the interferometer. The two beams are then mixed by a beam combiner. This has traditionally been accomplished for X rays by using Laue crystals (13), but the thickness of the crystals, coupled with their low efficiency, makes them impractical for astronomy.

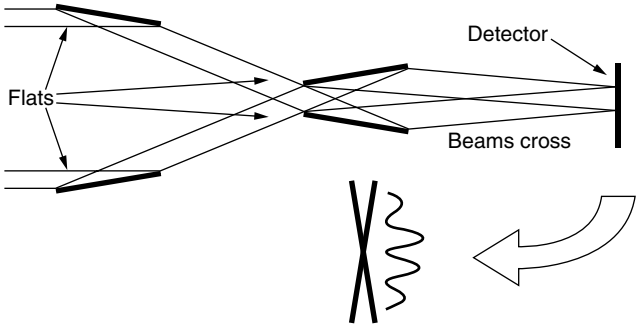


Figure 17. A simple X-ray interferometer, capable of synthetic aperture imaging may be built in this configuration. A pair of flat mirrors at grazing incidence causes the wave fronts to cross. Another pair of flat, grazing incidence mirrors redirects the light to an almost parallel geometry, where the beams cross at a large distance. Because of the low angles at which the wave fronts cross, fringes much larger than the wavelength of the X rays can be created.

A solution that provides broad spectral response and high efficiency is simply to use two more flat mirrors. The beams cross and then strike the second set of flat mirrors at a graze angle just slightly higher than the initial reflection. This brings the two beams (still plane parallel) back together at a low angle.

As Fig. 17 shows, if the two beams are coherent when they cross, fringes will appear. The spacing s of the fringes is given by

$$s = \frac{L\lambda}{d}, \tag{19}$$

where d is the separation of the second set of mirrors and L is the distance from the mirrors to the focal plane where the beams cross. If L/d is sufficiently large, then the fringes can be resolved by a conventional detector. For example, if L/d is 100,000, then the fringe spacing from 10-Å X rays will be 100 μ, easily resolved by most detectors. This approach to beam combination is highly practical because it uses flat mirrors and has high efficiency. It also works in a panchromatic way. Each wavelength of radiation creates fringes at its own spacing, so, if the detector can resolve the energy of each photon, the individual sine waves will be resolved, and the interferometer will function across a wide spectral band.

A single pair of mirror channels is inadequate in an interferometer for X-ray astronomy. In the early days of radio interferometry, a single pair was used, and the UV plane was sampled as the source drifted across the sky. However, many of the most interesting X-ray sources are highly variable, and it may not be possible to wait for a change of orientation. Thus, a substantial portion of the UV plane needs sampling simultaneously, effectively requiring that more than two channels mix at the focal plane. One attractive geometry that is being pursued is to use a ring of flats, as shown in Fig. 18. This can be considered a dilute aperture telescope. Because each of the many mirror paths interferes against all of the others, there is a large range of sampling in frequency space, and the beam pattern starts to resemble that of a telescope.

GRATINGS

The analysis of spectra is a central tool of the X-ray astronomer, just as it is in other bands of the spectrum.



Figure 18. The interferometer of Fig. 17 can be made more powerful by placing multiple sets of flat mirrors in a ring that feeds a common focus.

X-ray astronomy differs from other fields in that much of the spectroscopy is performed by using energy-sensitive, photon-counting detectors. The first X-ray observatories had no optics but could still perform low-resolution spectroscopy by using proportional counters. These devices measured the number of secondary ionization events caused by an X ray and thus led to an energy estimate of each photon. However, the spectral resolution ($R = \lambda/\delta\lambda$) was only about 5 at 1 keV. Solid-state systems, as exemplified by CCDs, now reach $R = 20$ at 1 keV, and quantum calorimeters have a resolution of several hundred and are still improving.

To achieve very high spectral resolution, the X-ray astronomer, just like the visible light astronomer, needs a diffraction grating. Similarly, diffraction gratings come in two categories. Transmission gratings allow the photons to pass through and diffract them in the process; reflection gratings reflect the photons at grazing incidence and diffract them in the process.

Transmission

Transmission gratings are, in essence, a series of thin, parallel wires, as shown in Fig. 19. The space between the wires is ideally empty to minimize absorption of X rays. The wires should be optically thick to absorb the flux that strikes them. This creates a wave-front amplitude on the far side of the grating that is shaped like a repeating square wave.

At a large distance from the grating, constructive interference can be experienced where the wave satisfies the "grating equation" (1),

$$n\lambda = d \sin \alpha, \tag{20}$$

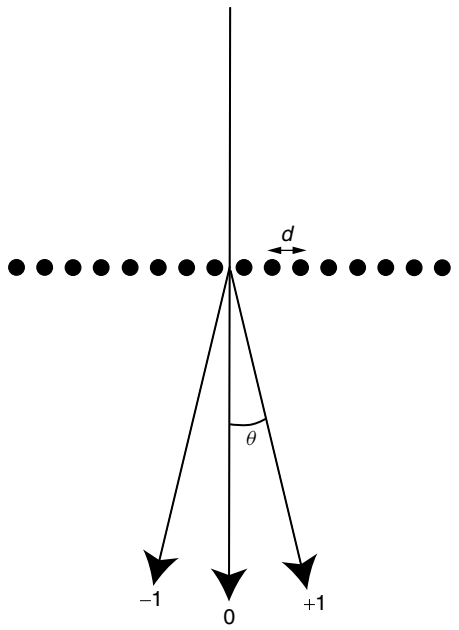


Figure 19. An X-ray transmission grating is made of many finely spaced, parallel wires. Diffraction of X rays results at wire densities of up to 10,000 per millimeter.

where d is the groove spacing, λ the wavelength of the radiation, and α is defined as in Fig. 19. The value of α is the dispersion angle, which is limited to about $\pm 1^\circ$ in the X ray.

Because the grating consists of simple wires that cast shadows, the grating cannot be blazed in the same way as a reflection grating. If the wires are fully opaque and they cover about half of the area, then the plus and minus first orders of diffraction have about 10% of the beam at a maximum signal of 20%. This is a low efficiency, but is still much higher than crystals, because the entire spectrum is dispersed with that efficiency.

To disperse the short wavelength X rays through a substantial angle, the grooves must be very close together, typically 100–200 nm. This in turn means that the gratings must be very thin, substantially under 1 μ. Making a single grating of this groove density and thickness is currently impractical, so the transmission grating used in a telescope consists of many small facets, each about a square centimeter in extent. These are arranged behind the Wolter telescope. The aberrations introduced off-axis by the grating array are smaller than those from the Wolter telescope, so a better spectrum is gained by placement behind the optic in the converging beam.

Reflection

A reflection grating is an alternative to transmission gratings. The same reflection gratings that are used in the ultraviolet and longer wavelength bands can be used in X ray as well, as long as they are used at grazing incidence.

When X rays approach a grating at grazing incidence, the symmetry of the interaction is broken. The rays can approach the grating in the plane that lies perpendicular to the grooves, and the diffracted light will also emerge in that plane, as shown in Fig. 20. This is known as the "in-plane" mount. However, it is also possible for the radiation to approach the grating quasi-parallel to the grooves, in which case it diffracts into a cone about the direction of the grooves, as shown in Fig. 21. This is known as conical diffraction in the "extreme off-plane" mount because the diffracted light no longer lies in the same plane as the incident and zero-order reflected light.

The grating equation can be written in general form,

$$n\lambda = d \sin \gamma (\sin \alpha + \sin \beta), \tag{21}$$

where n is the order number, λ the wavelength, and d the groove spacing. α and β are the azimuthal angles of the incident and diffracted rays about the direction of the grooves, and γ is the angle between the direction of the grooves and the direction of the incident radiation. For an in-plane mount, γ is 90°, and the grating equation simplifies to

$$n\lambda = d(\sin \alpha + \sin \beta), \tag{22}$$

which is the usual form of the equation. In the off-plane mount, γ is small and makes the change of azimuthal angle larger at a given groove density.

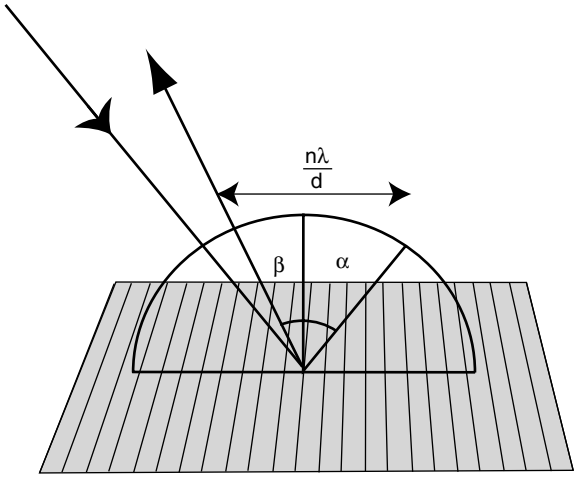


Figure 20. In the conventional grating geometry, a ray approaches in the plane that lies perpendicular to the groove direction. The reflected and diffracted rays emerge in the same plane. α measures the direction of the reflected, zero-order light, and β measures the diffracted ray. At grazing incidence, α and β must both be close to 90°.

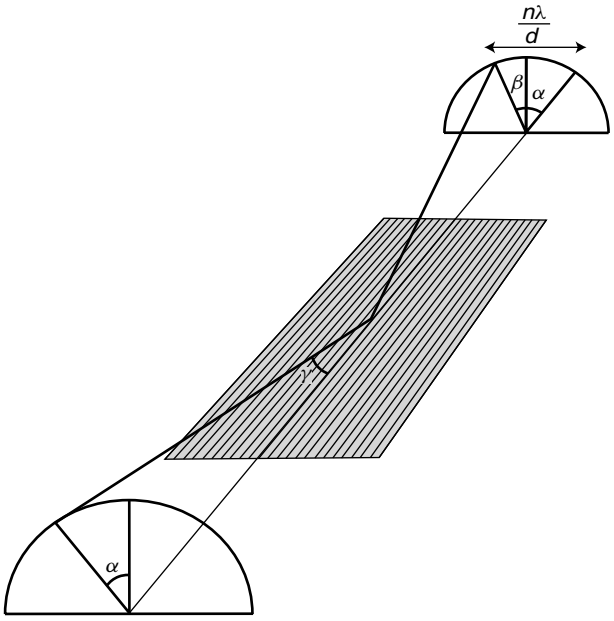


Figure 21. In the off-plane geometry, the ray approaches the grating nearly parallel to the grooves, separated by the angle γ. α and β are defined as in Fig. 20 but now are azimuthal angles. The diffracted light lies in the cone of half angle γ centered on the direction of the rulings in space.

Both approaches have been used in X-ray astronomy. The advantage of the in-plane mount is that the groove density can be low, easing the difficulty of fabrication. One advantage of the off-plane mount is that its dispersion can be greater at a given graze angle because all diffraction is at the same angle. Additionally, the efficiency of diffraction tends to be higher in the off-plane mount.

In both mounts, as with transmission gratings, it is preferable to place the gratings in the converging beam

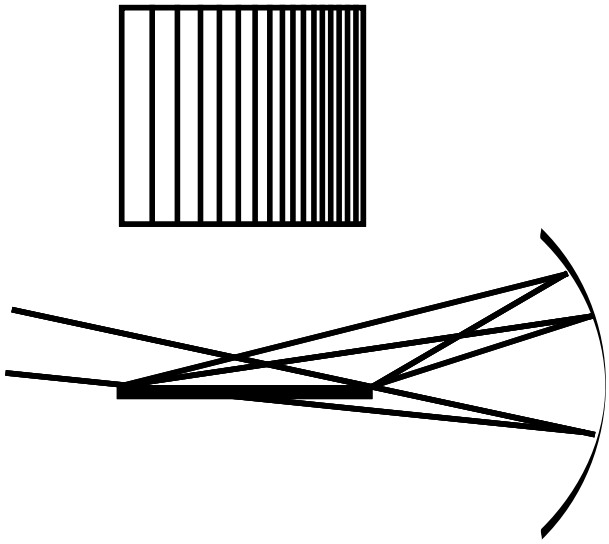


Figure 22. A varied line space grating features parallel grooves that become closer together along the length of the grating. This removes the coma from the spectrum.

behind the Wolter optic. If the gratings are placed ahead of the optic (14), then the aberrations of the telescope degrade the spectral resolution. However, a major advantage is that conventional, plane gratings can be used to disperse the light.

If the gratings are used in the in-plane mount, then the grooves may be parallel. To compensate for the changing distance between the point of reflection and the focus, the dispersion must increase as one goes farther down the grating (Fig. 22). This leads to a grating whose grooves are parallel, but change density (15). Such grating arrays were used on XMM-Newton (16).

An alternative that can lead to higher efficiency and greater resolution is to use the off-plane mount in the converging beam (17). Again, the gratings stretch from right behind the mirror to closer to the focal plane. Near the bottom of the gratings, the dispersion needs to be higher, so the grooves must be closer together. Such a geometry requires the grooves to be radial, fanning outward from a hub which is near the focus (Fig. 23). Such gratings can be more difficult to fabricate, given their requirement for higher groove density, but can give much better performance (18). Such a grating was used to obtain an extreme ultraviolet spectrum of a hot white dwarf in a suborbital rocket experiment (19).

When the gratings are placed in the converging beam, behind the optics, severe aberrations result if plane gratings are used. So, the grooves must not be parallel. Additionally, because there is a range of angles eight times the graze angle represented in the converging beam, eight or more gratings are required in an array, as shown in Fig. 24.

MAJOR MISSIONS

The history of X-ray telescopes can be tracked in the major missions that have flown and performed new astronomy

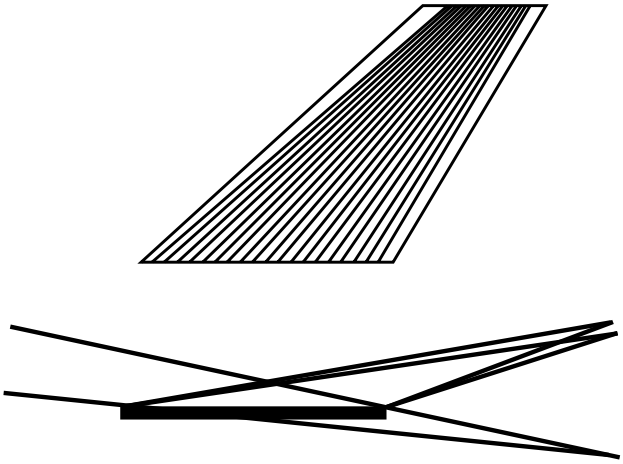


Figure 23. A radial groove grating features grooves that radiate from a hub off the edge of the grating. This removes the coma in the spectrum of the converging beam used in the off-plane mount.

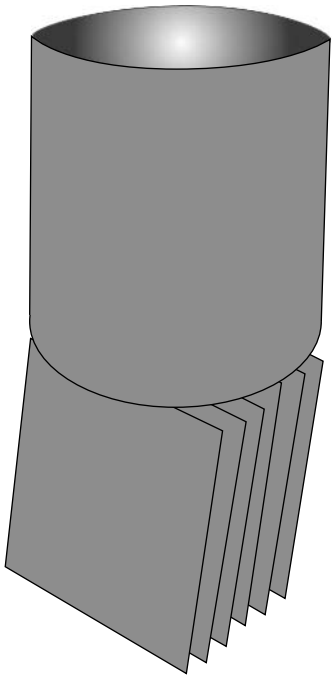


Figure 24. The radial groove and varied line space gratings must be placed in arrays behind the telescope to cover the entire beam.

using optics. There are too many missions to discuss in detail, but a few stand out as landmarks and show the evolution of X-ray telescopes.

Apollo Telescope Mount

The first high-quality X-ray images captured by telescopes were taken in an experiment called the Apollo Telescope Mount which was flown on Skylab, the orbiting space station of the 1970s. This featured a small Wolter telescope that had a few arcseconds of resolution for imaging the sun. This telescope was built before there was general access to digital, imaging detectors, and so it used film at the focal plane. Astronauts on the station operated

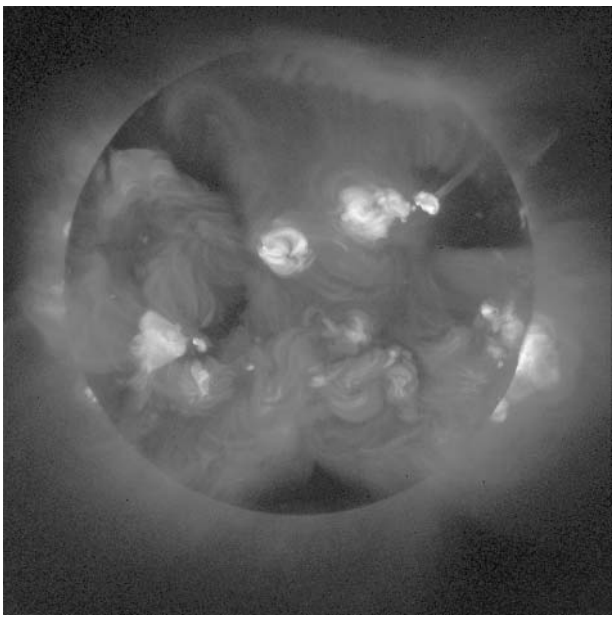


Figure 25. An image of the sun was taken at 1 keV using a paraboloid-hyperboloid telescope on the Yokoh mission. See color insert.

the telescope and returned the film to earth for analysis. Using this instrument, scientists could track the changing complexity of the solar corona in detail for the first time. In Fig. 25, we show an X-ray image of the sun captured by the Yokoh satellite in 1991 (20).

Einstein

The first major observatory for X-ray astronomy that featured a telescope was flown in 1978 (21). Dubbed the Einstein Observatory, the mission was based on a set of nested, Wolter Type I telescopes. Using a 4-m focal length and a maximum diameter of 60 cm, this observatory captured the first true images of the X-ray sky at a resolution as fine as 6 arcseconds. It had a low-resolution imaging proportional counter and a higher resolution microchannel plate at the focal plane. A transmission grating array was available for the study of bright sources. This observatory operated for just 3 years, but through the use of grazing incidence telescopes, moved X-ray astronomy onto the central stage of astrophysics.

Rosat

The next major advance came from the ROSAT (22), which featured nested Wolter telescopes of 3-arcsecond quality. Unlike the Einstein Observatory, which was used in a conventional observatory mode of pointing at selected targets, Rosat was first used to survey the entire sky. By sweeping continuously around the sky, it was able to build up an image of the entire sky in the X-ray region at unprecedented resolution. Then, it spent the ensuing years studying individual targets in a pointed mode.

Trace

Remarkable new images of the X-ray emission from the sun became available from the TRACE mission (10).

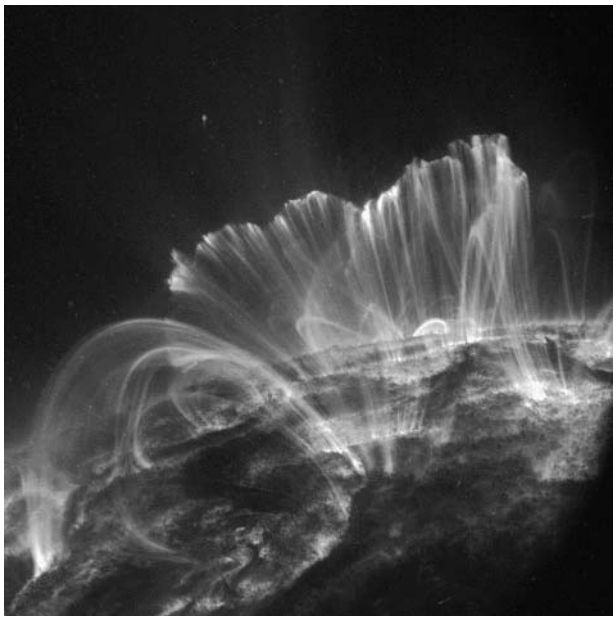


Figure 26. An image of the hot plasma on the limb of the sun was taken using a multilayer-coated telescope on the TRACE satellite. See color insert.

This satellite used a multilayer coating on a normal incidence optic to image the detailed structure of the sun's corona (Fig. 26). Because of its use of normal incidence optics, the telescope achieved exceptionally high resolution, but across a limited band of the spectrum. The structure of the X rays shows previously unsuspected detail in the sun's magnetic fields.

Chandra

NASA launched the Chandra Observatory in July 1999. This was the natural successor to the Einstein Observatory (23). Chandra features high-resolution Wolter Type I optics nested six deep. The telescope is so finely figured and polished that it achieves resolution better than one-half arcsecond, not far from the performance of the Hubble Space Telescope. Chandra, which is shown in Fig. 27, uses a CCD detector or a microchannel plate at the focal plane. It performs spectroscopy through the energy sensitivity of the CCD and by using a transmission grating array.

The improved resolution is providing dramatic new results. For example, the very first image taken was of the supernova remnant Cas-A. The telescope is so fine that it could immediately identify a previously unremarkable feature near the center as a stellar remnant of the supernova explosion. In Fig. 28, we compare the images of the Crab Nebula from Rosat, and Chandra, illustrating the importance of improved resolution from telescopes in X-ray astronomy.

XMM-Newton

In December 1999, the European Space Agency launched a major X-ray telescope into orbit. Called XMM-Newton, it is an X-ray Multi-Mirror Mission named in honor of Sir Isaac Newton (16). It features three high collecting area mirrors



Figure 27. The Chandra Satellite before launch. The 10-meter focal length of the X-ray mirror created a long, skinny geometry.

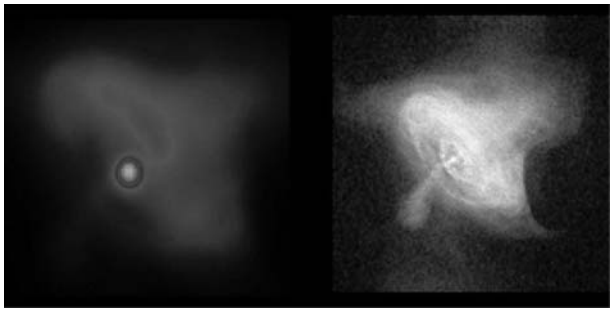


Figure 28. X-ray telescopes are improving. To the left is an image of the Crab Nebula acquired with the ROSAT in 1990. It shows the synchrotron emission in green and the bright spot which is the pulsar in the center. To the right is an image captured with Chandra in 1999. The resolution has improved from 3 arcseconds one-half arcsecond, and the level of detail is much higher. See color insert.

of the thin mirror variety. At the cost of some resolution (15 arcseconds), it can achieve a very high collecting area for studing faint objects. Spectroscopic studies using the CCDs and an array of reflection gratings are now starting to generate unique new information about the physics of X-ray sources.

Future Missions

Considerable effort has gone into the definition of future missions for X-ray astronomy. They appear to be splitting into two varieties, similarly to visible light astronomy. The first class of observatory will feature modest resolution

(1–30 arcseconds) but very high collecting area using thin mirror telescopes. NASA is planning a mission called Constellation-X (24) which will generate an order of magnitude more collecting area for spectroscopy than we currently enjoy with Chandra. The European Space Agency is studying a mission called XEUS (X-ray Evolving Universe Satellite), which will feature a huge X-ray telescope of modest resolution using thin mirrors in orbit near the Space Station Freedom (25).

The other area of development is toward high resolution. NASA is now studying a mission called MAXIM (Micro-Arcsecond X-ray Imaging Mission) which has the goal of using X-ray interferometry to improve our resolution of X-ray sources by more than a factor of 1,000,000(26). At resolution below a microarcsecond, it should be actually possible to image the event horizons of black holes in the centers of our galaxy and others.

BIBLIOGRAPHY

1. J. A. R. Samson, *Techniques of Vacuum Ultraviolet Spectroscopy*, Cruithne, Glasgow, 2000.
2. M. Born and E. Wolf, *Principles of Optics*, 7e, Cambridge University Press, Cambridge, 1999, pp. 292–263.
3. B. L. Henke, E. M. Gullikson, and J. C. Davis, *At. Data Nucl. Data Tables* **54**, 181 (1993).
4. W. Cash, *Appl. Opt.* **26**, 2,915–2,920 (1987).
5. D. K. G. de Boer, *Phys. Rev. B* **53**, 6,048–6,064 (1996).
6. H. Wolter, *Ann. Phys.* **10**, 94–114 (1952).
7. M. Lampton, W. Cash, R. F. Malina, and S. Bowyer, *Proc. Soc. Photo-Opt. Instrum. Eng.* **106**, 93–97 (1977).
8. H. Wolter, *Ann. Phys.* **10**, 286–295 (1952).
9. P. Kirkpatrick and A. V. Baez, *J. Opt. Soc. Am.* **38**, 766–774 (1948).
10. *http://vestige.lmsal.com/TRACE/*
11. D. L. Windt, *App. Phys. Lett.* **74**, 2,890–2,892 (1999).
12. W. Cash, A. Shipley, S. Osterman, and M. Joy, *Nature* **407**, 160–162 (2000).
13. U. Bonse and M. Hart, *App. Phys Lett.* **6**, 155–156 (1965).
14. R. Catura, R. Stern, W. Cash, D. Windt, J. L. Culhane, J. Lappington, and K. Barnsdale, *Proc. Soc. Photo-Opt. Instrum. Eng.* **830**, 204–216 (1988).
15. M. C. Hettrick, *Appl. Opt.* **23**, 3,221–3,235 (1984).
16. *http://sci.esa.int/home/xmm-newton/index.cfm*
17. W. Cash, *Appl. Opt.* **22**, 3,971 (1983).
18. W. Cash, *Appl. Opt.* **30**, 1,749–1,759 (1991).
19. E. Wilkinson, J. C. Green, and W. Cash, *Astrophys. J.* (Suppl.), **89**, 211–220 (1993).
20. *http://www.lmsal.com/SXT/*
21. *http://heasarc.gsfc.nasa.gov/docs/einstein/heao2_about.html*
22. *http://heasarc.gsfc.nasa.gov/docs/rosat/rosgof.html*
23. *http://chandra.harvard.edu*
24. *http://constellation.gsfc.nasa.gov*
25. *http://sci.esa.int/home/xeus/index.cfm*
26. *http://maxim.gsfc.nasa.gov*

This page intentionally left blank

INDEX

A
A or B wind, 1040
A roll, 1034
A scan, in ultrasonography, 1413
A&B cutting, 1040
Abbe (V) number, 234–235, 1081–1082, 1100–1102, 1109, 1122–1123
Abbe sine condition of light optics, in charged particle optics and, 88
Abbe, Ernst, 261, 1100, 1106, 1109
Aberrations, 91–92, 234, 559, 1095, 1098, 1114–1124
  accommodation and, 547
  in charged particle optics, 92–94, 98
  chromatic aberration, 1081–1082
  human eye, 547–548
  human vision and, 554–555
  microscopy and, 1106
  monochromatic, 1083–1085
  optical transfer function (OTF) and, 1095–1098
  PSF and, 1088–1089
  wave theory of, 1083–1084
Absolute fluorescence (ABF), 863
Absorptance, 527–529
Absorption, 253–256, 527–529, 803
  astronomy science and, 684–686
  atmospheric, 1497
  interstellar, 1496–1497
  microscopy and, 1126
  near resonance, 228–233
  photodetectors and, 1184–1187
  silver halide, 1273
Absorption band, 229
Absorption edge, 255
Absorption lines, 235
Academy aperture, 1040, 1359
Academy leader, 1040
Accommodation, 547, 1328
Acetate film, 1039, 1040
Acetone, PLIF and, 409
Achromatic doublet, 235
Acoustic antireflective coatings (AARC) in scanning acoustic microscopy (SAM), 1231
Acoustic impedance, in ultrasonography, 1415
Acoustic microscopy, 1128–1148
Acoustic reciprocity theorem, 1
Acoustic sources or receiver arrays, 1–9
Active centers, 1137
Active glasses, in three-dimensional imaging, 1331
Active matrix liquid crystal displays, 374, 857–858, 956
Active MATrix coating (AMAT), in electrophotography, 301
Active pixel sensors (APS), 1199–1200
Active snakes, thresholding and segmentation in, 644–645
Acuity, of human vision, 558–560
Acutance, 1357
Addition, image processing and, 590
Addition, Minkowski, 433, 612
Additive color films, instant photography and, 847–849
Additive color matching, 102
Additive color mixing, 126–127
Additive printing, 1040
Additivity, Grassmann's, 531
Addressability, cathode ray tube (CRT), vs. resolution, 32
Addressing displays, liquid crystal, 955–959
Admissibility condition, wavelet transforms in, 1446
Advance, 1040
Advanced Photo System (APS) in, 141
Advanced Technology Materials Inc. (ATMI), 377
Advanced Very High Resolution Radiometer (AVHRR), 759, 760–761
Advanced Visible Infrared Imaging Spectrometer (AVIRIS)
  geologic imaging and, 648, 649
  in overhead surveillance systems, 787
Aerial imaging, 350, 463–476, 1040
Aerial perspective, 1328
Aerosol scattering, lidar and, 880–882
Afocal systems, 1074, 1079
Agfa, 1024
Agfa Geaert, 839
Agfachrome Speed, 839
Agfacolor Neue process in, 128
Aging and human vision, 541, 549, 560
AgX (*See* Silver halide)
Airborne and Topographic SAR (AIRSAR/TOPSAR), 650
Airborne radar, 1471–1473
Aircraft in overhead surveillance, 773–802
Airy disk, 249
  flow imaging and, 399
  microscopy and, 1110, 1124
  telescopes, 688
Airy function, 94, 399
Airy pattern, 1082
Albedo, 610, 611
AlGaN/GaN, scanning capacitance microscope (SCM) analysis, 22
Algebraic opening/closing, 437
Algebraic reconstruction technique (ART), in tomography, 1407–1408
Algebraic theory, in morphological image processing, 430
Aliasing, 50, 59, 60, 69, 75, 84
  flow imaging and, 401
  in magnetic resonance imaging (MRI), 986
Along Track Scanning Radiometer (ATSR), 772
Alternating current scanning tunneling microscopy (ACSTM), 28
Alternating pairs, in three-dimensional imaging, 1338
Alternating sequential filters, 437
Alumina dopants, secondary ion mass spectroscopy (SIMS) in analysis, 487–489
Aluminum lithium alloys, secondary ion mass spectroscopy (SIMS) in analysis, 484
Ambient light, vs. cathode ray tube (CRT), 182–183
American National Standards Institute (ANSI), 1041
American Standards Association (ASA), 1023
Amines, 1179
Amino acids (*See* Biochemistry)
Ampere's law, 211, 218
Amplifiers, SQUID sensors, dc array, 12
Amplifying medium, radar, 223
Amplitude, 212, 226, 1098
  beam tilting, 5
  nonsymmetrical, 4
Amplitude modulation (AM), 383, 1362
Amplitude reflection coefficient, 235
Amplitude resolution, 151
Amplitude time slice imaging, ground penetrating radar and, 472–475, 472
Amplitude transmission coefficient, 235
Amplitude weighting, 3
Anaglyph method, in three-dimensional imaging, 1331

Analog technology, 1040
in endoscopy, 334
SQUID sensors using, 10–12
Analog to digital conversion (ADC), 49, 61
in field emission displays (FED), 382
in forensic and criminology research, 721
in overhead surveillance systems, 786
Analytical density, 1040
Anamorphic image, 1040
Anamorphic lens, 1031, 1040
Anamorphic release print, 1040
AND, 590
Anger cameras, neutron/in neutron imaging, 1062–1063
Angle, 1040
Angle alpha, 541
Angle of incidence, 234, 468
Angle of reflection, 234
Angle of refraction, 234
Angular dependency, in liquid crystal displays (LCDs), 184
Angular field size, 1080
Angular frequency, 213
Angular magnification, 1076–1077
Angular resolution, High Energy Neutral Atom Imager (HENA), 1010
Angular spectrum, 1099
Animation, 1022, 1040–1042
cel, 1042
meteor/in meteorological research, 763–764, 769
in motion pictures, 1035
Anisotropic coma, 94
Anisotropic distortion, 94
Anisotropic astigmatism, 94
Anode, in field emission displays (FED), 380, 386, 387
Anomalies, gravitational, 444–445
Anomalous dispersion, 228
Anomalous propagation, radar, 1454
Anomalous scattering, in biochemical research, 698
Ansco Color process in, 128
Answer print, 1041
Antennas, 220, 242
ground penetrating radar and, 464–465, 468–469
in magnetic resonance imaging (MRI), 979
radar and over the horizon (OTH) radar, 1142, 1151, 1450
terahertz electric field imaging and, 1394
Antifoggants, in photographic color display technology, 1216
Antihalation backing, 1041
Antireflective coatings (ARCs), 381
Aperture, 54, 57–59, 243, 247–249, 1041, 1101, 1352
academy, 1040, 1359
Fraunhofer diffraction in, 247–249
microscopy and, 1108, 1122, 1124, 1128
in motion pictures, 1028
numerical, 1081, 1115
in overhead surveillance systems, 783
radar and over the horizon (OTH) radar, 1147
relative, 1081
Aperture plate, 1041
Aperture stop, 1080
Aperture transmission function, 248
Aplanatic condensers, 1123
Apodization, 1086, 1088
Apollo Telescope Mount, 1507
Appliances, orthopedic, force imaging and, 422
Applications Technology Satellite (ATS), 757
Aqueous humor, 512
Arc lamps, 1041
Archaeology, ground penetrating radar and, 464
Archiving systems, 661–682
art conservation and analysis using, 661–682
in motion pictures, 1038–1039
Argon plasma coagulation (APC), in endoscopy, 340
Armat, Thomas, 1022
Array amplifiers, SQUID sensors, 12
Array of Low Energy X Ray Imaging Sensors (ALEXIS), lightning locators, 905, 929
Array theorem, 249
Arrival Time Difference (ATD), lightning locators, 890–904
Art conservation, 661–682
Artifacts
quality metrics and, 598–616
tomography/in tomography, 1410
Artificial intelligence, 371
feature recognition and object classification in, 351
in search and retrieval systems, 632
Artificial vision systems, 352
Arylazo, 135
ASA/ISO rating, film, 1023, 1041
ASOS Lightning Sensor (ALS), 907, 922
Aspect ratio, 1031, 1041
in motion pictures, 1022
television, 147, 1359
Associated Legendre functions, 253
ASTER, 660
Astigmatism, 92, 94, 1084–1085, 1089
in charged particle optics, 93
electron gun, 40–41
microscopy and, 1119
Astronomy (*See also* Telescopes), 682–693
Apollo Telescope Mount, 1507
Chandra Observatory, 1508
Constellation X mission, 1509
Einstein Observatory Telescope, 1507
magnetospheric imaging, 1002–1021
ROSAT telescopes, 1507
TRACE telescopes, 1507–1508
X-ray Evolving Universe Satellite, 1509
X-ray telescope, 1495–1509
XMM Newton telescope, 1508
Asynchronous transfer mode (ATM), 1382
Atacama Large Millimeter Array (ALMA), 693
Atmospheric pressure chemical vapor deposition (APCVD), 384
Atmospherics, 890
Atomic force microscope (AFM), 16
Atomic transitions, 215
ATSC Digital Television Standard, 1359
ATSC Digital Television Standard, 1382–1389, 1382
Attenuation, 229
gamma ray, 260
ground penetrating radar and, 466
human vision and, 559, 562
radar and over the horizon (OTH) radar, 1147
in ultrasonography, 1416–1417
Audio (*See* Sound)
Auroras, far ultraviolet imaging of, 1016–1020
Authentication
digital watermarking and, 161
in forensic and criminology research, 739–740
Autoassemble, 1041
Autochrome Plate process, color photography, 127
Autocorrelation, 1105
Autoexposure, 1355–1356
Autofluorescence, 1136
Autofocus, 1356
Automated Surface Observing System (ASOS) lightning locators, 907, 922

Autonomous System Lab, art conservation and analysis using, 664
Autostereoscopic displays, 1328, 1336–1341
Avalanche photoconductors, 1173–1174
Aviation meteorology, 767
Axis, camera, 1041
Azimuth range Doppler (ARD), radar and over the horizon (OTH) radar, 1148
Azohydroquinones, instant photography and, 834
Azomethines, instant photography and, 835
Azos, 1179
Azosulfones, instant photography and, 839

B

B frame, video, 1387
B roll, 1034
B scan, in ultrasonography, 1413, 1417, 1428–1429
Babinet's principle, 249
Background, 568–569
Background, in infrared imaging, 807, 810, 813
Background haze, 604
Background limited infrared photodetector (BLIP), 1189
Background uniformity, 605
Backing, 1041
Backlight systems, in three-dimensional imaging, 1339
Backpropagation algorithm, in neural networks, 372
Backscatter, 873, 1450, 1469
Backscattered electron (BE) imaging, in scanning electron microscopes (SEM), 276
Baffled circular piston, 8
Baffled rectangular piston, 8–9
Balance stripe, 1041
Balance, color, 114–116, 118
Ballistics analysis, in forensic and criminology research, 716
Balloons in overhead surveillance, 773
Band theory of matter, photoconductors and, 1170–1171
Bandpass filters, 593, 611
Bandwidth, 50
  cathode ray tube (CRT), 179–180
  digital watermarking and, 149–150
  expansion of, 149–150
  lightning locators, 905
  in magnetic resonance imaging (MRI), 985
  radar and over the horizon (OTH) radar, 1147
  television, 1362, 1367
Bandwidth compression, in overhead surveillance systems, 786
Barn door, 1041
Barney, 1041
Barrel distortion, 93
Baryta paper, 1209–1210
Base, for film, 1022–23, 1041, 1045
BASIC, 186
Batchelor scale, flow imaging and, 404
Beam candlepower seconds (BCPS) rating, 492
Beam conditioning, X-ray fluorescence imaging and, 1477
Beam expanders, holography, 509
Beam patterns and profiles, 1, 1426–1427
Beam splitters, in holography, 509
Beam tilting, amplitude, 5
Bechtold, M., 456
Beer-Lambert absorption law, 409
Beer's law, 344
Bell, Thomas, 455, 456
Benchmark comparison, 607
Bent contours, 282
Benzisoxazolone, in instant photography, 838
Benzoylacetanilides, 134
Bertrand polarization lens, 1132
Bessel function, 94, 1082
Best optical axis, in human vision, 540–541
Beta decay, 220
Bethe–Bloch equation, 1156
Biacetyl, PLIF and, 409
Bias voltage, scanning capacitance microscope (SCM) vs., 20
Biased transfer roll (BTR), in electrophotography, 301
Bidirectional reflectance distribution function (BRDF), 51, 528
Binary imaging, 584–589
Binder or dead layer, in field emission displays (FED), 381
Binocular disparity, 1328
Binoculars, 1074
Biochemistry and biological research, 693–709
  force imaging and, 422
  in scanning acoustic microscopy (SAM), 1228
  in scanning electrochemical microscopy (SECM), 1255–1256
  secondary ion mass spectroscopy (SIMS) in analysis, 484
  X-ray fluorescence imaging and, 1479–1482
Biochips, terahertz electric field imaging and, 1403
Bioorthogonal wavelet basis, 1446
Bipack filming, 1041
Bipedal locomotion, force imaging and analysis of, 419–430
Bipolar junction transistors, 1173
Birefraction, 1131–1132
Birefringence, 233, 1134
Bisazos, 1179–1180
Bistatic radar, 772
Bit rates, compressed vs. uncompressed, 152
Black-and-white film, 1023, 1041, 1356
Black-and-white images, 584–589, 830–833, 1347
Black-and-white TV, 1359
Black light, 1041
Black-white vision, 567
Black, James Wallace, 773
BLACKBEARD lightning locators, 905, 929
Blackbody radiation, 103, 211, 222, 525, 690, 803, 804, 813
  in overhead surveillance systems, 782, 789
  photodetectors and, 1184–1187
Blanking, television, 1360
Bleaches, 138–139, 1217
Blimp, 1041
Blind spot, human eye, 514, 516
Blobs, 646
Blobworld, 624, 630
Block based search and retrieval systems, 623
Blocking lens, in three-dimensional imaging, 1331
Blooping, 1041
Blow up, 1041
Blue light, 101
Blue screen, 1041
Blueshift (*See also* Doppler shift), 686, 772
Blur function, 596
Blurring, 50, 56, 58, 59, 72, 81, 82, 83, 84, 544, 577, 604
  digital watermarking and, 167
  flow imaging and, 399–400
  high-speed photography and, 491–492
  image processing and, 579, 581, 595–596
  in medical imaging, 756
  in overhead surveillance systems, 784
  quality metrics and, 598–616

Bohr magnetron, 288
Bolometers, 1193–1194, 1204–1206
Boltzmann's constant, 222
Bonvillan, L.P., 773
Boolean operations, 590
Boom, 1041
Borehole gravity meter, 450
Bormann effect, 282
Born Oppenheimer approximation, 216
Bouguer anomaly/correction, 447, 453
Bounce light, 1041
Boundaries, thresholding and segmentation in, 644–645
Boundary conditions, 233–234
Boundary hugging, thresholding and segmentation in, 642
Bragg reflection, 244
Bragg's law, 267
Brain and human vision, 513–514, 558, 569
Braking radiation (bremsstrahlung), 223
breakdown, 1041
Breast cancer detection, in infrared imaging, 812
Bremsstrahlung, 219, 223–224
Brewster angle, 237
Brewster window, 237
Bright field image, 268
Bright fringes, 243, 269–270
Brightness, 102, 618, 1037
  cathode ray tube (CRT), 34–35, 180
  in charged particle optics, electron gun, 88
  feature measurement and, 344–345
  in field emission displays (FED), 382
  in forensic and criminology research, 723–724
  image processing and, 580–584
Broad light, 1041
Broadatz textures, 622
Broadcast transmission standards, television, 1359–1393
Browsing, in search and retrieval systems, 617
Bucky diagrams, 572
Buffer rods, in scanning acoustic microscopy (SAM), 1233, 1234
Building block structures, in transmission electron microscopes (TEM), 271
Buried heterostructure laser, scanning capacitance microscope (SCM) analysis, 22–23
Butt splice, 1041

C

C scan ultrasonography, 1413
Cadmium sulfide photodetectors and, 1190, 1200
Calibration
  color image, 116–117
  DPACK, 27
  FASTC2D, 27
  force imaging and, 424
  Hall generators, 974
  scanners, 603
  scanning capacitance microscope (SCM), 27–28
  TSUPREM4, 28
Calotype, 1259–1309, 1345
Cambridge Display Technology (CDT), 818
CAMECA, 478, 479
Camera axis, 1041
Camera film, in motion pictures, 1026–1027
Camera log, 1041
Camera obscura, 1344–1345
Camera operator/animator, 1042
Cameras
  Anger, 1062–1063
  animation, 1041
  aperture in, 1028
  autoexposure, 1355–1356
  autofocus, 1356
  Captiva, 847
  digital imaging, 854–855
  dollies, 1030
  electronic flash in, 1348–1349
  energetic neutral atom (ENA) imaging, 1006–1010
  flow imaging and, 393–394
  in forensic and criminology research, 710–714
  frame and film gate in, 1028
  frame rates in, 495
  handheld cameras, 1031
  high-speed, 494–498, 1047
  I Zone, 847
  image dissection cameras, 498
  image formation in, 571
  instant photography and, 827–859
  intermittent action high-speed cameras for, 495
  large format, 1352–1354
  lens in, 1029, 1348
  medium format, 1351
  microimaging, 1351
  mirrors in, 1074, 1346, 1349–1350
  for motion pictures, 1022, 1027–1029, 1042
  motor drive in, 1348
  in overhead surveillance systems, 773–802
  photoconductors and, 1174
  pinhole, 1072–1073
  Pocket Camera, 847
  Polachrome, 848
  Polacolor, 843–844
  Polaroid, 844–847
  Polavision, 848
  pull down claw in, 1027
  rangefinder in, 1350–51
  reflex type, 1345, 1349–1350
  rotating drum and mirror cameras, 496–497
  rotating mirror framing cameras, 497–498
  rotating prism cameras, 495–496
  scintillation, 1313–1314
  shutter in, 1027–28, 1351–1352
  single lens reflex (SLR), 1349–1350
  speed of, 1028–1029
  Steadicam, 1031
  still photography, 1344–1358
  streak cameras, 499–500
  strip cameras, 500
  tripods, 1030
  trucks, 1030
  video assist, 1031
  video, 1029–1031, 1174
  viewfinders in, 1029
Campisi, George J., 375
Canadian Lightning Detection Network (CLDN), 890–904, 935
Cancer detection, using infrared imaging, 812
Candela, 1042
Candescent, 377–378
Canon Image Runner, 300
Capacitance electronic disk (CED), 16
Capacitance sensors, 17–18, 423–424
Capacitance–voltage (C–V) curve, scanning capacitance microscope (SCM), 21, 25
Capacitive probe microscopy (*See also* Scanning capacitance microscope), 16–31
Capacity, information, 1082–1083
Captiva, 847
Captured orders, microscopy and, 1109
Carbon arc lamps, in projectors, 1037
Carlson, Chester F., 299, 1174
Cascade development, in electrophotography, 312
Cassegrain mirrors, 783
Cassini Saturn Orbiter, 1020
Cataract, human eye, 548
Category scaling, quality metrics and, 608–609

Cathode
cathode ray tube (CRT), 44–45
electron gun, 39
in field emission displays (FED), 378–380, 384–387
Cathode ray direction finder (CRDF) lightning locators, 890, 912, 935
Cathode ray tube (CRT), 31–43, 44–48
in three-dimensional imaging, 1330
electron microscope use of, 262
in field emission displays (FED) vs., 374
graphics cards and, 174–175
meteor/in meteorological research, 757
oscilloscope using, 47
radar tubes using, 47
in three-dimensional imaging, 1333
Cathodluminescence (CL), 374
CAVE three-dimensional display, 1335
Cavity resonators, 1223–27
CD ROM compression, 151, 156–157
CD-2/3/4 developers, 131
Cel animation, 1042
Cellulose triacetate, 1042
Cement splice, 1042
Ceramics, secondary ion mass spectroscopy (SIMS) analysis, 484
Cerenkov radiation/counters, 1158, 1162
CERN Accelerator, 970
CGR3 flash counters, 907
Chain codes, in search and retrieval systems, 625
Chambers of human eye, 746
Chandra Observatory, 1508
Changeover, 1042
Channel buffer, 1388
Channel capacity, 65–66
Channel constancy, 173–174, 181, 184
Channel independence, 176, 181
Channeling, electron, 285–286
Channels, 1362, 1388
Character animation, 1042
Characterization of image systems, 48–86
Charge carrier, photoconductors and, 1170
Charge coupled devices (CCD), 1090
art conservation and analysis using, 663, 664, 671
astronomy science and, 688
color photography and, 141
digital photomicrography, 1138–1140
in endoscopy, 334
flow imaging and, 393–394
in forensic and criminology research, 722–723
high-speed photography and, 498–499
human vision and vs., 552–553
image formation in, 571
instant photography and, 827
multipinned phase (MPP), 1198
noise in, 396–397
in overhead surveillance systems, 784–787
particle detector imaging and, 1167
photoconductors, 1173–1174
photodetectors and, 1184, 1187, 1190, 1193, 1194–1198
radiography/in radiographic imaging, 1067–1069
terahertz electric field imaging and, 1398
virtual phase CCD (VPCCD), 1198
X-ray fluorescence imaging and, 1477
Charge cycle, in electrophotography, 300–304
Charge exchange process, 1004
Charge injection devices (CID), 1192–1193
Charge mode detectors, 1191–1193
Charge-to-mass ratio, in electrophotography, 313–315
Charge transfer devices (CTD), 1192
Charge transport layer (CTL), 1176
Charged area development (CAD), in electrophotography, 317
Charged particle optics, 86–100
Charging cycle, 1176
Charters of Freedom project, art conservation, 663
Chemical composition analysis, 684–686
Chemical sensitization, 1288–1293
Chemical shift imaging, MRI, 995–1000
Chemical shift saturation, MRI, 994–995
Chemical vapor deposition (CVD), 383–384
Chemiluminescence, 255
Chen operator, image processing and, 582
Chirped probe pulse, terahertz electric field imaging and, 1396–1397
Cholesteric liquid crystal displays, 965–966
Chomratic aberration, 92
Choppers, lidar, 873
Chroma in, 103
ChromaDepth three-dimensional imaging, 1332
Chromakey, 1042
Chromatic aberration, 94, 234, 1081–1082
in charged particle optics, 93, 98
in human vision, 544–545, 554–555
in microscopy, 1117–1118, 1123
Chromatic adaptation, human color vision and, 520
Chromatic difference of refraction, 544
Chromatic filters, human vision and, 560
Chromaticity, 119, 148, 380, 512, 534–535
Chromaticity diagrams, 107–108, 534–535
Chrome, in gravure printing, 457–458
Chrominance, 103, 109, 1042
Chromogenic chemistry, 129–139
Chromophores, instant photography and, 836
Cibachrome process, 127
CIECAM97, 538
CIELAB, 521, 533, 535, 536, 537, 619
CIELUV, 535, 537, 619
CIGRE flash counter lightning locators, 907
Cinch marks, 1042
Cinemascope, 1031, 1042
Cinematographe, 1022
Cinemiracle, 1042
Cineon digital film system, 1042
Cineorama, 1031
Cinerama, 1031, 1042
Cinex strip, 1042
Circ function, 542
Circaram, 1042
Circles of confusion, 1347
Circular pistons, 8, 1424–1426
Circular polarization, 231–232
Circularity, 347–348, 627
Clapsticks, 1042
Clarity, 71, 75
Classical electron radius, 250
Clausius–Mossotti relationship, 229
Claw, 1042
Clay animation, 1042
Climatic change, geologic imaging and, 655–656
Close up, 1042
Closed captioning, 1389
Closed interval, 434, 435
Closing, morphological, 430, 436–438, 585
Cloud classification, 761–764

Cloud to Ground Lightning Surveillance System (CGLSS), 890–904, 940–941
Clustering, in feature recognition and object classification, 366–370
Clutter, radar vs., 470–472
CMOS image sensors, 1139–1140, 1199–1200
Coated lens, 1042
Coaterless film, 832
Coatings, 528
Coded OFDM (COFDM), 1389
Codewords, 1388
Coding, retinex, 75–83
Codoping, secondary ion mass spectroscopy (SIMS) analysis of, 489
Coefficient of thermal expansion (CTE), 382, 384, 387
Coherent image formation, 242, 258, 504, 507, 874, 1086, 1098–1100
Coherent PSF, 1087–1088
Coherent transfer functions, 1098
Coils, deflection yoke, 41–42
Cold field emitter (CFE), in charged particle optics, 89
Cold pressure transfer, in electrophotography, 323
Collimation, 247, 255, 509, 1042
Collinear sources, in-phase signals, 2–3
Colonoscopes, 332
Color analyzer, 1042
Color appearance models, 537–538
Color balance, 114–116, 118, 1042
Color blindness, 522–523
Color burst reference, television, 1367
Color characterization, in field emission displays (FED), 380
Color circle in, 109–110
Color codes, 705–706, 1116
Color compensation in, 116–117
Color constancy, in human vision, 520–521
Color coordinate systems, 105–114, 531–536, 619, 537, 630, 641
Color correction, 114–116, 1042
Color difference signals, 106–107, 536–537, 1367
Color duplicate negative, 1035, 1043
Color electrophotography, 325–329
Color film, 1023, 1356–1357
Color image calibration in, 116–117
Color image fidelity assessor (CIFA) model, 612
Color image processing, 100–122, 325–329
Color imaging, 51, 56, 75, 219, 524, 641
  in biochemical research, 705–706
  cathode ray tube (CRT), 35, 46, 47
  in endoscopy, 335–336
  in forensic and criminology research, 714
  geologic imaging and, 654
  in infrared imaging, 809
  instant photography and, 833–842
  liquid crystal displays, 967–968
  quality metrics and, 598–616
  in search and retrieval systems, 618–622
  in ultrasonography, 1431–1432
Color internegative, 1043
Color lookup tables (CLUTs), 575
Color master, 1034
Color matching, 102, 104–105, 531–534
Color negative, 1043
Color photography, 122–145, 1208–22
Color positive, 1043
Color print film, 1043
Color purity, 388
Color reproduction, in dye transfer printing, 194
Color reversal film, 1043
Color reversal intermediate film, 1043
Color saturation, 1043
Color separation negative, 1043
Color shifted dye developers, 835
Color space, 512, 521, 535–536, 619, 641
Color television, 1365–67
Color temperature, 103, 525
Color ultrasound, 1431–1432
Color vision, 512–514, 518–521, 551–552, 560–561, 564, 567, 747, 1328
  deficiencies of, 522–523
Coloray, 377
Colored masking couplers for, 136–137
Colorimetry, 103–105, 512, 523–524
Coma, 92–94, 1084, 1089, 1119
Comb filters, television, 1366
Combined feature, in search and retrieval systems, 630–631
Combined images, image processing and, 589–591
Combined negative, 1043
Commission Internationale de l'Eclariage (CIE), 102, 618
Common intermediate format (CIF), 156–157
Communication theory, 49, 75–83, 161, 168–171
Compensating eyepieces, microscopy and, 1115, 1120
Compensation, color, 116–117
Complementary color, 1043
Complex cells, 567–568
Component analog video (CAV), 1373–74
Component analysis for dimensionality reduction, 363–364
Component digital television 1374–75
Component TV systems, 149–150
Component video standards, 1380–1382
Composite digital television, 148–149, 1377–1380
Composite materials, in scanning acoustic microscopy (SAM) analysis for, 1228
Composite print, 1043
Composite video, 1043
Compound, 1043
Compressed video, 1385–1386
Compression, 151, 156–157, 519
  advantages of, 151
  audio, 1388
  bit rate, vs. uncompressed, 152
  CD ROM, 151, 156–157
  common intermediate format (CIF) in, 156–157
  comparison of techniques in, 155
  complexity/cost of, 156
  digital compression standards (DCS) in, 156
  digital watermarking and, 150–157
  discrete cosine transform (DCT) in, 154
  displaced frame difference (DFD) in, 155
  entropy coding in, 154–155
  error handling in, 156
  in forensic and criminology research, 740, 741
  high definition TV (HDTV), 151, 153, 157
  human vision and color vision, 521
  interactivity and, 152
  MPEG and, 156–157
  multiple encoding and, 152–153
  noise in, 152
  in overhead surveillance systems, 786
  packet communications and, 151, 152
  packetization delay and, 152
  perceptual factors and, 155
  perceptual redundancy and, 151
  predictive coding (DPCM) in, 153, 155

quality of, 151
quantization in, 154
redundancy in, 155
requirements of, 151–153
resolution vs., 151
robustness of, 152
sampling in, 153
scalability of, 153
in search and retrieval systems, 631
standard definition TV (SDTV), 157
standards for, 156–157
statistical redundancy and, 150
subband/wavelet coding in, 154
symmetry of, 152
temporal processing in, 155
transform coding in, 153–154
transformation in, 153
vector quantization in, 154
video, 1385–86
Videophone standard for, 156–157
wavelet transforms in, 154, 1449–1450
Compton scattering, 211, 250, 256–258, 260, 1323–1324
Compton, Arthur, 211
Compur shutters, 1351
Computed tomography (CT), 219, 1057–1071, 1404
image formation in, 571–572
in magnetic resonance imaging (MRI) vs., 983
in medical imaging, 743
single photon emission computed tomography (SPECT), 1310–1327
Computer modeling, gravity imaging and, 452–453
Computer simulations of human vision, 50
Computers
in biochemical research, 695–696
holography in, 510–511
in magnetic resonance imaging (MRI), 999–1000
meteor/in meteorological research, 769–771
in search and retrieval systems, 616–637
X-ray fluorescence imaging and, 1477–1478
Concentration density, flow imaging and, 408–411
Concentration measurements, planar laser induced fluorescence (PLIF) in, 863–864
Condensers, microscope, 1112–1124
Conduction electrons, 230
Conductivity, 220–231
developer, in electrophotography, 315–317
photoconductors and, 1171–1172
Cone coordinates, 176–178
Cone opsins, 517, 561
Cones, of human eye, 122, 513–517, 551–554, 558, 560–563, 746–747
Conform, 1043
Conjugate images, holography in, 505–506
Conjunctive granulometry, 441
Constant deltaC or closed loop mode, SCM, 21
Constant deltaV or open loop mode, SCM, 20
Constant luminance, television, 1367
Constant separation imaging, SECM, 1254–1255
Constellation X Observatory, 693, 1509
Constructive interference, 241, 243, 245, 507
Contact heat transfer, in electrophotography, 324–325
Contact print, 1043
Content based indexing, 617–618
Continuity, 1032–1033, 1043
Continuous contact printer, 1043
Continuous line sources, 6–7
Continuous planar sources, 7–8
Continuous wave (CW), 288, 293, 391
Continuous wavelet transform (CWT), 1444–1446
Contours, 615, 644–645
Contrast, 81, 117, 270, 605, 606, 611, 612, 623, 1094
in cathode ray tube (CRT), 34–35, 183
in field emission displays (FED), 387–388
in forensic and criminology research, 723–724
human vision and, 521, 558–560
liquid crystal displays, 967
in magnetic resonance imaging (MRI), 983, 988–991
in medical imaging, 752–755
microscopy and, 1110, 1113–1124, 1127, 1132–1134
in motion pictures, 1043
Rose model and medical imaging, 753–754
in scanning acoustic microscopy (SAM), 1229, 1235–1244
silver halide, 1261–1262
in transmission electron microscopes (TEM), 269
in ultrasonography, 1418, 1423–1424, 1432–33
Contrast agents, in ultrasonography, 1418
Contrast mechanism, scanning capacitance microscope (SCM), 17–18
Contrast resolution, in ultrasonography, 1423–1424
Contrast sensitivity, 167, 521–522, 558–560, 605
Contrast transfer function (CTF), 401
Control strip, 1043
Conventional fixed beam transmission electron microscopy (CTEM), 277
Convergence, 34, 1328, 1330
Convergent beam diffraction, 268
Convex, 438
Convexity, 438
Convolution, 594–597, 1085–1092, 1105
Cooling, of photodetectors, 1188–1190
Coordinate registration, radar, 1144, 1147–1148
Copal shutters, 1351
Copiers, 574
electrophotographic/xerographic process in, 299
photoconductors, 1174–1183
quality metrics and, 598–616
Copper, in gravure printing, 457–458, 463
Copper phthalocyanines, 301
Copy control, digital watermarking and, 159
Copycolor CCN, 839, 855
Core, film, 1043
Corey Pauling Koltun (CPK) models, 694
Cornea, 512, 539, 541, 547
Corona, in electrophotography, 302–303
CORONA project, 777, 780
Corotron charging, in electrophotography, 302–303
Corpuscles, in light theory, 211
Correction filter, 1043
Corrections, optical, microscopy and, 1115
Correlated color temperature (CCT), 525
Correlation, image processing and, 594–597
Correspondence principle, 215
Corrosion science, scanning electrochemical microscopy (SECM) in, 1255
Coulombic aging, 380

Couplers, coupling, 1043
in color photography, 129, 131–137
in photographic color display technology, 1216
scanning capacitance microscope (SCM) vs., 20
in scanning acoustic microscopy (SAM), 1230
Coupling displacement, in instant photography dyes, 839
Cover glass, microscope, 1115–1116
Coverslip, microscope, 1115–1116, 1118
Crane, 1043
Credits, 1043
Criminology (*See* Forensics and criminology)
Critical angle, 238
Cronar films, 1023
Cropping, 1043
Cross talk, in three-dimensional imaging, 1330, 1333
Cross viewing, in three-dimensional imaging, 1336
Cryptography, digital watermarking and, 161–164
Crystal growth and structure of silver halide, 125–126, 1262–1268
Crystal spectrometers, 244
Crystal structure, Bragg reflection and, 244
Crystallography, 282–285, 696–699
Cube root systems in, 107–108
Cubic splines, tomography/in tomography, 1405
Current density, in charged particle optics, 87, 93
Curvature of field, 92, 93, 1085, 1118–1119
Cut, cutting, 1043
Cyan, 1043
Cyan couplers, color photography, 133–135
Cycles of wave, 213
Cyclic color copying/printing, 328
Cycolor, 827
Cylinders, in gravure printing, 457–463
Cylindrical color spaces in, 109–110
Cylindrical coordinates in, 110–111
Czochralksi vertical puller, 1197

D

D log E, 1043
D log H curve, 1043
Daguere, Louis, 1345
Daguerreotype, 1259–1309, 1345
Dailies, 1033–1034, 1044
Damping force, 226, 231
Dancer roll, in gravure printing, 459–460
Dark adaptation, 519
Dark current, 396, 1188–1190
Dark discharge, 311–312
Dark field image, 269
in scanning transmission electron microscopes (STEM), 277
in transmission electron microscopes (TEM), 264
Dark fringes, 243, 269–270
Dark noise, 785–786
Dark signal, 397
Darkfield microscopy, 1127–1128
Darkness, overall, 604
Darkroom techniques, in forensic and criminology research, 724–725
Darwin–Howie–Whelan equations, 282, 284
Data acquisition
in infrared imaging, 807
in scanning acoustic microscopy (SAM), 1230
Data rate, 50, 63–64, 67, 81, 84
Data reduction processes, in feature recognition, 353–358
Data transmission, over TV signals, 1389
Databases
in biochemical research, 699
in search and retrieval systems, 617–637
Dating, (EPR) imaging for, 296–297
Daylight, 524, 525, 1044
DC restoration, cathode ray tube (CRT), 180
De Broglie, Louis, 261
De-exicitation of atoms, 253
De Morgan's law, 431, 435
Dead layer, 381
Decay, 254, 260, 980–981
Decibels, 1044
Decision theory, 161, 359–363
Decomposition, in search and retrieval systems, 627
Deconvolution, 595–596
Defect analysis and detection
electron paramagnetic resonance (EPR) imaging for, 295–296
feature recognition and object classification in, 355
microscopy and, 1116–1117
scanning capacitance microscope (SCM) for, 22
in scanning acoustic microscopy (SAM) analysis for, 1228
silver halide, 1268–1269
Defense Meteorological Satellite Program (DMSP), 890–904, 929
Definition, 1044
Deflection angle, cathode ray tube (CRT), 35–36
Deflection yoke, CRT, 31, 41–43, 46, 48
Deflectometry, flow imaging and, 412
Defocus, 272, 555, 1088
Deinterlacing, in forensic and criminology research, 725–726
Delphax Systems, 301, 322
Delrama, 1044
Delta function, 1103–1104
Delta rays, particle detector imaging and, 1156
Denisyuk, N., 504
Denoising, wavelet transforms in, 1448
Densitometer, 1044
Density
film, 1357
gravity imaging and, 449–450
human vision and, 513, 598–616, 598
in motion pictures, 1044
Density gradients, flow imaging and, 405
Depth cues, in three-dimensional imaging, 1327–1328
Depth of field (DOF), 403, 1044, 1124, 1347
Depth of focus, 403, 1044, 1124
Derived features, 357–358
Desertification, geologic imaging and, 655–656
Designer, 1044
Desktop publishing, 300
Destructive interference, 241
Detection theory, digital watermarking and, 161
Detectivity, photodetector, 1188
Detector arrays, photoconductive, 1203–1205
Detectors
background limited infrared photodetector (BLIP), 1189
charge mode detectors, 1191–1193
diamond, 1168
fabrication and performance of, 1194–1198
in infrared imaging, 805, 806
lidar, 873
magneto/ in magnetospheric imaging, 1007
neutron/in neutron imaging, 1058–1062
in overhead surveillance systems, 785, 790–793
particle detector imaging, 1154–1169
photoconductor, 1169–1183

photodetectors, 1183–1208
photoelectric, 1169
pixel, 1167–1168
radiography/in radiographic imaging, 1067
scintillator, 1062–1064
semiconductor, 1064–1065, 1163–1165, 1168
silicon drift, 1167
silver halide, 1259–1309
strip type, 1165–1167
in ultrasonography, 1418–1419
X-ray fluorescence imaging and, 1477
Deuteranopia, 522–523
Developers and development processes, 1356–1357
color photography, 129–131
in electrophotography, 300, 312–322
in forensic and criminology research, 724–725
holography in, 509
instant photography and, 827, 830, 834–842
in motion pictures, 1034, 1044
silver halide, 1299–1302
Development inhibitor releasing (DIR) couplers, 129, 137–139
Diagnostic imaging, art conservation and analysis using, 665–680
Dialogue, 1044
Diamond detectors, 1168
Diaphragm, 1044, 1080
Dichroic coatings, 1044
Dichroism, 233
Dicorotron charging, 303
Diderot, Denis, 455
Dielectric constants, 225
Dielectrics, speed of light in, 224–225
Difference of Gaussian (DOG) function, 80
Difference threshold, 612
Differential absorption lidar (DIAL), 869, 874, 882
Differential interference contrast (DIC), 1106, 1116, 1134–1135
Differential pulse code modulation (DPCF), 67, 786
Differential scattering, 250, 410
Diffraction, 3, 58, 246–253, 278–286, 1073, 1082, 1083, 1087
in charged particle optics, 94
convergent beam, 268
flow imaging and, 403
image formation in, 571
microscopy and, 1107–1108, 1110, 1126, 1128
in motion pictures, 1044
in overhead surveillance systems, 792–794
in transmission electron microscopes (TEM), 266–274
in ultrasonography, 1424–1428
Diffraction contrast imaging, TEM, 268–270
Diffraction limited devices, 249
Diffraction limited PSF, 1087
Diffraction theory, 3
Diffractometers, 244
Diffuse reflection, 234
Diffusers, 509, 528
Diffusion, 404, 525, 528, 1044
Diffusion coefficients, electron paramagnetic resonance (EPR), 294
Diffusion etch, gravure printing, 460
Diffusion transfer process, 456
Diffusion transfer reversal (DTR), 827, 833
Diffusivity, 404
Digital cameras, 854–855
Digital cinema, 1039–1040
Digital compression standards (DCS), 156
Digital effects, 1044
Digital imaging, 49–50, 516
art conservation and analysis using, 662–665, 680
CMOS sensors and, 1139–40
in forensic and criminology research, 722–723
instant photography and, 827, 854–855
in medical imaging, 754
microscopy and, 1106
in overhead surveillance systems, 786
in photographic color display technology, 1216–1217
photography, 1358
in photomicrography, 1137–1140
quality metrics and, 602–603
in search and retrieval systems, 616–637
television, 1374–1380
video, 1374–80
wavelet transforms in, 1447–1450
Digital intermediate, in motion pictures, 1035
Digital Library Initiative (DLI), 616
Digital light processor (DLP), 183, 185
Digital photography, 141–142, 827
Digital photomicrography, 1138–1140
Digital processing
Cineon digital film system, 1042
in forensic and criminology research, 725
in motion pictures, 1039–1040, 1044
in ultrasonography, 1429–1430, 1429
Digital rights management (DRM), 159
Digital sound, 1031, 1033, 1037, 1044
Digital television, 1374–1380
Digital to analog conversion (DAC), 721
Digital versatile disk (DVD), 159
Digital video (*See also* Compression), 146–157
Digital Video Broadcast (DVB), 1392
Digital watermarking, 158–172, 158
Digitizers, in forensic and criminology research, 709, 722–723
Dilation, 430, 433, 584–589, 1446
Dimension Technologies Inc. (DTI), 1338–1339
Dimensionality reduction, in feature recognition, 363–364
Dimmer, 1044
Dipole moment matrix element, 254
Dirac functions, 57
Direct image holography, 505
Direct mapping, in RF magnetic field mapping, 1125–1126
Direct thermal process, 196, 853
Direct transfer imaging, gravure printing, 461
Direction finding (DF) lightning locators, 890, 906–907
Directional response characteristics, 1
Directivity patterns, 1
Director, 1044
Dirichlet conditions, 1103
Disaster assessment, geologic imaging and, 651–655
Discharge cycle, 310–312, 1176
Discharge lamps, 222
Discharged area development (DAD), 317
Discontinuities, SAM, 1243–44
Discounting of illuminant, 520–521, 558, 561–562
DISCOVERER mission, 777
Discrete cosine transform (DCT)
compression, 154
digital watermarking and, 171
in search and retrieval systems, 631–632
video, 1387–88
Discrete element counters, 1060
Discrete Fourier transform (DFT), 57, 61, 67–68, 80, 263, 1405, 1448
Discrete wavelet transform (DWT), 1446–1447

Discrimination, 607
Disjunctive granulometry, 441
Dislocation contrast, TEM, 270
Dispersion, 225–229, 234–235, 466, 1081
Dispersive power, 234
Displaced frame difference (DFD), 155
Displacement current, 212
Displays, 172–199
  autostereoscopic, 1336–1341
  characterization of, 172–199
  field emission display (FED) panels, 374–389
  flat panel display (FPD), 374
  holography in, 510
  liquid crystal (LCD), 955–969
  in medical imaging, 754
  photographic color technology, 1208–1222
  secondary ion mass spectroscopy (SIMS) in, 482–484
  in three-dimensional imaging, 1331–1335
  in ultrasonography, 1413–1415
Dissolves, 1034, 1044
Distortion, 82, 92–94, 398, 1085, 1106
Distributors, 1044
DNA (*See* Genetic research)
Doctor blade, in gravure printing, 454–455, 458
Documentation, in forensic and criminology research, 716–717
Dolby Digital Sound, 1033, 1044
Dollies, 1030, 1044
Dolph–Chebyshev shading, 3
Dopants and doping
  in field emission displays (FED), 384
  germanium photoconductors, 1204–1205
  in photographic color display technology, 1214
  photoconductors and, 1171
  secondary ion mass spectroscopy (SIMS) in analysis, 487–489
  silver halide, 1293–1294
Doppler lidar, 870
Doppler radar, 223, 758, 764, 766, 767, 768, 772, 1142, 1458–1468
Doppler shift, 764, 772
  astronomy science and, 685–686
  flow imaging and, 415–416
  lidar and, 879–881
  magneto/ in magnetospheric imaging, 1017
  planar laser induced fluorescence (PLIF) in, 863
  radar and over the horizon (OTH) radar, 1142, 1145
  in ultrasonography, 1418, 1430–1432
Doppler ultrasound, 1430–32
Dot matrix printers, 300
Dots per inch (DPI), 602
Double exposures, silver halide, 1288
Double frame, 1044
Double refraction, 1131–1132
Double system recording, 1033, 1044
Doublet, achromatic, 235
Downward continuation, gravity imaging and, 452
DPACK calibration, 27
DPCM, 83
Drag force, 226
DRAM, scanning capacitance microscope (SCM), 22
Drift correction, gravity imaging and, 446
Drift tube scintillation tracking, 1161
Drive mechanisms,
  display, 382–383
  printer, 193
Driven equilibrium, in magnetic resonance imaging (MRI), 992
Driving force, 226
Dry developers, in electrophotography, 300
Dry process printers, 195
Dryers, gravure printing, 459
Dual color polymer light emitting diodes (PLED), 820–822
Dual mapping, 431
Dubbing, 1044
Dupe, dupe negative, 1044
Duplexers, radar, 1452
DuraLife paper, 1211
Dwell coherent integration time, radar and over the horizon (OTH) radar, 1147
Dwells, radar, 1146
Dyadic set, 1446
Dye bleach, 1217
Dye desensitization, 1298
Dye diffusion thermal transfer (D2T2), 853
Dye lasers, 885
Dye sublimation printers, 189–190, 189, 194–195, 194
Dye transfer printing, 188–197, 827
Dyes, 1356
  in color photography, 123–125, 131–133, 140
  in dye transfer printing, 188–197
  in instant photography, 827, 834–842
  in motion pictures, 1045
  silver halide, 1295–1296
Dynamic astigmatism, electron gun, 40–41
Dynamic spatial range (DSR), 404–405, 415

E

E field Change Sensor Array (EDOT), 890–904
Earth Observing System (EOS), 659, 772
Earth Probe, 660
Earth Resources Technology Satellite (ERTS), 778–779
Earthquake imaging, 453
Eastman Color film, 1024
Eastman Kodak, 498
Eastman, George, 1022
Echo, in magnetic resonance imaging (MRI), 981–983, 981
Echo planar imaging (EPI), 989, 992
Echo signal processing, in ultrasonography, 1419–1420
Edge detection, 517
Edge enhancement, 580–583
Edge finding, 582
Edge following, 642
Edge histograms, 626
Edge numbering, edge codes, film, 1026–1027, 1045
Edge sharpness, 1357
Edge spread function (ESF), 751–752, 1091–1092
Edgerton, Harold, 774
Edison, Thomas, 1022
Edit decision list (EDL), 1035, 1045
Editing, motion pictures, 1034–1035, 1045
Effective/equivalent focal length (EFL), 1078
EI number, film, 1023
Eigenfunctions, eigenstates, 285–286
8 mm film, 1025
Einstein Observatory Telescope, 1507
Einstein, Albert, 211
Einstein's coefficient of spontaneous emission, 253
Einstein's Theory of Special Relativity, 228
Ektapro process, 498
Elastic scattering, 249, 338
Elastic theory, liquid crystal displays, 959
Elasticity imaging, in ultrasonography, 1433–1434
ELDORA radar, 1457, 1471
Electric dipole radiation, 218, 220
Electric discharge sources, 222–223
Electric field imaging, X-ray fluorescence imaging and, 1489–1494
Electric Field Measurement System (EFMS), 890–904, 908
Electric permittivity, 212

Electric polarization, 227
Electrical conductivity, 230
Electrical discharge, 910–911
Electrical fields, magnetospheric imaging, 1002–1021
Electrical Storm Identification Device (ESID), 907, 922–924
Electro–optic effect, 233
Electro optics, terahertz electric field imaging and, 1394–1396
Electrocardiogram (ECG), 198
Electrochemical microscopy (SECM), 1248–1259
Electroencephalogram (EEG), 198–210, 744
Electroluminescent display, 817–827
Electromagnetic pulse (EMP), 909, 941
Electromagnetic radiation, 210–261, 682, 803, 1072, 1393
Electromagnetic spectrum, 218–220, 1072
Electromechanical engraving, 461–462
Electromyography, force imaging and, 424
Electron beam, cathode ray tube (CRT), 32–34
Electron beam gravure (EBG), 462
Electron beam induced current (EBIC), SEM, 276
Electron channeling, 285–286
Electron gun, 31, 39–45, 48, 87–88, 173
Electron magnetic resonance (EMR), 287
Electron microscopes, 87, 261–287, 573, 590, 594
Electron paramagnetic resonance (EPR) imaging, 287–299, 1223–1227
Electron positron annihilation, 220
Electron radius, classical, 250
Electron sources, in charged particle optics, 87–91
Electronic endoscopes, 334
Electronic flash, 1348–1349
Electrophotography, 299–331, 574, 598–616, 1174–1183
Electroplating, gravure printing, 457–458
Electrosensitive transfer printers, 195
Electrostatic energy analyzers (ESA), 482
Electrostatic image tube cameras, 498
Electrostatic transfer, 323
Electrostatics and lightning locators, 911–912
Elementary bit streams, 1382
Elliptical polarization, 232
Emission, 223, 253–256, 259, 379, 383–384, 387, 804
Emission computed tomography (ECT), 743
Emission electron microscope (EMM), 479
Emission ion microscope (EIM), 478
Emissivity, 804, 1072
Emittance, 804
  in charged particle optics, electron gun, 88
  in infrared imaging, 807–810, 813–814
Emitters, in charged particle optics, 89–91
Emmert's law, 1330
Empty magnification, microscopy and, 1121
Emulsion speed, 1045
Emulsions, 1023, 1039, 1045, 1208–1222, 1268
Encircled energy, 1090
Encoding, 50, 61–62, 152, 152–153, 1045
Encryption (*See also* Steganography), 160
Endogenous image contrast, MRI, 988
Endoscopy, 331–342
Energetic neural atom (ENA) imaging, 1003–104, 1006–1016
Energy and momentum transport by, 213–214
Energy density, in transmission electron microscopes (TEM), 213
Energy exploration, geologic imaging and, 650
Energy flux, astronomy science and, 688–690
Energy levels and transitions in, 215
Engraving, 454–462
Enhanced definition TV (EDTV), 1391–1392
Enhancement of image
  color/in color image processing, 117–119
  feature recognition and object classification in, 351–353
  in forensic and criminology research, 722
  in medical imaging, 756
  in overhead surveillance systems, 787
  SPECT imaging, 1316–1321
Enteroscopes, 331–332
Entrance pupil, 1080, 1354
Entropy, tomography/in tomography, 1408
Entropy coding, 83, 154–155
Environmental issues, color photography, 141
Eötvös correction, 448
Equalization, 725, 755, 1360
Equatorial anomaly, 1144
Ergonomics, force imaging and, 424
Erosion, 430, 432–434, 584–589
Error correction, television, 1390
Error handling, 156, 174–175
Estar films, 1023, 1045
ETANN neural network, 371–373
Etching
  in field emission displays (FED), 384
  gravure printing, 456, 460–461
Ethylenediaminodisuccinic acid, 141
Euclidean distance functions, 624, 646
Euclidean granulometries, 439–442
Euclidean mapping, 586–587
Euclidean properties, 439
Euclidean set theory, 430
Euler's formula, 226
European Broadcasting Union (EBU), 1374
European Radar Satellite, 649
European Spallation Source (ESS), 1057
Evanescent waves, 238
Evaporation, 383
Event frequency, high-speed photography and, 493
Evoked brain activity, evoked potential (EEG), 199, 201–202
Ewald diffraction, 267
Ewald's sphere, 267, 279, 280
Exchange, 1045
Excimer lasers, 391
Excitation, in magnetic resonance imaging (MRI), 979–980
Excitation error, 268
Excitation of atoms, 253
Existing light, 1045
Exit pupil, 1080
Expanders, holography in, 509
Expectation maximum (EM), tomography/in tomography, 1408–1409
Exposure, 1045, 1352
  autoexposure, 1355–1356
  in electrophotography, 304–310
  electrophotography, 1176–77
  in forensic and criminology research, 723–724
  gravure printing, 460–461
  high-speed photography and, 492
  in motion pictures, 1043–1045
  silver halide, 1288
Exposure latitude, 1045
Extended definition TV (EDTV), 1382
External reflection, 236

Extinction contours, 282
Extinction distance, 280
Extraction, in search and retrieval systems, 622, 625
Extraction of features, 353–358
Extraneous marks, 604, 605
Extreme ultraviolet imaging (EUV), 1005–1006
Eye (*See* Human vision)
Eye tracker systems, 522
Eyepieces, microscope, 1108–1109, 1114–1124

F

f-number, 1045
f-stops, 1352, 1354
Fabry Perot devices, 246
Fade, 1034, 1045
Fakespace CAVE three-dimensional imaging, 1335
Fakespace PUSH three-dimensional imaging, 1334
Far distance sharp, 1347
Far field, 220, 911–912, 1425–1426
Far field approximation, 251
Far field diffraction, 246
Far ultraviolet imaging of proton/electron auroras, 1016–1020
Far Ultraviolet Spectographic Imager (FUV-SI), 1017–1020
Faraday effect, 975
Faraday, Michael, 1022
Faraday's laws, 211, 212
Fast, 1045
Fast Fourier transform (FFT), 1095, 1151, 1405
Fast On Orbit Recording of Transient Events (FORTE), 890–904, 929
Fast spin echo, MRI, 992
FASTC2D calibration, 27
Fawcett, Samuel, 456
Feature extraction, in feature recognition, 353–358
Feature Index Based Similar Shape Retrieval (FIBSSR), 626
Feature measurement (*See also* Measurement), 343–350
Feature recognition and object classification, 350–374
Fermat's principle of least time, 234
Fermi energy levels, 1170–1171
Ferric ethylendiaminetetraacetic acid (ferric EDTA), 138, 141
Ferric propylenediaminetetraacetic acid (ferric PDTA), 141
Ferricyanide, 138
Ferroelectric liquid crystal displays (FLC), 964–965
Feynman diagrams, 259
Fiber optics, 333–334, 509, 1063–1064
Fidelity, 50, 71–74, 81, 84, 598, 611, 612, 615
Field angle, 1080
Field curvature, 1085
Field effect transistors (FET), 1173, 1199
Field emission, in charged particle optics, 89
Field emission display (FED) panels, 374–389
Field emission guns (FEG), 277
Field emitter arrays (FEA), 89, 375, 376
Field number (FN), 1121
Field of view (FOV), 60
  lidar and, 870
  in magnetic resonance imaging (MRI), 986, 996
  microscopy and, 1121
  in motion pictures, 1045
  scanning capacitance microscope (SCM), 19
  in three-dimensional imaging, 1330, 1333
Field points, 251
  in three-dimensional imaging, 1330–1331, 1333
Field stop, 1080
Fields, television, 1359
Figures of merit (*See also* Quality metrics), 50, 62–64, 1409
Filaments, in charged particle optics, 87–88
Fill, 604
Fill light, 1045
Filling-in phenomenon, 516
Film
  acetate, 1039, 1040
  additive color films, 847–849
  Agfa, 1024
  antihalation backing, 1041
  art conservation and analysis using, 661–662
  ASA/ISO rating for, 1023, 1041
  aspect ratio, 1022
  backing, 1041
  balance stripe, 1041
  base, 1041
  black-and-white, 1023, 1041, 1356
  camera type, 1026–1027
  cellulose triacetate, 1042
  coaterless, in instant photography, 832
  color photography, 124, 139–142
  color reproduction in, 139
  color reversal intermediate, 1043
  color reversal, 1043
  color, 1023, 1043, 1356–1357
  containers for, 1039
  core for, 1043
  density of, 1357
  dichroic coatings, 1044
  dye stability in, 140
  Eastman Color, 1024
  edge numbering, edge codes in, 1026–1027
  EI number for, 1023
  8 mm, 1025
  emulsion in, 1023, 1039
  Fuji instant films, 849–851
  Fuji, 1024
  Fujix Pictrography 1000, 851–852
  granularity in, 140
  gravure printing, 461
  high-speed photography and, 498
  holography in, 509
  image formation in, 571
  image structure and, 140–141
  imbibition (IB) system for, 1024
  instant photography and, 827–829
  integral, for instant photography, 828
  intermediate, 1026, 1047
  Kodachrome, 1024
  Kodak instant films, 849
  Kodak, 1024
  laboratory, 1048
  length of, in motion pictures, 1025
  magazines of, 1026
  in medical imaging, 754
  modulation transfer function (MTF) in, 140–141
  in motion pictures, 1022–1023, 1045
  negative and reversal, 1023
  negative, 1049
  negative, intermediate, and reversal, 1023–1024
  neutron/in neutron imaging, 1065
  nitrate, 1023, 1039, 1049
  orthochromatic, 1050
  in overhead surveillance systems, 790
  Panchromatic, 1050
  peel apart, for instant photography, 828
  perforations in, in motion pictures, 1025–1026, 1050
  photographic color display technology, 1208–1222
  photomicrography, 1137–1138
  Pictrography 3000/4000, 852–853
  Pictrostat 300, 852–853
  Pictrostat Digital 400, 852–853
  pitch in, 1026
  Pocket Camera instant films, 847
  Polachrome, 848
  Polacolor, 843–844
  Polavision, 848

polyester, 1023, 1039, 1050
print type, 1024
projection type, 1023
quality metrics and, 598–616
rem jet backing, 1051
reversal type, 139, 1052
root mean square (rms) granularity in, 140
safety acetate, 1023, 1024, 1052
sensitivity or speed of, 124, 139
seventy/70 mm, 1025
sharpness and, 140
silver halide and, 140
sixteen/16 mm, 1024, 1025, 1052
sixty/ 65 mm film, 1024
Spectra instant film, 847
speed of, 1023
still photography, 1344–1358
Super *xxx*, 1025
SX70 instant film, 844–847
Technicolor, 1024
thirty/35 mm, 1022, 1024, 1054
Time Zero, 846–847
Type 500/600 instant films, 847
vesicular, 662
width of, in motion pictures, 1024–1025
Film base, 1045
Film can, 1045
Film cement, 1045
Film gate, 1028, 1036, 1045
Film gauge, 1045
Film identification code, 1045
Film perforation, 1045
Film to tape transfer, 1045
Filtered backprojection (FBP), 1405
Filtered Rayleigh scattering (FRS), 411, 412, 415
Filters and filtering, 55, 56, 59, 65, 68, 69, 70, 71, 73, 76, 80, 100, 262, 437, 1092–1100
alternating sequential, 437
color/in color image processing, 118–119
comb, 1366
digital watermarking and, 150, 167
downward continuation, 452
extreme ultraviolet imaging (EUV), 1006
flow imaging and, 394
in forensic and criminology research, 717
Gabor, 623
gravity imaging and, 450–452
haze, 1047
holography in, 509
human vision and, 548, 558–560, 566–567
image processing and, 578, 579, 589, 593–597
incoherent spatial, 1096
lidar and, 872
light, 1048
linear, 756
liquid crystal displays, 968
logical structural, 442
in medical imaging, 755–756
microscopy and, 1113–1114
in motion pictures, 1043, 1045
multispectral image processing, 101
neutral density, 1049
open-close, 437
polarizing, 1050
quadrature mirror filter (QMF), 622
quality metrics and, 611
Ram Lak, 1405
in search and retrieval systems, 623
spatial, 1100
SPECT imaging, 1322
strike, 452
upward continuation in, 451
vertical derivatives, 452
wavelet transforms in, 1447–1450
Final cut, 1046
Fine grain, 1046
Fine structure, 218, 254
Fingerprinting, digital watermarking and, 159
Finite ray tracing, 1083
First hop sky waves, 912
First order radiative processes, 253
First print, 1046
Fisher's discriminant, 364–366
Fixers, 1345
Fixing bath, 1046
Fixing process, 138–139, 324–325, 324
Flaking, 1046
Flame imaging, 409
Flange, 1046
Flare, in cathode ray tube (CRT), 182–183
Flash photography, 104, 492, 1348–1349
Flashing, 1046
Flat, 458–459, 1046
Flat-bed editing tables, 1034
Flat panel display (FPD), 374
Flat, motion pictures, 1031
Floating wire method, in magnetic field imaging, 975
Flooding disasters, geologic imaging and, 651–655
Flow imaging, 390–419, 501, 989–991
Fluid dynamics
flow imaging and, 390
gravity imaging and, 453
in magnetic resonance imaging (MRI), 991
Fluorescence and fluorescence imaging, 210, 223, 255, 259, 529
absolute fluorescence (ABF), 863
art conservation and analysis using, 661, 676–677
flow imaging and, 397
laser induced fluorescence (LIF), 408, 861–869
phosphor thermography, 864–867
planar laser induced (PLIF), 391, 408–409, 411–416, 861–864
pressure sensitive paint, 867–868
thermally assisted fluorescence (THAF), 863
X-ray fluorescence imaging, 1475–1495
Fluorescence microscopy, 1106, 1135–1137
Fluorescent lifetime, 254
Fluorescent sources, 524, 525, 529
Fluorescent yield, 255
Fluorochrome stains, microscopy and, 1137
Flux density, 214
Flux measurement, 972–973
Flux, lens-collected, 1081
Flux, photon, 215
Flux, reflected, 527–529
Fluxgate magnetometer, 974–975
Fluxmeters, 971–973
Focal length, 1073, 1078, 1354–1355, 1347
Focal plane, 1046, 1352
Focal plane array (FPA), 804, 805
Focal plane shutters, 1352
Focal point, 54
Focus, 1073, 1347
autofocus, 1356
flow imaging and, 403
ground penetrating radar and, 471
human vision and, 513, 555
in infrared imaging, 809–810
microscopy and, 1122, 1124
in ultrasonography, 1427–1428
Focus variation, holographic, 272
Focused ion beam (FIB) imaging, 90–91, 93, 479
Fog, 1046
Fold degeneration, 217
Foley, 1046, 1053
Follow focus, 1046
Foot, human, force imaging and analysis of, 421
Footage, 1046
Footlambert, 1046
Force imaging, 419–430

Force process, 1046
Forecasting and lightning locators, 909
Foreground, 1046
Forensics and criminology, 709–742, 1393
Foreshortening, 1328
Forgery detection, art conservation and analysis using, 661
Format, 1046
Format conversion, video, 720–722
Formation of images (*See* Image formation)
Forward error correction (FEC), 1390
Foundations of morphological image processing, 430–443
Four field sequence, television, 1366–1367
Fourier analysis, 1102–1106
  in forensic and criminology research, 731–732
  gravity imaging and, 448
Fourier descriptors, in search and retrieval systems, 625–626
Fourier series, 698, 1102–1103
Fourier transform infrared (FTIR) microscope, 667
Fourier transforms, 50–58, 77, 280, 285, 1073, 1088, 1092–95, 1098–1099, 1102–1104, 1448
  human vision and, 542
  image processing and, 591–594, 596
  in magnetic resonance imaging (MRI), 985, 987
  in medical imaging, 751, 756
  periodic functions and, 1104–1105
  tomography/in tomography, 1405, 1406
  in transmission electron microscopes (TEM), 263
  two dimensional, 1104–1105
Fovea, 513, 515, 522, 561, 566, 746–747
Fowler–Nordheim plot, 379
Fox-Talbot, William Henry, 455, 492, 1345
Foxfet bias, 1166
Fractals, feature measurement and, 349–350
Fractional k space, in magnetic resonance imaging (MRI), 992
Frame, 1046
Frame and film gate, 1028
Frame by frame, 1046
Frame grabbers, 709, 722–723
Frame line, 1046
Frame rates, high-speed photography and cameras, 495
Frame transfer CCD, 393
Frames, television, 1359
Frames per second (FPS), 1046
Fraud detection, digital watermarking and, 159
Fraunhofer diffraction, 246, 247–249, 264, 280
Fraunhofer lines, 235
Free air correction, gravity imaging and, 447
Free electron gas, 230
Free electron lasers, 223
Free induction decay, in magnetic resonance imaging (MRI), 980–981
Free viewing, in three-dimensional imaging, 1336
Freeze frame, 1046
Frei operator, image processing and, 582
Frenkel equilibrium, silver halide, 1270
Frequency, 213, 227, 293, 559
Frequency band power mapping, 200–201
Frequency domain, image processing and, 591–595
Frequency interference, ground penetrating radar and, 470–471
Frequency modulation (FM), television, 1362
Frequency multiplexing, television, 1366–1367
Frequency response, 50, 1046
Frequency spectrum, 1103
Fresnel approximation, 285
Fresnel diffraction, 246
Fresnel equations, 235–236, 1498
Fresnel rhomb, 239
Fresnel sine integral, 1091
Fringes, 243, 1102
Frit, in field emission displays (FED), 387
Frustrated total internal reflection (FTIR), 238
Fuji, 1024
Fuji Colorcopy, 840
Fuji instant films, 849–851
Fujix Pictrography 1000, 851–852
Full frame CCD, 393
Fuming, 1345
Functional MRI (fMRI), 744
Functional parallelism, 562, 563, 565
Fundamental tristimulus values, 534
Fur brush development, in electrophotography, 312
Fusing process, in electrophotography, 324–325
Futaba field emission displays (FED), 377

G

G strings, in search and retrieval systems, 628
Gabor filters, 623
Gabor, Dennis, 504
Gabriel graph (GG), 370
Gain, screen, 1046
Gallium arsenic phosphorus (GaAsP) photodiodes, 1200
Gallium arsenide (GaAs) photodiodes, 1172, 1200
Gamma, 1046
  in cathode ray tube (CRT), 176, 177, 179
  in liquid crystal displays (LCDs), 184
  in motion pictures, 1043
  silver halide and, 1261–62
  in television, 1362
Gamma radiation, 218–220, 257, 803
  art conservation and analysis using, 677–680
  astronomy science and, 682, 683, 688
  attenuation in, 260
  neutron imaging and, 1057–58
  photoconductors, 1169
  SPECT imaging, 1311–1313, 1323
Ganglion cells, human vision and, 517–518, 562, 563
Gastroscopes, 331
Gate, frame and film, in camera, 1028, 1046
Gauge, 1046
Gauss' law, 25, 211, 212
Gaussian distribution, 1157
Gaussian operator/function, 542, 581, 593, 637
Gaussian optics, 1078–1079
Gaussian rays, 1083–1084, 1085
Gelatin filter, 1046
Generalized functions, 1103
Generation, 220–224
Genetic research (*See also* Biochemistry and biological research), 694, 745–746
Geneva movement, 1046–1047
Geodesy, gravity imaging and, 444
Geodetic Reference System, 445
Geographic Information System (GIS), 453
Geoid, 445–446
Geological Survey, 453

Geology, 647–661
ground penetrating radar and, 464
in magnetic field imaging, 970
instant photography and, 855
magnetospheric imaging, 1002–1021
radar and over the horizon (OTH) radar, 1149
Geostationary Meteorological Satellite (GMS), 760
Geostationary Operational Environmental Satellite (GOES), 760, 778
Germanium photoconductors, 1190, 1197, 1204–1205
Gettering materials, in field emission displays (FED), 380
Ghosting
in radar, 1452
in three-dimensional imaging, 1330, 1333
Giant Segmented Mirror Telescope, 693
Gibbs phenomenon, 71
Gladstone–Dale relationship, flow imaging and, 406
Global change dynamics, 655–656
Global circuit and lightning locators, 910
Global field power (GFP), EEG, 203
Global Position and Tracking System (GPATS), 890–904
Global Positioning System (GPS), 445
Global processing, in human vision, 567–568
Glossy surfaces, 528
Gobo, 1047
Godchaux, Auguste, 456
Gold sensitization, in silver halide, 1292–1293
Gradient, in dye transfer printing, 193–194
Gradient echo, MRI, 981, 992–993, 998–999
Grain boundary segmentation, SIMS analysis, 487–489
Graininess, 598–616, 1047, 1303–1304, 1357
Granularity, 75, 140, 439, 604–605
Granulometric size density (GSD), 442
Granulometries, 439–442
Graphics cards, cathode ray tube (CRT), 174–175
Grasp, microscopy and, 1108
GRASP software, 706
Grassmann's law, 531
Graticules, microscope, 1113, 1121
Gratings, 244–246, 560, 1092, 1094, 1099, 1102
holography in, 508
human vision and, 565, 567
lobes of, 5
microscopy and, 1108–1109
X-ray telescopes, 1504–1506
Gravimeter, 446
Gravitation imaging, 444–454
Gravity, 444, 445
Gravity anomalies, gravity imaging and, 444–445
Gravure multicopy printing, 454–463
Gravure Research Institute, 459
Gray card, 1047
Gray levels, 68, 618, 622
color/in color image processing, 114–115
in medical imaging, 752
monochrome image processing and, 100
multispectral image processing and, 101
thresholding and segmentation in, 638–641
Gray scale, 589–591, 961, 1347
Gray surface, 804
Gray value, 103, 646, 1421
Gray, Henry F., 375
Graybody radiation, 222
Grazing incidence, X-ray telescopes, 1497–1499
Great Plains 1 (GP-1) lightning locators, 924–928
Green print, 1047
Greene, Richard, 376
Gridding, gravity imaging and, 450
Ground clutter, radar, 1453
Ground coupling, ground penetrating radar and, 468
Ground instantaneous field of view (GIFOV) overhead surveillance systems, 783, 791
Ground penetrating radar, 463–476
Ground reaction force (GRF), force imaging and, 419–420
Ground resolvable distance (GRD), 790
Ground sampled distance (GSD), 790, 791–792
Groundwater detection, 454, 464
Group of pictures (GOP), 1386
Group range, 1144
Group velocity, 228
Guide rails, 1047
Guide roller, 1047
Guillotine splice, 1047
Gutenberg, Johannes, 455
Gyromagnetic ratio, 217, 978

H

Haar wavelet, 1446
Hadronic cascades, 1158
Halation, 1047
Half-wave antennas, 220
Half-wave plates, 233
Halftone screening, 455
Halide, 1047
Hall effect/Hall generators, in magnetic field imaging, 973–974
Halogen, 135, 222
Halos, 1135
Hamiltonians, 285
Hamming windows, 611
Handheld cameras, 1031
Hanover bars, 1371
Hard, 1047
Hard light, 1047
Harmonic imaging, 1432
Harmonics, 212, 225–226, 250, 972, 1092, 1098, 1104
Hartmann–Shack sensors, 547
Hazard assessment, geologic imaging, 651–655
Haze, 604
Haze filters, 1047
Head-end, 1047
Head recording, 1047
Head-up displays (HUD), 509
Heat transfer, in infrared imaging, 809
Heater, electron gun, 39
Heidelberg Digimaster, 300
Heisenberg uncertainty principle, 215
Helio Klischograph, 456–457, 461
Helium neon lasers, 508
Hell machine, 456, 461
Helmholtz invariant, 1080
Helmholtz–Kirchoff formula, 1486
Helmholtz–Lagrange relationship, 481
Hermitian transforms, 987
Herschel, John, 1022, 1345
Hertz, 213, 1047
Hertz, Heinrich, 211
Heterojunction photoconductors, 1173
Hewlett-Packard Laserjet, 302
Heyl, Henry, 1022
Hi Vision television, 1391
High-definition TV (HDTV), 41, 42, 47, 147, 151, 153, 157, 1039, 1047, 1382, 1390
High-Energy Neutral Atom Imager (HENA), 1007–1010
High-energy radiation (*See* X-ray; Gamma radiation)
High-frequency voltage, scanning capacitance microscope (SCM) vs., 20

High-frequency waves, radar and over-the-horizon (OTH) radar, 1142
High-pass filters, image processing, 593–597
High-resolution electron microscopy (HREM), 273
High-resolution images, TEM, 270
High-resolution secondary ion mass spectroscopy, 477–491
High-resolution visible (HRV) imaging systems, 649, 655
High-speed cameras, 1047
High-speed photographic imaging, 491–504
High-voltage regulation, cathode ray tube (CRT), 180
Highlights, 1047
Hindered amine stabilizers (HAS), 296
HinesLab three-dimensional imaging, 1341
HIRES geologic imaging, 660
Histograms
  art conservation and analysis using, 666
  in forensic and criminology research, 723
  image processing and, 584
  in search and retrieval systems, 619–620, 626, 629
  in medical imaging, 755
  thresholding and segmentation in, 637–638, 640
Hit or miss transform, 434–436
HLS coordinate system, 619
HMI lights, 1047
Hoffman Modulation Contrast, 1106, 1132–1134
Hoffman, Robert, 1106, 1133
Hold, 1047
Holographic optical elements (HOE), 509, 510
Holographic PIV (HPIV), 417
Holography, 223, 262, 504–512, 1328
  flow imaging and, 417
  image formation in, 571
  inverse X-ray fluorescent holographic (IXFH) imaging, 1486–1489
  normal X-ray fluorescent holographic (NXFH) imaging, 1484–1486
  stereograms using, 1336–1337
  in three-dimensional imaging, 1336–1337
  in transmission electron microscopes (TEM), 272
  X-ray fluorescence imaging and, 1484–1489
Homogeneity, in cathode ray tube (CRT), 181–182
Homologous points, in three-dimensional imaging, 1329
Horizontal gradients, gravity imaging, 452
Horn, Hermann, 456
Hot, 1047
HSB color coordinate system, 641
HSI color coordinate system, 641
HSI coordinate system, 112–114
HSV color coordinate system, 619, 641
Hubble telescope, 595
Hue, 103, 111, 117–119, 578, 580, 618, 1047
Human vision, 49–51, 84, 122–142, 512–570, 746–748
  color vision in, 122–142
  in color image processing, 101
  display characterization in and, 186
  feature recognition and object classification in, 353–358
  image processing and, 583
  optical geometry of, 54
  persistence of vision, 1021–1022, 1050
  in three-dimensional imaging, depth cues and, 1327–1328
Humidity, 1047
Huygen's principle, 242–243, 246
Huygenian eyepieces, 1120
Huygens, Christian, 243
Hybrid image formation, 571, 574
Hybrid ink jet printing technology (HIJP), 818–819
Hybrid scavengeless development (HSD), electrophotography, 312, 320–321
Hydrazone, 1179
Hydrology, ground penetrating radar, 464
Hydroquinone, instant photography, 834–839
Hydroxylamine, instant photography, 839
Hyperfine splitting, electron paramagnetic resonance (EPR), 288
Hyperfocal distance, 1347
Hyperspectral imaging, in overhead surveillance systems, 787
Hypo, 1046, 1047
Hyposulfite, 1345

I

I frame, video, 1387
iWERKS, 1031
I-Zone camera, 847
Idempotent operators, 436–437
Identification systems, instant photography, 855–856
Idle roller, 1047
IHS coordinate system, 111–112
Infrared light, 356
IKONOS satellite, 780
Ilfochrome, in photographic color display technology, 1217
Illuminance, 1345
Illuminants, 103–104, 524–527, 558, 561, 610–611
  discounting, in human vision, 520–521
Illumination, 243, 528, 558, 1072, 1101
  in art conservation and analysis, 665
  in endoscopy, 335–336
  in holography, 504
  Kohler, 1126
  matched, 1099
  in microscopy and, 1110–1014, 1125–1128
  in monochrome image processing, 100
  Rheinberg, 1128
  standard illuminants and, CIE, 103–104
Illuminators, microscope, 1107, 1125–1127
Image, 1047
Image aspect ratio (*See* Aspect ratio)
Image authentication, digital watermarking, 161
Image chain, in overhead surveillance systems, 782–800
Image combination, image processing, 589–591
Image correction, flow imaging, 397
Image dissection cameras, 498
Image enhancement (*See* Enhancement of image)
IMAGE EUV, 1006
Image fidelity, 598, 611, 612, 615
Image formation, 571–575
  in human eye, 541–544
  in microscope, 1107–1114
Image gathering, 49, 54–67, 84
Image integrity, in forensic and criminology research, 740–741
Image manipulation, instant photography, 856
Image-on-image or REaD color printing, 328–329, 328
Image plates, in neutron imaging, 1065–1066

Image processing, 575–598
color (*See* Color image processing)
in dye transfer printing, 193–194
in endoscopy, 336–338
feature recognition and object classification in, 351–353
in forensic and criminology research, 719–732
in infrared imaging, 807
instant photography and, 828, 829–830, 833–842
in magnetic resonance imaging (MRI), 987–988
in medical imaging, 754–756
monochrome, 100–101
morphological, 430–443
multispectral, 101
in overhead surveillance systems, 787
in scanning acoustic microscopy (SAM), 1231–1233
silver halide, 1299–1303
wavelet transforms in, 1448
X-ray fluorescence imaging and, 1477–1478
Image processing and pattern recognition (IPPR), 351
Image quality metrics (*See* Quality metrics)
Image restoration, 49, 51, 60, 67–75, 84, 118–119, 167, 722–756
IMAGE satellite, 1018
Image search and retrieval (*See* Search and retrieval systems)
IMAGE SEEK, 618, 621
Image sensors, 1199–1200
Image space, 1075
Image vectors, in tomography, 1407
Imagebase, 630
Imagery Resolution Assessment and Reporting Standards (IRARS), 795
ImageScape, 633
Imaging arrays, photodetector, 1194–1198
Imaging satellite elevation angle (ISEA), 791–792
IMAX, 1031, 1335
Imbibition (IB) system, 1024, 1047
Immersion medium, microscopy, 1116
Impedance, boundaries, non-planar radiators, 9
Impedance, acoustic, in ultrasonography, 1415
Impression rolls, gravure printing, 458–459
Improved Accuracy from Combined Technology (IMPACT) sensors, 935, 937, 941
Impulse response, in transmission electron microscopes (TEM), 265
In-phase signals, 1–3
In-plane switching (IPS), liquid crystal displays, 963–964
In the can, 1048
In vivo imaging
electron paramagnetic resonance (EPR) imaging for, 297–298
endoscopy, 331–342
fluorescence microscopy, 1136–1137
terahertz electric field imaging and, 1398–1399
INCA, 1020
Incandescent sources, 222, 242, 524, 1037
Incoherent light, 232, 242, 1085
Incoherent spatial filtering, 1096
Incoherent transfer functions, 1098
Index of refraction, 225, 1075, 1079
in charged particle optics, 88
complex, 227
flow imaging and, 405, 412
human vision and, 551
microscopy and, 1109
radar, 1453
Indexed color, cathode ray tube (CRT), 176
Indexing, in search and retrieval systems, 617–618
Indium antimonide (InSb), 806–807, 1201
Indium gallium arsenide (InGaAs), 806, 1200
Indium tin oxide (ITO), 381, 819, 957
Indoaniline, 138
Indophenol, instant photography, 835
Induced dipole moment, 225
Induction coils, in magnetic field imaging, 971–972
Inelastic scattering, 249
Infinity corrected microscopes, 1106–1107
Infinity space, microscopy, 1107
Information capacity, 1082–1083
Information efficiency, 50, 64, 84
Information hiding, digital watermarking vs., 160
Information processing, holography, 510–511
Information rate, 50, 62–67, 72–74, 81, 84
Information theory, 99, 161, 168–171
Infrared imaging, 218–219, 230, 1393
art conservation and analysis using, 668–672
astronomy science and, 690, 691–693
geologic imaging and, 648, 660
in overhead surveillance systems, 789
National Imagery Interpretability Rating Scale (NIIRS), 795–800
phosphor thermography vs., 866–867
pigments and paints in, 668–672
quantum well infrared photodetector (QWIP), 1190, 1205
satellite imaging systems and, 758
Space Infrared Telescope Facility (SIRTF), 690, 691–692
Stratospheric Observatory for Infrared Astronomy (SOFIA), 692–693
Television and Infrared Observational Satellite (TIROS), 757, 777
thermography, 802–817
Initial phase, 213
Ink-jet printers
hybrid systems, 818–819
instant photography and, 853
organic electroluminescent display and, 817–827
quality metrics and, 598–616
shadow masks in, 825–826
Inks
in electrophotography, 321–322
in gravure printing, 454, 455, 458, 459
InP lasers, 22
InP/InGaAsP buried hetero structure laser, 22–23
Instant color photography, Polaroid Corp., 127
Instant photography, 827–859
Instantaneous field of vision (IFOV), 54, 805
Intaglio, gravure printing, 454–463
Integrity, image (*See* Image integrity)
Intensified CCD (ICCD), 393–394, 862
Intensity, 102, 110–111, 214, 241, 610, 611, 1047
in cathode ray tube (CRT), 177, 178–179
in charged particle optics, 89
feature measurement and, 344–345
flow imaging and, 397–398
image processing and, 578, 580
in television, 147–148
Interface contrast, in transmission electron microscopes (TEM), 269

Interference, 239–246
in cathode ray tube (CRT), 38–39
ground penetrating radar and, 470–471
Interferometry, 242, 246
astronomy science and, 691
flow imaging and, 405, 412
geologic imaging and, 648
holography in, 510
lightning locators, 907, 946–947
X-ray interferometric telescopes, 1503–1504
Interlace, 146–147, 725–726, 1048, 1359
Interline transfer CCD, 393
Interlock, 1047
Intermediate atomic state, 258
Intermediate films, 1026, 1047
Intermediate sprocket, 1047
Intermittent, 1048
Intermittent action high-speed cameras, 495
Intermittent contact SCM, 21–22
Intermittent movement, 1048
Internal conversion, 255
Internal reflection, 236, 237–239
International Color Consortium (ICC) standards, 172
International Commission on Illumination (CIE), 523
International Electrochemical Commission (IEC), 602
International Standards Organization (ISO), 139, 602, 1048
International Telecommunications Union (ITU), 102, 1362
Internegative, 1043, 1048
Interocular distance, 1329
Interplanetary magnetic fields (IMF), magnetospheric imaging, 1002–1021
Interposition, 1328
Interpositive films, 1034, 1048
Intersystem crossing, 255
Intertropical Convergence Zone (ITCZ), 655–656
Invariant class, in morphological image processing, 437
Inverse continuous wavelet transform, 1446
Inverse Fourier transform (IFT), 1095
Inverse square law, 214
Inverse X-ray fluorescent holographic (IXFH) imaging, 1486–1489
Inversion recovery, in magnetic resonance imaging (MRI), 993–994
Iodate, 1213
Iodide, 1212–1214
Ion beam-induced chemistry, in charged particle optics, 87
Ion fraction, secondary ion mass spectroscopy (SIMS), 478
Ion-induced secondary electrons (ISE), 482–483
Ion-induced secondary ions (ISI), 482–483
Ion selective electrodes (ISE), 1248–1259
IONCAD, 482
Ionic conductivity, silver halide, 1271
Ionization, 1154–1157, 1159–1162, 1173
Ionographic process, 301, 322
Ionospheric analysis, 219, 1141, 1149
Ionospheric plasma outflow, ENA imaging, 1012–1016
Ionospheric propagation, radar and over-the-horizon (OTH) radar, 1143–1145
IRE units, television, 1361
Iridium, in photographic color display technology, 1214
Iris, 112–113, 512
Irradiance, 50–55, 75, 80–83, 214, 524–525, 530, 782, 1079–1081, 1085, 1092, 1094, 1102, 1287–1288
Isostatic correction, gravity imaging, 448
Isotope decay, art conservation and analysis using, 679–680
Issacs, John D., 1022

J

Jacobian transformation, 97, 98
Jagged edges, 75
Johnson, C.L., 775
Joint motion, force imaging, 420
Josphson tunnel junctions, SQUID sensors, 9–15
JPEG, 154, 171, 519, 521, 631, 741
Jumping development, in electrophotography, 301, 312, 318–320
Just noticeable difference (JND), 747

K

K, 1048
K nearest neighbor classification, 366–370
K space, 574, 987–988, 992
Kell factor, television, 1362
Kernel operations, image processing, 577–578, 580
Kerr cells, high-speed photography, 492–493
Kerr effect, 233
Ketocarboxiamdie, 134
Keykode number, 1048
Keystoning, 1048, 1331
Kinematical theory, 286, 278–281
Kinematics, 421, 1154–1169
Kinematoscope, 1022
Kinescope, 1048
Kinetics, force imaging, 424
Kinetograph, 1022
Kinetoscope, 1022
Kirchoff's laws, 759
Kirkpatrick–Baez telescopes, 1502–1503
Klein–Nishina formula, 257
Klietsch, Karl, 456
Klystrons, radar, 1452
Knife edge technique, 402–403, 501
Knock-on electrons, 1156
Kodachrome process, 127–128, 1024
Kodacolor process, 127–128
Kodak, 165–168, 829–830, 849, 1024
Kohler illumination, 1110–1114, 1126
Kohler, August, 1106, 1110–1111
Kolmogorov scale, flow imaging, 404
KONTRON, 483
Kramers–Heisenberg formula, 1476
Kronecker delta, 78
Kuwahara filter, 582

L

L*a*b* coordinate system, 108–109, 630
L*u*v* coordinate system, 109, 537
Laboratory film, 1048
Laboratory image analysis, in forensic and criminology research, 732–740
Lagrange invariant, 1079–1081
Lambertian diffusers, 525
Lambertian operators, 525, 528, 782
Lambertian surfaces, 51
Laminar flow, resolution, 402–403
Lamps, in projectors, 1037
Land cameras, 827
Land, Edwin, 1331
Landau distributions, 1155–1156, 1159
Landmark-based decomposition, 627
Landsat, 353, 444, 453, 648, 778, 787
Laplacian of Gaussian (V2G) operator, 71
Laplacian operator, 580–581, 593
Large area density variation (LADV), 604
Large format cameras, 1352–1354
Larmor frequency, 979
Larmor's formula, 220, 221
Laser engraving systems, 462–463
Laser-induced fluorescence (LIF), 338, 408, 861–869

Laser printers, 195–196, 302, 306–310
Laser pumps, 223
Laserjet printer, 302
Lasers, 223, 255, 1177
  continuous wave (CW) lasers, 391
  dye lasers, 885
  in electrophotography, 306–310
  engraving systems using, 462–463
  excimer lasers, 391
  flow imaging and, 391–394, 416
  free electron, 223
  gravure printing, 462–463
  helium neon lasers, 508
  in holography, 507–508, 507
  laser-induced fluorescence imaging, 861–869
  lidar, 869–889
  liquid crystal displays, 958
  NdYAG, 391, 416
  in phosphor thermography, 866–867
  Q switching in, 391–392
  terahertz electric field imaging and, 1402–1403
  yttrium aluminum garnet (YAG), 391
Latent image, 1047, 1048, 1276–1284
Latent impressions, in forensic and criminology research, 712–714
Lateral geniculate nucleus (LGN), 563, 564
Lateral hard dot imaging, gravure printing, 461
Lateral inhibition, 1093
Latitude, 1048
Latitude correction, gravity imaging, 447
Lateral geniculate nucleus (LGN), 516, 518
Lattice, sampling, 50, 54, 57, 58, 59, 65, 71
Laue patterns, 244, 279
Launch Pad Lightning Warning System (LPLWS), 890–904, 908, 921–922
Law of reflection, 234
Law of refraction, 234
Law, H.B., 31
Layout, 1048
Lead chalcogenide photoconductors, 1200–1201
Leader, 1048
Leaf shutters, 1351
Least mean square (LMS) algorithm, 372
Least time, Fermat's principle, 234
Legendre functions, 253
Leith, E.N., 504
Lens, 54, 58, 59, 65, 91–100, 541, 1073, 1345, 1347, 1348, 1354–1355
  anamorphic, 1031, 1040
  Bertrand polarization lens, 1132
  charged particle optics, 86–100
  coatings on, 1042
  electron gun, 40
  eyepiece, 1119–1122
  flow imaging and, 392–394, 399
  flux in, 1081
  Gaussian optics in, 1078–1079
  in holography in, 505–506, 508–509
  in human eye, 54, 512–514, 540, 547, 560, 746
  in microscopy, 1106, 1107, 1117
  for motion pictures, 1029, 1048
  in overhead surveillance systems, 783
  in projectors, 1036–1037
  in scanning acoustic microscopy (SAM), 1233–1234
  in secondary ion mass spectroscopy (SIMS) in, 480–481
  in terahertz electric field imaging, 1397
  for wide-screen motion pictures, 1031
  zoom, 1029
Lenticular method, color photography, 127
Lenticular sheets, in three-dimensional imaging, 1337–1338
Letter press, 455
Leyden jars, 492
Lichtenberg, 299
Lidar, 223, 804, 869–889
Lif converter, 1064–1065
Lifetime, radiative, 254
Light
  speed of, 224–225
  wave vs. particle behavior of light in, 210–211
Light adaptation, human vision, 519–520
Light amplification by stimulated emission of radiation (*See* Lasers)
Light axis, 1048
Light detection and ranging (*See* Lidar)
Light emitting diodes (LED), 1177
  in electrophotography, 300
  in electrophotography, 305–306, 305
  multicolor organic, 822–825, 822
  in three-dimensional imaging displays, 1340
Light filter, 1048
Light in flight measurement, holography, 510
Light intensity, 1048
Light levels, human vision, 558
Light meter, 1048
Light microscope, 261
Light output, 1048
Light sensitive microcapsule printers, 195
Light sources, 524–527
Light valve, 1048
Lighting
  monochrome image processing and, 100–101
  in motion pictures, 1033
Lighting ratio, 1048
Lightness, 102–103, 618
Lightning Detection and Ranging (LDAR), 890–904, 914, 941–945
lightning direction finders, 906–907
Lightning Imaging Sensor (LIS), 890–904, 929, 932–935
Lightning locators, 572, 890–955
Lightning Mapping System (LMS), 890–904, 929
Lightning Position and Tracking System (LPATS), 890–904, 935
Lightning warning systems, 907
LightSAR, 660
Ligplot software, 706
Line metrics, 603–604
Line of sight, 541
Line scan image systems, 60, 1482
Line sources, continuous, 6–7
Line spread function (LSF), 402, 751–752, 1090–1091, 1097
Line spread profiles, cathode ray tube (CRT), 33–34
Linear attenuation coefficient, 260
Linear filtering, in medical imaging, 756
Linear operations, image processing, 577–578
Linear perspective, 1328
Linear polarization, 231, 233
Linearity, 1085
Lines per picture height, television, 1362
Linogram method, in tomography, 1406
Lip sync, 1048
Liquid crystal display (LCD), 176, 184–186, 955–969
  in field emission displays (FED) vs., 374
  in three-dimensional imaging, 1331, 1338–1339
  response time in, 387
Liquid gate, 1048

Liquid immersion development (LID), 312, 321–322
Liquid metal ion sources (LMIS), 90–91, 93, 478, 479
Liquid mirror telescopes (LMT), 872
Lithography, 456
  electron beam, 87
  quality metrics for, 598–616
  secondary ion mass spectroscopy (SIMS) in, 479
Lithosphere, gravity imaging, 444
Live action, 1048
Lobes, grating, 5
Local (residual) gravity anomaly, 448
Local motion signal, human vision, 566
Local processing, human vision, 567–568
Localization, 1074–1077
  in human vision and, 569
  in magnetic resonance imaging (MRI), 984–987
Locomotion, force imaging and analysis, 419–430
Log, camera, 1041
Logical granulometry, 442
Logical structural filters, 442
Long Range Lightning Detection Network (LRLDN), 890–904
Long shot, 1048
Longitudinal magnification, 1076
Lookup table (LUT), 575, 612
Loop, 1048
Lorentz electromagnetic force, 1157
Lorentzian, 229
Lossless coding, 50
Lossless compression, 741
Lossy compression, 741
Low Energy Neutral Atom Imager (LENA), 1013–1016
Low key, 1048
Low-pass filters, 593, 594–597, 611
Low resolution electromagnetic tomography (LORETA), 204–208
Lowe, Thaddeus, 773
LPE, scanning capacitance microscope (SCM) analysis, 22
Lubrication, 1048
Luma, 103
Lumen, 1048
Lumiere, Louis and Auguste, 1022
Luminance, 102, 109, 530, 611, 618, 1081
  in cathode ray tube (CRT), 173, 183
  in field emission displays (FED), 380, 387
  human vision and, 521, 564
  in liquid crystal displays, 957, 968
  in motion pictures, 1048
  in television, 148, 1365, 1367, 1375
Luminescence, 255
  in field emission displays (FED), 375
  in forensic and criminology research, 717–719
Luminiferous ether, 211
Luminosity, flow imaging, 397–398
Luminous efficiency function, 529–531
Luminous flux, 1345
Lux, 1048
Lyman alpha signals, 1017
Lythoge, Harry, 456

M

M scan ultrasonography, 1414–1415
M way search, in search and retrieval systems, 626
Mach bands, 81
Macro lenses, 1354–1355
Macula, 513, 514, 549, 746–747
Macular degeneration, 549
Magazines, film, 1026, 1049
Magenta, 1049
Magenta couplers, color photography, 135–136
Magic lasso, thresholding and segmentation, 644–645
Magic wand, thresholding and segmentation, 643–645
Magnetic brush developer, in electrophotography, 312–318
Magnetic dipole moment, 217
Magnetic direction finders (MDF), 906
Magnetic field imaging, 211–212, 803, 970–977
  cathode ray tube (CRT), 38
  electron paramagnetic resonance (EPR) imaging for, gradients in, 289–292
  energetic neutral atom (ENA) imaging, 1006–1010
  magnetospheric imaging, 1002–1021
  particle detector imaging and, 1157
  RF magnetic field mapping, 1223–1227
Magnetic permeability, 467
Magnetic permittivity, 212
Magnetic quantum number, 217
Magnetic resonance imaging (MRI), 210, 977–1002
  electron paramagnetic resonance (EPR) imaging for vs., 293–294
  image formation in, 571, 573, 574
  in magnetic field imaging, 970
  in medical imaging, 744, 745
  in RF magnetic field mapping, 1224
Magnetic resonance measurement, 970–971
Magnetic sector mass spectrometer, 481–482
Magnetic sound, 1037, 1049
Magnetic splitting, atomic, 217
Magnetic storms, 1145
Magnetic striping, 1049
Magnetic susceptibility effects, 297
Magnetic tape, 1049
Magnetic toner touchdown development, 312
Magnetic track, 1049
Magnetization transfer imaging, 991
Magnetoencephalograms, 12–15, 744
Magnetoinductive technology, 976
Magnetometers, 1015, 974–975
Magnetopause to Aurora Global Exploration (*See* IMAGE)
Magnetoresistivity effect, 975
Magnetospheric imaging, 1002–1021
Magnetron, 288, 1452
Magnets, for MRI, 980–981, 997–998
Magnification, 1076, 1079, 1080
  angular, 1076–1077
  longitudinal, 1076
  in microscopy, 1107, 1121, 1124
  in overhead surveillance systems, 783
  transverse, 1076
  visual, 1077
Magnification radiography, 673–674
Magnitude, 226, 241
Magoptical, 1049
MagSIMS, 479–484
Mahanlanobis distance measure, 624
Makeup table, 1049
Manhattan Project, 744
Mapping
  in biochemical research, 699
  geologic imaging and, 647–661
  gravity imaging and, 444–454
  in infrared imaging, 812–815
  lightning locators, 909
  in magnetic field imaging, 970
  RF magnetic field mapping, 1223–1227
  secondary ion mass spectroscopy (SIMS) in, 477
Marcus electron transfer theory, 1297
Markov random field, in search and retrieval systems, 623
Mars Express mission, 1020–1021
MARS systems, 622, 632
Masking, 129, 136–137, 139, 1049
Mass, gravity imaging, 444
Mass attenuation coefficient, 260

Mass spectrometry, secondary ion (SIMS), 477–491
Master, 1034, 1049
Master positive, 1049
Matched illumination, 1099
MATLAB, 186, 542
Matte, 1049
Matte surfaces, 528
Mauguin condition, 960
Maxima, 245
Maximum realizable fidelity, 73, 81, 84
Maxwell pairs, in magnetic resonance imaging (MRI), 984
Maxwell, James Clerk, 123, 211
Maxwell's equations, 212, 225, 233, 234
Maxwell's theory, 211–214
Mean size distribution (MSD), 442
Mean square restoration error (MSRE), 69–70
Measurement, 343–350
  in astronomy science, 686–688
  in color image processing, 120
  in colorimetry, 528
  in display characterization, 185–186
  in endoscopy, lesion size, 336–338
  in gravity imaging, 453–454
  in holography, 510
  in infrared imaging, 814–815
  in magnetic field imaging, 970–976
  in meteorological research, 758–759
  in planar laser-induced fluorescence (PLIF), 863
  quality metrics and, 598–616
  in radar imagery, 764–765
  in satellite imaging systems, 758–759
  in terahertz electric field imaging, 1396–1397
Media relative colorimetry, 533
Medial axis transform (MAS), 625
Medical imaging, 742–757
  infrared imaging for, 812
  instant photography and, 856
  lens in, 1354
  magnetic resonance imaging (MRI) in, 977–1002
  scanning acoustic microscopy (SAM) in, 1228
  single photon emission computed tomography (SPECT) in, 1310–1327
  terahertz electric field imaging, 1393–1404
  three-dimensional imaging in, 1327
  ultrasound, 1412–1435
  X-ray fluorescence, 1479–1482
Medium Energy Neutral Atom Imager (MENA), 1010
Medium format cameras, 1351
Medium shot, 1049
MegaSystems, 1031
MEH PPV polymer, 819
Memory color, 520–521, 612
Mercury cadmium telluride (HgCdTe or MCT), 807, 1190, 1201–1205
Mercury discharge tubes, 222
Mercury vapor lamps, 222
Mesopic vision, 515
Metal alloys, SAM analysis for, 1228
Metal halide lamps, 1037
Metal matrix composites, SIMS analysis, 484, 487
Metal oxide semiconductors (MOS), 18–19
Metal semiconductor diodes, 1173
Meteorology, 757–773
  airborne radar, 1471–1473
  ASOS Lightning Sensor (ALS), 907, 922
  Automated Surface Observing System (ASOS), 907, 922
  Defense Meteorological Satellite Program (DMSP), 890–904, 929
  Doppler radar, 1458–1468
  Electrical Storm Identification Device (ESID), 907, 922–924
  lidar and, 869–870
  lightning locators, 890–955
  magnetospheric imaging, 1002–1021
  mobile radar, 1471–1473
  Thunderstorm Sensor Series, 907
  Tropical Rainfall Measuring Mission (TRMM), 660, 771–772, 890–904, 929, 932–935, 1473
  weather radar, 1450–1474
  wind profiling radar, 1469–1471
Meter candle, 1049
Meteorological Satellite (METEOSAT), 760
Methodology for Art Reproduction in Color (MARC), 664
Methyliminodiacetic acid (MIDA), 141
Metz filters, 1322
Mexican hat wavelet, 1444
Michelson interferometer, 242
Michelson, Albert, 211
Micro dry process printers, 195
Microchannel plates (MCP), in neutron imaging, 1066–1067
Microcrystal preparation, silver halide, 1262–1268
Microcrystals, SIMS analysis, 484–486
Microfilm, art conservation and analysis using, 661–662
Microimaging, 1351
Micromachining, 87
Micron Display, 377
Micropattern gas detector scintillation tracking, 1162–1163
Micropolarizers, in three-dimensional imaging, 1334–1335
Microreagant action in SECM, 1256–1257
Microscopy, 210, 261, 1072, 1106–1141
  alternating current scanning tunneling microscopy (ACSTM), 28
  art conservation and analysis using, 666–668
  atomic force microscope (AFM), 16
  biochemistry research and, 693–709
  capacitive probe microscopy, 16–31
  in charged particle optics and, 87
  conventional fixed beam transmission electron microscopy (CTEM), 277
  darkfield microscopy, 1127–1128
  differential interference contrast (DIC) microscopy, 1106, 1134–1135
  digital photomicrography, 1138–1140
  electron microscopes, 261–287
  emission electron microscope (EMM), 479
  emission ion microscope (EIM), 478
  fluorescence, 1106, 1135–1137
  Fourier transform infrared (FTIR) microscope, 667
  high-resolution electron microscopy (HREM), 273
  Hoffman Modulation Contrast, 1106, 1132–1134
  infinity corrected, 1106–1107
  instant photography and, 855
  Nomarksi interference microscopy, 1106
  phase contrast microscopy, 265–266, 1106, 1128–1130
  photomicrography, 1106, 1124, 1137–1138
  reflected light microscopy, 1124–1127
  scanned probe microscopy (SPM), 1248

Microscopy, (*continued*)
scanning acoustic microscopy (SAM), 1128–1148
scanning capacitance microscope (SCM), 16–31
scanning electrochemical microscopy (SECM), 1248–1259
scanning electron microscope (SEM), 23, 87–88, 262, 274–278, 477, 1243
scanning evanescent microwave microscope (SEMM), 28
scanning ion microscope (SIM), 477
scanning Kelvin probe microscope (SKPM), 16, 28
scanning microwave microscope (SMWM), 16, 28
scanning transmission electron microscope (STEM), 87, 93, 262, 276–278
scanning transmission ion microscope (STIM), 479
secondary ion mass spectroscopy (SIMS), 477–491
stereomicroscope, 1106
transmission electron microscope (TEM), 23, 87, 93, 262–274, 262
Microstrip gas chambers (MSGC), 1060–1062
Microstrip, silicon/Gd, 1065
Microtips, molybdenum, in field emission displays (FED), 375–376
Microwave radar ducting, 1141
Microwaves, 218, 220, 230, 242, 288, 803, 1223–1227, 1393, 1452
Mid-foot key number, 1049
Mie scattering, 253, 548, 875, 1451
Miller indices, 267
Mineral resource exploration, geologic imaging, 650
Mineralogy, electron paramagnetic resonance (EPR) imaging for, 296–297
Minima, 245
Minimum density, 1049
Minimum ionizing particles (MIPs), 1155
Minkowski addition/subtraction, 433, 612
Mirrors, 1073, 1074, 1075
astronomy science and, 691
camera, 1346, 1349–1350
extreme ultraviolet imaging (EUV), 1006
Giant Segmented Mirror Telescope, 693
in holography, 508–509
liquid mirror telescopes (LMT), 872
in overhead surveillance systems, 783–784
quadrature mirror filter (QMF), 622
rotating drum and mirror cameras, 496–497
rotating mirror framing cameras, 497–498
thin mirror telescopes, 1501–1502
in three-dimensional imaging, 1341–1343
X-ray telescopes, 1498–1499
Mix, 1049
Mixed high-signals, television, 1365
Mixing studies, flow imaging, 409, 410
Mobile radar, 1471–1473
Modeling
in biochemical research, 699
quality metrics and, 610–615
three-dimensional imaging and, 1327
MODIS, 659–660
MODTRAN, 783, 785
Modulation, 1094
amplitude modulation (AM), 383, 1362
differential pulse code modulation (DPCF), 67, 786
frequency modulation (FM), 1362
pulse code modulation (PCM), 150
pulse width modulation (PWM), 383
quadrature amplitude (QAM), 149
quadrature modulation, 1365–1366
quantization index modulation (QIM), 170–171
sequential frequency modulation, 1367–1368
Modulation transfer function (MTF), 57–58, 95–96, 99, 140–141, 749, 750, 1091, 1094–1098, 1358
cathode ray tube (CRT), 34, 182
flow imaging and, 400–402
human vision and, 521, 544, 555
Moiré patterns, 38–39, 75
Molecular biology spectroscopy, 1403
Molecular medicine, 745–746
Molecular polarization, 227, 252
Molecular replacement, in biochemical research, 698–699
MOLMOL software, 706
MolScript software, 706
Molybdenum, in field emission displays (FED), 375–376
Moment-based search and retrieval systems, 626
Mondrian values, 52–53
Monitors (*See* Displays)
Monochromatic aberrations, 545–548, 1083–1085
Monochromatic primaries, 104
Monochrome image processing, 100–101
Morley, Edward, 211
Morphological image processing, 430–443, 584–589
MOSFETs, 22–23
Motion analysis, 424, 494
Motion detection, human vision, 566–567
Motion parallax, 1328
Motion pictures (*See also* Television; Video), 1021–1056
color in, 142
electrostatic image tube cameras, 498
high-speed photography and, 494–498
image dissection cameras, 498
intermittent action high-speed cameras for, 495
photographic color display technology, 1208–1222
Polachrome, 848
Polavision, 848
rotating drum and mirror cameras, 496–497
rotating mirror framing cameras, 497–498
rotating prism cameras, 495–496
Motion studies
in meteorological research, 769–771
in overhead surveillance systems, 784
Motorboating, 1049
Motorola, 377
Mottle, 604
Moving slit parallax barrier, 1338
Moviola, 1034, 1049
MOVPE, 22
MPEG, 156–157, 171, 631
MPS PPV, 821–822
MSMS software, 706
MSP software, 707
Mug shots, 715–716
Multi-up processing, 384
Multiangle Imaging Spectroradiometer (MISR), 772
Multiangle viewing instruments, 772
Multichannel scaler (MCS), 873
Multichannel sound, 1362, 1365
Multicolor organic light emitting diodes (OLED), 822–825

Multihop operations, radar and over-the-horizon (OTH) radar, 1141
Multilayer perceptron for pattern recognition, 371–373
Multilayer telescopes, 1503
Multipinned phase (MPP) CCD, 1198
Multiplanar imaging, 1328, 1341
Multiple beam interference, 244–246
Multiple carrier systems, television, 1365
Multiple coulombic scattering (MCS), 1157
Multiple isomorphous replacement (MIR), 698
Multiple wavelength anomalous dispersion (MAD), 698
Multiplex sound, television, 1362, 1365
Multiplexed analog component (MAC), television, 1374
Multiplexing, 1049
  in liquid crystal displays, 956, 957
  in television, 1366–1367, 1382–1385, 1389
Multislice method for crystals, 284–285
Multispectral imaging, 56, 101
  geologic imaging and, 660
  National Imagery Interpretability Rating Scale (NIIRS), 801
  in overhead surveillance systems, 787
Multispectral scanner (MSS), 352, 360, 779
Multispectral sensors, 356
Multivariate granulometry, 441
Multiwave proportional chamber (MWPC), 1160
Multiwavelength imaging, 682–686
Multiwire proportional chambers (MWPC), 1059–1060
Munsell system, 110
MUSE TV transmission systems, 1391
Muybridge, Edweard, 1022
Mylar, in electrophotography, 301

N

Naphthol, 133, 134
Narration, 1049
National Imaging and Mapping Agency, 445
National Imagery Interpretability Rating Scale (NIIRS), 795–800
National Integrated Ballistics Information Network (NIBIN), 716
National Lightning Detection Network (NLDN), 890–904, 924, 925, 935–941
National Television Standards Committee (NTSC) standards, 102, 146–149, 382, 519, 521, 619, 1049–1050, 1359–1393
NdYAG lasers, 391, 416, 867
Near distance sharp, 1347
Near field lightning locators, 911–912
Near field diffraction, 246
Near field imaging, 1400–1403, 1425–1426
Near instantaneous companded audio multiplex (NICAM), 1365
Near point, 1077
Nearest neighbor classification, 366–370
Negative and reversal film, 1023
Negative film, 1049
Negative image, 1043, 1049
Negative positive process, 1049, 1347–1348
Negative, intermediate, and reversal film, 1023–1024
Neighborhood operations
  in image processing, 576–578, 583–584
  in medical imaging, 755
  in search and retrieval systems, 623
  in thresholding and segmentation, 640
Network locator lightning locators, 905, 908
Network mapper lightning locators, 905, 909
Neural activity, electroencephalogram (EEG), 198–199
Neural networks, 371–373, 510–511, 632
Neurons and human vision, 513, 560, 563, 565
Neutral density filters, 1049
Neutral test card, 1049
Neutron activation autoradiography (NAAR), 678–680
Neutron imaging, 1057–71
Neutron radiography, art conservation and analysis using, 678–680
New York Times, 456
Newton's law of gravitation, 444
Newton's rings, 1049
Next Generation Space Telescope (NGST), 693
Niepce, Joseph, 455, 1345
Nier–Johnson geometry, 481
NIMBUS satellites, 778
Niobium, in SQUID magnetometers, 10
Nitrate film, 1023, 1039, 1049
Nitride oxide silicon (NOS), 23
Nitrogen, 135
Nitroxyl radicals, 289, 296
NMOS, scanning capacitance microscope (SCM), 22
Noise, 50, 60, 61, 62, 63, 64, 68, 69, 72, 74, 75, 80, 1049, 1357
  compression, 152
  in forensic and criminology research, 727–731
  in human vision and color vision, 517
  image processing and, 593–594
  in infrared imaging, 809
  in magnetic resonance imaging (MRI), 996
  in medical imaging, 748–749
  in overhead surveillance systems, 785
  photodetectors and, 1187–1188
  quality metrics and, 598–616
  radar and over-the-horizon (OTH) radar, 1142
  Rose model and medical imaging of, 753–754
  SPECT imaging, 1322
  thresholding and segmentation in, 639
  in ultrasonography, 1416–1417
  wavelet transforms in, 1448
Noise equivalent power (NEP), 1188
Noise equivalent temperature difference (NETD), 814
Noise factor, flow imaging, 395, 396
Nomarksi interference microscopy, 1106
Nomarski, Georges, 1106, 1134
Nondestructive evaluation (NDE)
  infrared imaging for, 811–812
  ultrasonography for, 1412
  X-ray fluorescence imaging for, 1478–1479
Nonlinear editing, 1035, 1230–1231
Nonlinearity, in ultrasonography, 1417–1418
Nonparametric decision theoretic classifiers, 359–363
Nonradiometric IR imaging systems, 804
Nonsymmetrical amplitude, 4
Normal dispersion, 227
Normal X-ray fluorescent holographic (NXFH) imaging, 1484–1486
Normalization, 532, 604–605, 1156
Normalized colorimetry, 533
Normally black (NB) liquid crystal displays, 963
Normally white (NW) liquid crystal displays, 960

North American Lightning Detection Network (NALDN), 890–904, 935
Notching, 1049
Nuclear magnetic resonance (NMR)
in biochemical research, 694, 699, 705, 706
electron paramagnetic resonance (EPR) imaging for vs., 287, 293–294
in magnetic field imaging, 971
RF magnetic field mapping, 1223–1227
Nuclear systems, 215
Nucleation and growth method, for silver halide, 1276–1280
Nucleic Acid Database (NDB), 699
Numerical aperture, 1081, 1115
Nyquist frequency, 790
Nyquist limit, 401, 561
Nyquist's criterion, MRI, 986
Nyquist's sampling theorem, 552

O

O software, 707
Object classification, human vision, 569
Object recognition, 520–521
Object shaped search and retrieval systems, 625–628
Object space, 1075
Objective quality metrics, 602–606
Objectives, microscope, 1114–1124
Observed gravity value, 446
Ocean wave height, radar and over-the-horizon (OTH) radar, 1149
Oceanography, 760
Oculars (*See* Eyepieces)
Odd and even field scanning, 146
Off axis aberration, 547
Off axis holography, 272
Off line editing, 1050
Offset gravure, 460
Oil exploration, 454, 650
Oil immersion microscopy, 1109–1110
Omnimax, 1031
On axis aberration, 545–546
Online editing, 1050
OPAL jet chamber, 1155
Opaque, 1050
Open cascade development, 312
Open-close filters, 437
Openings, morphological, 430, 436–438, 585
Optic nerve, 513, 552, 562
Optical activity, 233
Optical axis, 54, 540–543
Optical character recognition (OCR), 680
Optical coherence tomography (OCT), 339
Optical density, 234
Optical design, 1074
Optical disks storage 483
Optical effects, 1050
Optical fibers, 509
Optical geometry, 54
Optical image formation, 571, 1072–1106
Optical microscopy, 1106–1141
Optical power, 1074
Optical printer, 1050
Optical resonators, 223
Optical sectioning, microscopy, 1133–1134
Optical sound, 1031, 1037, 1050
Optical thickness, 1135
Optical transfer function (OTF), 57, 1092–1100
aberrations and, 1095–1098
flow imaging and, 399–402
in overhead surveillance systems, 784
Optical Transient Detector (OTD), 890–904, 929–932
Optics (*See also* Lens)
astronomy science and, 691
flow imaging and, 392–393
holography in, 508–509
in human vision, 558–560
microscopy and, 1106–1107
paraxial, 1073–1074
secondary ion mass spectroscopy (SIMS) in, 480–481
terahertz electric field imaging and, 1394
Optimization criterion, in tomography, 1407
Optimum classification, 364–366
OR, 590
Orbital angular momentum, 216
Orders, microscopy, 1108
Organic electroluminescent display, 817–827
Organic ligands, in photographic color display technology, 1214
Organic light emitting diodes (OLED), 817, 822–825
Original, 1050
Orientation, human vision, 565
Original negative, 1050
Orthochromatic film, 1050
Orthogonal frequency division multiplexing (OFDM), 1389
Orthogonal wavelet basis, 1446
Orthopedic medicine, force imaging, 422
Oscillating planar mirror displays, 1341–1342
Oscillations of waves, 226, 803
Oscillator strength, 229
Oscillators, 242
Oscilloscope, cathode ray tube (CRT), 47
Ostwald system, 110
Out of phase signals, 2
Out take, 1050
Outgassing, in field emission displays (FED), 370–380
Over-the-horizon (OTH) radar, 1141–1153
Overall darkness, 604
Overexposure, in electrophotography, 318
Overhead surveillance (*See also* Aerial imaging), 773–802
Overlap splice, 1050
Overlay, 1050
Overshoot tolerance, 912–914
Oxichromic developers, 835
Oxidation of dyes, 836–839
Oximetry, EPR imaging for, 297–298

P

P-frame, video, 1387
p–n photodiodes, 1172
Packet communications, 151, 152, 1382–1385
Packetization delay, compression, 152
Packetized elementary streams (PES), 1382–1385
Paints
infrared light and, 668–672
Pressure sensitive paint, 867–868
Pair production, 256, 260
PAL standard, 146–149, 382, 1050, 1371–1373
PALplus television systems, 1392
Pan, 1050
Panavision 35, 1050
Panchromatic film, 1050
Panchromatic imaging, 55, 61, 85
Papers, photographic, 141, 605–606, 1209–1211
Parallacticscope, 1338
Parallax, 1050, 1330
Parallax barrier displays, 1337
Parallax, motion, 1328
Parallel viewing, 1336
Parallelism, functional, 562, 563
Parallelism, spatial, 562, 564
Paraxial marginal ray (PMR), 1080, 1083, 1085
Paraxial optics, 1073–74
Paraxial pupil ray (PPR), 1080, 1083, 1085
Paraxial ray tracing, 1074–1075, 1078, 1083, 1085
Paraxial rays, 97

Parietal lobe, 569
Parseval's theorem, 1105
Particle accelerators, 970
Particle beam measurement, 975
Particle detector imaging, 1154–1169
Particle form factor, 252
Particle image velocimetry (PIV), 391, 413–416
Particle polarization, 252
Particle sizing, in holography, 510
Passband sampling, 50, 56–61, 65
Passive matrix liquid crystal displays, 956
Pathogen dispersal, geologic imaging, 656–659
Pattern classification, 358–371
Pattern recognition, 350–353, 371–373
  human vision and, 522, 559, 565
  in search and retrieval systems, 625, 633
Pattern scanning, 573–574
Pattern spectrum, 442
Pattern spectrum density (PSD), 442
Pauli exclusion, in silver halide, 1272
Pediatric medicine, force imaging, 422
PEDOT polymer, 819–820, 823
Peel apart films, 828
Pendellousung effect, 282, 286
Penetration depth, in scanning acoustic microscopy (SAM), 1129–1130
Perceptron, pattern recognition using, 371–373
Perceptual redundancy and compression, 151
Perfect reflecting diffuser, 528
Perforations, film, 1025–1026, 1045, 1050
Period of wave, 212
Periodic functions, 1102–1105
Permeability, ground penetrating radar, 467
Permittivity, 212, 225, 466, 467
Persistence of vision, 1021–1022, 1050
Perspecta sound, 1050
Perspective, 1328, 1330, 1345–1347
Persulfate bleaches, 139
Perturbing spheres, in RF magnetic field mapping, 1223–1224
Perylenes, 1179, 1181
Phase, 241
Phase contrast imaging, in scanning acoustic microscopy (SAM), 1230
Phase contrast microscope, 265–266, 1106, 1116, 1128–1130
Phase object approximation, 278
Phase of wave, 213
Phase retarders, 233
Phase shift, 238–239, 1094, 1130
Phase transfer function (PTF), 1094
Phase velocity, 213, 228
Phasor diagrams, 241
Phasors, 246
Phenols, 133
Phenylenediamine (PPD) developer, color photography, 130
Phosphor primaries, 104
Phosphor screen, CRT, 31, 44–46, 173, 380
PhosPhor thermography, 864–867
Phosphoresence, 255
Photo-induced discharge curve (PIDC), 301–302, 318
Photobook, 623, 627
Photoconductivity, 1169–83
Photoconductors, 300–302, 310–312, 1169–83, 1190, 1200–05
Photodetectors, 49–61, 68, 70, 75, 80, 1087, 1090, 1172–1174, 1190–1191, 1198–1201
  quantum well infrared photodetectors (QWIP), 807
Photodetectors, 1183–1208
Photodiodes, 1172–1174, 1190–1191, 1198–1201
Photodynamic therapy (PDT), in endoscopy, 339
Photoelastic materials, force imaging, 420
Photoelectric detectors, 1169
Photoelectric effect, 211, 255–256
Photoemission, 1169–1170
Photoemulsion microcrystals, SIMS analysis, 484–486
Photofinish photography, 500–501
Photogrammetry, 737–739, 1327
Photographic color display technology, 1208–1222
Photography, 456, 1344–58
  art conservation and analysis using, 661–682
  color (*See* Color photography)
  digital, 141–142
  electro- (*See* Electrophotography)
  in forensic and criminology research, 709–714
  high-speed imaging (*See* High-speed photographic imaging)
  instant (*See* Instant photography)
  in medical imaging, 754
  motion picture (*See* Motion pictures)
  mug shots, 715–716
  in overhead surveillance systems, 773–802
  photofinish photography, 500–501
  process screen photography, 1051
  quality metrics and, 598–616
  in search and retrieval systems, 616–637
  silver halide detectors and, 1259–1309
  still, 1344–1358
  stroboscopic, 493–494
  surveillance imaging using, 714–715
  synchroballistic photography, 500–501
Photointerpretation, 774
Photolithography, 383, 384
Photoluminescence, 255
Photometer, 185–186, 1050
Photomicrography, 855, 1106, 1124, 1137–1138
Photomultiplier, 1062, 1183–1208
Photomultiplier tubes (PMT), 873
Photon absorbers, in infrared imaging, 806
Photon counting, lidar, 873–874
Photon flux, 215
Photons, 211, 214–215, 223
Photopic luminous efficiency function, 555
Photopic vision, 515
Photopigments, 101, 563
Photopolymers, holography, 509
Photoreceptors, 49, 50, 58, 65, 71, 72, 513–517, 549–552, 558–562, 746–747, 1178–1183
  in copiers, 1175–1183
Photoreconnaissance, 774
Photostat, 299
Photothermographic imaging, 851–853
Phototransistors, 1172–1174
Photovoltaics, 1190, 1204
Phthalocyanines, 1179, 1181
Physical vapor deposition (PVD), 383
PicHunter, 632
Pickup coils, in RF magnetic field mapping, 1224
Pictorial interface, in search and retrieval systems, 618
Pictrography 3000/4000, 852–853
Pictrostat 300, 852–853
Pictrostat Digital 400, 852–853
Piezoelectric effect, 420
Piezoelectric sensors, 423–424
Piezoelectric transducers, 1233, 1234
Pigment epithelium, 513
Pigments, infrared light, 668–672
Pin photodiodes, 1173
Pin registration, 1050
Pin scorotron, in electrophotography, 303

Pincushion distortion, 42–43, 93
Pinhole camera, 1072–1073
Piston transducers, 8–9, 1424–26
Pitch, 1050
Pitch, film, 1026
Pivaloylacetanilides, 134–135
Pixel detectors, particle detector imaging, 1167–1168
Pixel independence, cathode ray tube (CRT), 176, 179–181
Pixels, 1050
  in cathode ray tube (CRT), 173
  feature extraction using, 357
  feature measurement and, 343–344
  image processing and, 575–578
PixTech, 376
Planar Doppler velocimetry (PDV), flow imaging, 415–416
Planar imaging, flow imaging, 390–391, 397
Planar laser-induced fluorescence (PLIF), 391, 408–409, 411–416, 861–864
Planar neutron radiography, 1067
Planar sources, continuous, 7–8
Planck distribution, 804
Planck, Max, 211
Planck's constant, 287, 288, 394, 1184, 1273, 1476
Planck's equation, 525, 782
Planck's function, 759
Planck's law, 683
Planckian radiation (*See also* Blackbody radiation), 525
Plane of incidence, 233, 234
Plane of vibration, 231
Plasma displays, display characterization, 185
Plasma enhanced chemical vapor deposition (PECVD), 384
Plasma frequency, 227, 231
Plasma sheet and cusp, ENA imaging, 1010–1012
Plasma wave generators, terahertz electric field imaging, 1403
Plasmasphere, extreme ultraviolet imaging (EUV), 1005–1006
Plasmatrope, 1022
Plateau, Joseph A., 1022
Platinum silicide Schottky barrier arrays, 1201
PMOS, scanning capacitance microscope (SCM), 22
p–n junctions, scanning capacitance microscope (SCM) analysis, 22
Pockels cells, high-speed photography, 492–493
Pockels effect, 233, 1394, 1397
Pocket Camera instant films, 847
Podobarograph, force imaging, 421
Point of denotation (POD), in search and retrieval systems, 621–622, 630
Point of subjective equality (PSE), 609
Point spread function (PSF), 55, 1082, 1085–1092, 1097, 1098, 1099
  flow imaging and, 399, 402
  human vision and, 542, 543
  image processing and, 596
  in magnetic resonance imaging (MRI), 987
  in medical imaging, 750
  in overhead surveillance systems, 784
  telescopes, 688
  in ultrasonography, 1421–1423
Poisson probability, 53
Poisson's equation, 26–27
Polachrome, 848
Polacolor, 834, 840, 843–844
Polarization, 212, 223, 226, 227, 231–233, 235, 241, 250, 251, 252, 255, 257
  art conservation and analysis using, 666–668
  in forensic and criminology research, 719
  in ground penetrating radar, 468
  in high-speed photography, 492–493
  in holography, 509
  in human vision, 550–551
  in lidar, 873
  in magnetic resonance imaging (MRI), 978, 996
  in meteorological research, 772
  in microscopy and, 1116, 1131–1132, 1134
  modifying materials vs., 233
  in radar, 1453, 1468
  reflectance and, 237
  in terahertz electric field imaging, 1397
  in three-dimensional imaging, 1330, 1331
Polarization and Directionality of Earth's Reflectance (POLDER), 772
Polarization angle, 237
Polarization diversity radar, 772, 1468–69
Polarization rotators, holography, 509
Polarizing filter, 1050
Polaroid, 127, 829, 831–833, 844–849, 1331
Polaroid sheet, 233
Polavision, 848
Polyaromatic hydrocarbons, 1179
Polyester film, 1023, 1039, 1050
Polyethylene terephthalate films, 1023
Polymer degradation, EPR imaging, 296
Polymer films, gravure printing, 461
Polymer light emitting diodes (PLED), 817, 820–822
Polymer light emitting logo, 819–820
Polymeric research, scanning electrochemical microscopy (SECM) for, 1255–1256
Polyvinylcarbazole (PVK), 301, 823–825, 1178
Population inversion, 223
Position of objects, feature measurement, 345–347
Positive film, 1051
Positron emission tomography (PET), 210, 220, 743, 1407, 1324–1326
Postdevelopment processes, silver halide, 1303
Postproduction phase, in motion pictures, 1034, 1051
Potential well (POW) Scorotron, in electrophotography, 303
POV Ray software, 707
Powder cloud development, in electrophotography, 312
Power amplification, radar and over-the-horizon (OTH) radar, 1151
Power consumption, in field emission displays (FED), 388
Power K, 1078
Power law, human vision, 747
Power spectral density (PSD), 52, 53, 61–75, 80
Power, optical, 1074
Poynting vector, 213
PPP NEt3 polymer, 820–822
Prandtl number, flow imaging, 404
Precipitation, in silver halide, 1266
Predictive coding (DPCM), compression, 153, 155
Preproduction, in motion pictures, 1031–1032
Presbyopia, 540
Pressure effects
  flow imaging and, 411–416
  force imaging and, 420, 421, 424–426
  planar laser-induced fluorescence (PLIF) in, 863
  pressure sensitive paint, 867–868
Pressure garments, force imaging, 420
Pressure sensitive paint, 867–868
Primary colors, 102, 104, 126–127, 358, 531–534, 1051
Principal maxima, 245

Principal planes, 1078
Principle of superposition, 239–240
Print film, 1024, 1051
Printed edge numbers, 1051
Printer lights, 1051
Printers, 300
  cyclic color copier/printers, 328
  direct thermal printers, 196
  display characterization in, 183
  dots per inch (DPI) in, 602
  drive mechanisms in, 193
  dye sublimation, 189–190, 194–195
  in dye transfer printing, 188–197
  electrosensitive transfer printers, 195
  image-on-image or REaD color printing, 328–329
  laser printers, 195–196, 302, 306–310
  light sensitive microcapsule printers, 195
  micro dry process printers, 195
  in motion pictures, 1022
  quality metrics and, 598–616
  REaD color printing, 328–329
  tandem color printers, 328
  thermal transfer printers, 189
  wax melt, 190–191, 195
  wet gate, 1055
Printing, 454–463, 1051
Printing press, 455
Prisms, 495–496, 1073, 1106
Prisoner's problem, 160
Process screen photography, 1051
Processing, 1051
Processing chemicals (*See also* Developers), in holography, 509
Processing of images (*See* Image processing)
Product theorem, 4
Production phase, in motion pictures, 1032
Production supervisor, 1051
Progressive scanning, 146–147
Projection, image formation, 571–574
Projection film, 1023
Projection matrix, in tomography, 1407
Projection speed, 1051
Projection theorem, in tomography, 1406
Projectors, 184–185, 1022, 1035–1038, 1330
Prontor shutters, 1351
Propagation of waves, 220–231
  in ground penetrating radar, 466, 467–469
  in radar and over-the-horizon (OTH) radar, 1147, 1453–1454
Propagative speed, in ultrasonography, 1415
Proportionality, Grassmann's, 531
Proportioned bandwidth, television, 1367
Protanopia, 522–523
Protein Data Bank (PDB), 699
Proteins (*See* Biochemistry)
Proton/electron auroras, far ultraviolet imaging, 1016–1020
Proximity sensors, for lightning locators, 907
Pseudocolor display, 120–121
Pseudostereo imaging, 1330
Psychological quality metrics, 614–615
Psychometrics in quality metrics, 609–610
Psychophysical quality metrics, 607, 614–615
Ptychography, 262
Pulfrich technique, in three-dimensional imaging, 1333–1334
Pull down claw, 1027, 1051
Pulse code modulation (PCM), 67, 150
Pulse echo imaging, in ultrasonography, 1412
Pulse wave mode, in scanning acoustic microscopy (SAM), 1232–1234
Pulse width modulation (PWM), 383
Pulsed EPR imaging, 293
Pulsed repetition frequency (PRF), 872, 1452
Pump, laser, 223, 391
Punch-through bias, particle detector imaging, 1166
Pupil, 112–113, 514, 519, 540–543, 547, 553–554, 558
Pupil functions, 1086–1087
Pupil rays, 1081
Purple Crow lidar, 871
Push-broom scanners, 806
PUSH three-dimensional imaging, 1334
Push processing, 1051
Pyrazolin, 135
Pyrazolinone, 135
Pyrazolotriazole, 133, 135–136
Pythagoras' theory, 1094

Q

Q switching, 391–392, 508
QBIC, 618–630
Quad tree, in search and retrieval systems, 629
Quad tree method, thresholding and segmentation, 645–646
Quadrature amplitude modulation (QAM), 149
Quadrature mirror filter (QMF), 622
Quadrature modulation, 1365–1366
Quality metrics and figures of merit, 598–616, 1081–1085, 1357–1358
  human vision and, 543–544
  in medical imaging, 748–754
  in overhead surveillance systems, 789
  PSF and, 1089–90
  in tomography, 1409–1410
  in ultrasonography, 1420–1424
Quantitative imaging, flow imaging, 391, 397
Quantitative SPECT, 1322–1324
Quantitative structure determination, in transmission electron microscopes (TEM), 273–274
Quantization, 61, 64, 211
  compression, 154
  digital watermarking and, 150
  electron paramagnetic resonance (EPR) imaging for, 287
  flow imaging and, 402
  liquid crystal displays (LCDs), 184
  television, 1375
  video, 1388
Quantization index modulation (QIM), digital watermarking, 170–171
Quantization noise, 62, 63, 64, 68, 69, 80
Quantum detection efficiency/quantum efficiency, 691, 748, 1190
Quantum electrodynamics (QED), 256, 259
Quantum mechanical cross section, 255–256
Quantum nature, 214–218
Quantum sensitivity, in silver halide, 1282, 1284–1285
Quantum step equivalence (QSE), 786
Quantum terahertz biocavity spectroscopy, 1403
Quantum theory, 211, 214–218, 253, 1170
Quantum well arrays, 1205
Quantum well infrared photodetector (QWIP), 807, 1190, 1205
Quantum yield, 255, 862
Quarter wave plates, 233
Quasi 3D, scanning capacitance microscope (SCM), 25–26

Query by color, in search and retrieval systems, 621–622
Query by example, in search and retrieval systems, 620–621, 624, 628, 630
Query by sample, in search and retrieval systems, 624
Query by sketch, in search and retrieval systems, 618, 627, 630
Quinone, instant photography, 837–839
Quinonedimine (QDI) developer, color photography, 130

R

R 190 spool, 1051
R 90 spool, 1051
Radar, 1450–1474
  airborne, 1471–1473
  bistatic, 772
  cathode ray tube (CRT) in, 47
  coordinate registration, 1144
  Doppler (*See* Doppler radar)
  Doppler shift, 1145
  dwells, 1146
  equatorial anomaly, 1144
  geologic imaging and, 648
  ground penetrating, 463–476
  group range, 1144
  lidar vs., 869–870
  magnetic storms vs., 1145
  measurements in, 764–765
  meteorology, 757–773
  microwave radar ducting, 1141
  mobile systems, 1471–73
  National Imagery Interpretability Rating Scale (NIIRS), 795–800
  over-the-horizon (OTH) radar, 1141–1153
  in overhead surveillance systems, 775–802
  polarization diversity radar, 772, 1468–1469
  range folding, 1145
  scattering, 1145–1146
  shortwave fadeout, 1145
  skip zone, 1141
  sporadic E, 1144
  spotlights, 1146
  storm tracking, 765–769
  surface wave radar, 1141
  synthetic aperture radar (SAR), 356, 789
  terminators, 1145
  traveling ionsopheric disturbance (TID), 1144
  weather radar, 1450–1474
  wind profiling radar, 1469–1471
Radar cross section (RCS), 1145
Radar reflectivity factor, 765, 1453, 1454–1458
RADARSAT, 649
Radial velocity, 764
Radiance, 61, 524–525, 530, 785, 968, 1081, 1092
Radiance field, 51–54
Radiant heat transfer, in electrophotography, 325
Radiated fields, lightning locators, 909–914
Radiation, lightning locators, 911–912
Radiation belts, energetic neutral atom (ENA) imaging, 1006–1010
Radiation damping, 226
Radiation oncology, in medical imaging, 744
Radiation pressure, in transmission electron microscopes (TEM), 214
Radiation zone, 220
Radiative lifetime, 254
Radiators, non-planar, impedance boundaries, 9
Radio astronomy, 210
Radio frequency (RF), 220, 909
Radio interferometry, lightning locators, 946–947
Radio waves, 218, 230, 242, 803
  astronomy science and, 682, 693
  Atacama Large Millimeter Array (ALMA), 693
  lightning locators, 890
  RF magnetic field mapping, 1223–27
Radiography, 1057–71, 1067
Radiometric IR imaging systems, 804
Radiometry, 524, 810
  Advanced Very High-Resolution Radiometer (AVHRR), 759, 760–761
  Along Track Scanning Radiometer (ATSR), 772
  Multiangle Imaging Spectroradiometer (MISR), 772
  in overhead surveillance systems, 783
Radiosity, 804
Radiotracers, SPECT imaging, 1310–1314
Radon transform, in tomography, 1404–1406
Raggedness, 604
Rainbow schlieren, flow imaging, 412
Ram Lak filters, in tomography, 1405
Raman scattering, 259–260, 391, 411, 874, 882–885
Ramsden disk, 1111, 1120
Ramsden eyepiece, 1120
Random dot autostereogram, in three-dimensional imaging, 1336
Range folding, 1145
Range normalized signal strength (RNSS), lightning locators, 937
Range of use magnification, 1121
Rangefinders, 1350–1351
Rank ordering, quality metrics, 607–608
Ranking operations, image processing, 578–580, 583
RasMol software, 707
Raster, 31, 35, 42–43, 1051
Raster effect, 72
Raster3D, 707
Rate conversion, television, 1380
Rating techniques, quality metrics, 608–609
Raw stock, 1051
Ray tracing, 1076
  finite, 1083
  paraxial, 1074–1075, 1078, 1083, 1085
Rayleigh criterion, 245–246, 249
Rayleigh function, 94
Rayleigh–Gans–Debye scattering, 251–252
Rayleigh range,
  flow imaging and, 392
Rayleigh resolution limit, 1082, 1087
Rayleigh scattering, 252–253, 257–259, 391
  flow imaging and, 410, 411, 412, 416
  human vision and, 548
  lidar and, 874, 876–880
  flow imaging and, 415
Reach-through bias, particle detector imaging, 1166
REaD color printing, 328–329
Read noise, flow imaging, 395, 396
Reagents, instant photography, 828–829
Real time, 1051
Real time imaging, ultrasonography, 1428–1429
Receivers
  lidar and, 872–873
  radar and over-the-horizon (OTH) radar, 1147, 1151
Reception, in magnetic resonance imaging (MRI), 979–980
Receptive field, human vision, 563
Reciprocity failure, silver halide, 1286–1288
Reciprocity law, 1051
Recombination, in silver halide, 1275, 1281
Recommended practice, 1051

Reconstruction of image
algebraic reconstruction technique (ART), 1407–1408
in magnetic resonance imaging (MRI), 987–988
SPECT imaging, 1316–1321
in tomography, 1407–1408
wavelet transforms in, 1448
Reconstructive granulometries, 441–442
Recording systems, 504–505, 873
Rectangular piston, baffled, 8–9
Red green blue (RGB) system, 105–106, 531–534
in cathode ray tube (CRT), 173, 174
in color image processing, 101–102
in color photography, 123–142
in feature recognition and object classification in, 358
HSI conversion in, 112–114
image processing and, 578, 580
in motion pictures, 1052
in search and retrieval systems, 619
in television, 147–148
in thresholding and segmentation, 641
Redshift (*See also* Doppler shift), 686, 772
Reduction dye release, in instant photography, 839–840
Reduction printing, 1051
Reduction sensitization, silver halide, 1293
Redundancy, compression, 155
Reed Solomon coding, 1390
Reel 3D Enterprises, 1333
Reel band, 1051
Reference carriers, television, 1365
Reference white, 102
Reflectance, 50–53, 61, 68, 236–237, 527–530, 558, 610, 611, 803, 1072
in forensic and criminology research, 717
in scanning acoustic microscopy (SAM), 1235–1244
Reflected light microscopy, 1124–1127
Reflecting prisms, 1073
Reflection, 233–239, 267, 527–529, 1072, 1075
Bragg, 244
conducting surface, 239
ground penetrating radar and, 464–466, 468
holography in, 508
human vision and, 553–554, 561
in scanning acoustic microscopy (SAM), 1235–1244
in ultrasonography, 1415–1416
Reflection coefficient, in scanning acoustic microscopy (SAM), 1235–1244
Reflection grating, X-ray telescopes, 1505–06
Reflection holograms, 508
Reflective liquid crystal displays, 965–966
Reflectivity factor, radar, 1453, 1454–1458
Reflex cameras, 1345
Reflex radiography, art conservation and analysis using, 675–676
Refraction, 233–239, 1075, 1131–1132
dispersion vs., 234–235
ground penetrating radar and, 468
index of (*See* Index of refraction)
in motion pictures, 1051
Refractive error, 548
Refractive index (*See also* Index of refraction), 512, 1075, 1079, 1109, 1453
Region growing, thresholding and segmentation, 643–645
Regional gravity anomaly, gravity imaging, 448
Regions of influence, feature recognition and object classification, 370–371
Register tolerances, cathode ray tube (CRT), 37
Registration of image, in three-dimensional imaging, 1331
Regularization parameters, in tomography, 1407
Regularized least squares, in tomography, 1408
Rehabilitation, force imaging, 422, 424
Rehalogenating bleaches, 138
Reich, Theodore, 456
Relative aperture, 1081
Relative colorimetry, 533
Relative dielectric permittivity (RDP), 467
Relative edge response (RER), 800
Relative neighborhood graph (RNG), 370–371
Relative quantum efficiency (RQE), silver halide, 1296–1298
Relaxation parameters, in tomography, 1408
Release print, 1023, 1051
Rem jet backing, 1051
Rembrandt Intaglio Printing Co., 456
Remote sensing, 210, 356
geologic imaging and, 648–650
lidar and, 869, 870
in overhead surveillance systems, 780, 782
in three-dimensional imaging, 1327
Remote Sensing Act, 1992, 780
Rendering, 1051
Repetitive flash (stroboscopic) photography, 493–494
Reset noise, flow imaging, 396
Residual gravity anomaly, 448
Residual index, in biochemical research, 699
Resin coated (RC) paper, 1210–1211
Resolution, 60, 71, 75, 84, 1073, 1082, 1087, 1100–1103, 1357, 1358
Abbe numbers in, 1100–1102
amplitude, 151
astronomy science and, 686–688
cathode ray tube (CRT), 32, 33–34
charged particle optics, 86–100
in charged particle optics, 94–100
compression and, 151
detector, 790–793
electron paramagnetic resonance (EPR) imaging for, 292
in endoscopy, 334–336
flow imaging and, 398–405
in forensic and criminology research, 712–713
ground penetrating radar and, 469–471
High-Energy Neutral Atom Imager (HENA), 1010
human vision and, 561
liquid crystal displays (LCDs), 184, 858–859
in magnetic resonance imaging (MRI), 987
in magnetospheric imaging, 1010
in medical imaging, 750–752
microscopy and, 1106
in motion pictures, 1051
in overhead surveillance systems, 789–790, 792–794
photoconductors, 1181–82
Rose model and medical imaging, 753–754
scanning capacitance microscope (SCM), 19
in scanning acoustic microscopy (SAM), 1229, 1244–1245
spatial, 151
television, 1362
temporal, 151
in tomography, 1410

Resolution, (*continued*)
in transmission electron microscopes (TEM), 263, 266
in ultrasonography, 1421–1424
wavelet transforms in, 1448
X-ray fluorescence imaging and, 1482–1484
Resolution limit, 1073, 1087
Resolving power, 1051, 1358
Resonance, 226, 228–233, 287
Resonance curve, 226
Resonance lidar, 883–885
Resonant coils, RF magnetic field mapping, 1223–1227
Resonant fluorescence, 259
Resonant frequency, 228
Resonators, 223
Response time, in field emission displays (FED), 387
Responsitivity, 54–56
human vision and color vision, 529–531
photodetectors and, 1187
Restoration of images (*See* Image restoration)
Restoring force, 225
Reticulation, 1051
Retina, 65, 513–519, 522, 543, 547, 549, 552, 558–564, 746–747
Retinal image size, 1328
Retinex coding, 61, 65, 75–83
Retrace lines, television, 1362
Retrofocus, 1354
Return beam vidicon (RBV), in overhead surveillance systems, 779
Reversal film, 139, 1052
Reversal process, 1052
Reverse modeling, scanning capacitance microscope (SCM), 27
Reverse perspective, 1330
Reynolds number, flow imaging, 404, 410, 412
RF coils, in magnetic resonance imaging (MRI), 999
RF magnetic field mapping, 1223–27
RF spoiling, in magnetic resonance imaging (MRI), 992–993
Rheinberg illumination, 1128
Rhodopsin, 517, 561, 747
Ring current, energetic neutral atom (ENA) imaging, 1006–1010
Ring imaging, particle detector imaging, 1162
Ring imaging Cerenkov counters (RICH), 1162
Ring opening single electron transfer (ROSET) dye release, 840, 851
Ringing, 75
Rise time, in lightning locators, 913
Ritchey-Chretein mirrors, 783
Ritter von Stampfer, Simon, 1022
Robustness, 74
Rocking curve, 280, 282
Rods, 122, 513–519, 530–531, 551–554, 558, 560–561, 562, 746–747
Roget, Peter Mark, 1022
Roller charging, in electrophotography, 303
Ronalds, 299
Rontgen, Wilhelm C., 1475
Root mean square (rms) aberration, 1090
Root mean square (rms) granularity, 140
ROSAT telescopes, 1507
Rose model, in medical imaging, 753–754
Rotating drum and mirror cameras, 496–497
Rotating mirror framing cameras, 497–498
Rotating mirrors, in three-dimensional imaging, 1343
Rotating prism cameras, 495–496
Rotation, 167, 1052
Rotation, molecular, 216
Rotational frequency, in magnetic resonance imaging (MRI), 979
Rotational mapping, in RF magnetic field mapping, 1225
Rotogravure, 454–463
Rough cut, 1052
RS strings, in search and retrieval systems, 629
Run-length encoding, 516
Ruska, Ernst, 261
Rutherford scattering, particle detector imaging, 1157

S

Saddle coils, deflection yoke, 41–42
Safety acetate film, 1023, 1024, 1052
SAFIR lightning locators, 914, 945–950
Sampling, 56, 61
compression, 153
digital watermarking and, 150
human vision and, 552, 554, 560
lightning locators, 918–921
in magnetic resonance imaging (MRI), 986–987
passband, 50, 56–61, 65
thresholding and segmentation in, 640
Sampling lattice, 50, 54, 57, 58, 59, 65, 71
Sanyo three-dimensional display, 1340–41
Satellite imaging systems, 350, 356
Advanced Very High-Resolution Radiometer (AVHRR), 759, 760–761
Advanced Visible Infrared Imaging Spectrometer (AVIRIS), 650, 787
Along Track Scanning Radiometer (ATSR), 772
Applications Technology Satellite (ATS), 757
Array of Low Energy X Ray Imaging Sensors (ALEXIS), 905, 929
cloud classification, 761–764
Defense Meteorological Satellite Program (DMSP), 890–904, 929
Earth Observing System (EOS), 772
Earth Resources Technology Satellite (ERTS), 778–779
Fast On Orbit Recording of Transient Events (FORTE), 890–904, 929
feature recognition and object classification in, 352–353
geologic imaging and, 647–661
Geostationary Meteorological Satellite (GMS), 760
Geostationary Operational Environmental Satellite (GOES), 760, 778
gravity imaging and, 444–454
IKONOS satellite, 780
image processing and, 589
IMAGE, 1018
imaging satellite elevation angle (ISEA) in, 791
in overhead surveillance systems, 773–802
infrared in, 758
Landsat, 778, 787
Lightning Imaging Sensor (LIS), 890–904, 929, 932–935
lightning locators, 890, 928–935
magnetospheric imaging, 1002–1021
measurement in, 758–759
meteorology, 757–773
Meteorological Satellite (METEOSAT), 760
Multiangle Imaging Spectroradiometer (MISR), 772
multiangle viewing instruments, 772
multispectral image processing and, 101
NIMBUS, 778
oceanography, 760

Optical Transient Detector (OTD), 890–904, 929–932
Polarization and Directionality of Earth's Reflectance (POLDER), 772
Systeme Probatoire d'Observation de la Terre (SPOT), 779–780
Television and Infrared Observational Satellite (TIROS), 757, 777
Thematic Mapper, 648, 653, 654, 657, 779
time delay integration (TDI), 1018
Tropical Rainfall Measuring Mission (TRMM), 660, 771–772, 890–904, 929, 932–935, 1473
X-ray Evolving Universe Satellite, 1509
Saturation, 103, 111, 117, 119, 578, 580, 618, 994–995, 1043, 1052
Sawtooth irradiance distribution, 1092
Scaling, digital watermarking, 167
Scaling factors, television, 1367
Scaling methods, quality metrics, 607–608
Scalograms, 1444
Scan a Graver, 456, 461
Scanned probe microscopy (SPM), 1248
Scanners, 574
art conservation and analysis using, 663
calibration of, 603
in forensic and criminology research, 709
in infrared imaging, 804, 805–806
in meteorological research, 769
multispectral, 360
in overhead surveillance systems, 779
push broom type, 806
quality metrics and, 602–603
whisk broom type, 806
Scanning, 1072
high-definition TV (HDTV), 147
image formation in, 571, 573–574
interlaced, 146–147
k space, 574
odd and even field, 146
pattern, 573–574
progressive, 146–147
television, 146–148, 1359, 1362, 1366–1367
ultrasonography, 1413–1415
X-ray fluorescence imaging and, 1478–1479, 1482
Scanning acoustic microscopy (SAM), 1128–1148
Scanning capacitance microscope (SCM), 16–31
Scanning capacitance spectroscopy, 21
Scanning electrochemical microscopy (SECM), 1248–1259, 1248
Scanning electron microscope (SEM), 23, 87–88, 262, 274–278, 477, 1243
Scanning evanescent microwave microscope (SEMM), 28
Scanning ion microscopy (SIM), 477
Scanning Kelvin probe microscope (SKPM), 16, 28
Scanning lines, television, 1359
Scanning microwave microscope (SMWM), 16, 28
Scanning transmission electron microscope (STEM), 87, 93, 262, 276–278
Scanning transmission ion microscope (STIM), 479
Scattering, 51, 242, 244, 249–253, 256–260, 282, 283, 285, 286, 391, 1072
in biochemical research, 698
flow imaging and, 397–398, 410, 411, 415–416
ground penetrating radar and, 471
holography in, 504
human vision and, 548, 549
image formation in, 571
lidar and, 874–875
multiple coulombic scattering (MCS), 1157
in overhead surveillance systems, 785
particle detector imaging and, 1157
radar, 1145–1146, 1451
Rutherford scattering, 1157
in scanning acoustic microscopy (SAM), 1243
SPECT imaging, 1323–1324
in transmission electron microscopes (TEM), 269
in ultrasonography, 1415–1416
Scattering angle, 250
Scattering plane, 251
Scene, 1052
Scherzer defocus, 272
Schlieren images, 405–408, 412, 501–504
Schottky diode, 19, 1172–1174, 1190, 1201
Schottky thermal field emitter, in charged particle optics, 90
Schulze, Johnann Heinrich, 1345
Science (*See also* Astronomy; Biochemistry; Medicine and medical research), 742
Scientific Working Group on Imaging Technologies (SWGIT), 719, 741
Scintillation, 688, 1158, 1313–1314
Scintillation cameras, SPECT imaging, 1313–1314
Scintillation tracking devices, particle detector imaging, 1158–1168
Scintillator detectors, in neutron imaging, 1062–1064
Scintillators, in radiographic imaging, 1067, 1068
Scope, 1052
Scorotron charging, in electrophotography, 303
Scorotrons, 1176
Scotopic (rod) vision, human, 122, 515, 747
Scotopic luminous efficiency function, 555
Scrambling, 60
Screened images, 455
Screw threads, microscopy, 1116
Scrim, 1052
Script, 1052
Sealing glass (frit), in field emission displays (FED), 387
Seaphone three-dimensional display, 1339–1340
Search and retrieval systems, 616–637
Search engines, in search and retrieval systems, 633
SECAM standards, 146–148, 1052, 1367–1371
Second order radiative process, 256
Secondary electron (SE) imaging, in scanning electron microscopes (SEM), 275–276
Secondary ion mass spectroscopy (SIMS), 477–491
Secondary maxima, 245
Security, digital watermarking, 159–161
Segmentation, 615
in color image processing, 119–120
human vision and, 568–569
image processing and, 587
in search and retrieval systems, 622, 625
Seidel polynomials, 542
Selection rules, atomic, 253
Selectivity, human vision, 565–567
Selenium, 300, 301, 1170
Sellers, Coleman, 1022

Semiconductor detectors, 1064–65, 1163–1165, 1168
Semiconductors, 22–23, 1183–1208
Semigloss surfaces, 528
Sensitivity, 1052
  astronomy science and, 688–690
  in magnetic resonance imaging (MRI), 983
  in radar and over-the-horizon (OTH) radar, 1147
  silver halide, 1261, 1282, 1284–1285
  in SQUID sensors, 14
Sensitivity or speed of film, 124, 139
Sensitization, in photographic color display technology, 1215–1216, 1288–1293
Sensitizers, color photography, 124–125
Sensitometer, 1052
Sensitometry, silver halide, 1262
Sensors, 101, 356
  active pixel sensors (APS), 1199–1200
  capacitance, 17–18
  CMOS image sensors, 1199–1200
  force imaging and, 420, 422–424
  monochrome image processing and, 100
  in overhead surveillance systems, 787–789
  scanning capacitance microscope (SCM), 17–18
Separation light, 1052
Separation masters, 1052
Sequence, 1052
Sequential color TV, 1365
Sequential frequency modulation, 1367–1368
Series expansion methods, in tomography, 1406–1409
Serrations, television, 1360
Set, 1052
Set theory, 430
Setup level, television, 1362
70 mm film, 1025
Sferics, 890
Shading, 3, 58, 65, 82, 1328
Shadow mask, 31, 36–38, 44, 47, 173, 825–826
Shadowgraphs, 405–408, 501, 743
Shadowing, 1328
Shadows, 51, 53, 58, 75, 82
Shallow electron trapping (SET) dopants, 1215
Shannon's law, 49, 63
Shannon's theory of information, 99
Shape analysis-based search and retrieval systems, 625–628
Shape factor, 357–358
Shape of objects, in feature measurement, 347–350
Sharpness, 71, 75, 81, 140, 1052, 1347, 1357
  in color photography, 137
  in high-speed photography, 492
  in image processing, 590–591
  quality metrics and, 598–616
  silver halide and, 1304
Sheet-fed gravure, 460
Shielding, ground penetrating radar, 468
Shore A scale, gravure printing, 459
Short, 1052
Shortwave broadcast, radar and over-the-horizon (OTH) radar, 1142
Shortwave fadeout, 1145
Shot, 1052
Shot noise, flow imaging, 395
Show Scan, 1031
Shutter, 492–493, 1027–1028, 1036, 1052, 1351–1352
Shuttle Imaging Radar, 649
Shuttle Radar Topographic Mapping Mission (SRTM), 660
Sibilance, 1052
SiC, scanning capacitance microscope (SCM) analysis, 22
Sidebands, television, 1362, 1366, 1389
Signal coding (*See also* Encoding), 49–51, 61–62, 84
Signal detection, particle detector imaging, 1168
Signal-induced noise (SIN), lidar, 873
Signal levels, television, 1361–1362
Signal processing,
  in digital watermarking, 161, 171
  in human vision and color vision, 516–518
  in radar and over-the-horizon (OTH) radar, 1147, 1151–1152
signal propagation model, lightning locators, 937
signal to noise ratio, 50
signal to noise ratio (SNR), 60, 64–66, 74, 81, 84
  in charged particle optics, 99
  electron paramagnetic resonance (EPR) imaging for, 289
  in flow imaging, 393, 394–397
  in magnetic resonance imaging (MRI), 987, 996
  in medical imaging, 749
  in overhead surveillance systems, 794–795
  in radar and over-the-horizon (OTH) radar, 1147, 1150
  silver halide, 1303
  in sound systems, 1388
  in tomography, 1410
  in X-ray telescopes, 1497
Signal transduction, in scanning electrochemical microscopy (SECM), 1249–1253
Silicon dioxide, scanning capacitance microscope (SCM) analysis, 23
Silicon drift detectors, particle detector imaging, 1167
Silicon nitride, scanning capacitance microscope (SCM) analysis, 23
Silicon photoconductors, 1204–1205
Silicon technology, 384–385
Silicon transistors, scanning capacitance microscope (SCM) analysis, 23
Silicon Video Corp. (SVC), 377
Silver assisted cleavage dye release, instant photography, 840, 851
Silver clusters and development, silver halide, 1280–1281
Silver Dye Bleach, 127
Silver halide, 140, 1259–1309, 1345, 1356–1357
  art conservation and analysis using, 661–662
  in color photography, 123, 125–126, 129–130
  detectors using, 1259–1309
  in holography, 509
  in instant photography, 827, 830–833
  in motion pictures, 1052
  in photographic color display technology, 1208–1222
  secondary ion mass spectroscopy (SIMS) analysis of, 484–486
Silver nitrate, 1345
Silver oxide, 381
SIMION, 482
Simulated images, in transmission electron microscopes (TEM), 271–273
Simulations, 769–771, 1282, 1327
Simultaneous autoregressive (SAR) model, 623–624
Single electron transfer dye release, 840, 851
Single frame exposure, 1052
Single hop mode, radar and over-the-horizon (OTH) radar, 1141
Single lens reflex cameras, 1349–1350
Single perforation film, 1052
Single photon emission computed tomography (SPECT), 743, 1310–1327
Single pixel image processing, 575–576

Single poSitive imaging, gravure printing, 461
Single station lightning locators, 907–908
Single system sound, 1052
16 mm film, 1024–1025, 1052
Size distributions, 442
Size of image, 1074–77
Size of objects, 343–344, 686–688
Sketch interface in search and retrieval systems, 618
Skin depth, 229, 230
Skip frame, 1052
Skip zone, 1141
Skunk Works, 775, 776
Sky waves, 912
Slew rate, cathode ray tube (CRT), 179–180
Slides, microscope, 1124
Slitting, 1052
Slow motion, 1052
Smith, Willoughby, 1170
Smoothing, 577, 580–583, 593, 598–616, 755–756
Snell's law, 234, 238, 468, 1076
Snellen chart, human vision, 747
Sobel operator, image processing, 582
Society for Information Display (SID), 818
Society of Motion Picture and Television Engineers (SMPTE), 102, 1052, 1374
Sodium arc lamps, 222
Sodium double, 222
Sodium lidar, 884–885
Soft, 1053
Soft light, 1053
Soil moisture mapping, geologic imaging, 656–659
Solar wind, magnetospheric imaging, 1002–1021
Solid state detectors (SSD), 1007, 1477
SOLLO lightning locators, 905, 908, 909
Sound drum, 1053
Sound editing, in motion pictures, 1035
Sound effects, 1053
Sound gate, 1053
Sound head, 1037–38, 1053
Sound navigation and ranging (SONAR), 1412
Sound pressure level (SPL), 3
Sound recorder, 1053
Sound speed, in motion pictures, 1028–1029
Sound sprocket, 1053
Sound systems
  in motion pictures, 1031, 1033, 1037
  in television, 1362, 1365, 1388–1389, 1388
Sound track, 1053
Sounding, radar and over-the-horizon (OTH) radar, 1142
Source points, 251
Space-based imaging technology, astronomy science, 691
Space exploration, magnetospheric imaging, 1002–1021
Space Infrared Telescope Facility (SIRTF), 690, 691–692
spacers, in field emission displays (FED), 381–382, 386
Spallation Neutron Source (SNS), 1057
Sparrow resolution limit, 1087
Spatial domain, image processing, 577, 594
Spatial filters, 509, 1100
Spatial frequency, 248, 559–562, 565–566, 1098, 1103
Spatial frequency response (SFR), 50, 56–61, 62, 63, 66, 68, 70–74, 72, 79, 80
Spatial homogeneity, in cathode ray tube (CRT), 181–182
Spatial parallelism, 562, 564
Spatial relationship, in search and retrieval systems, 628–630
Spatial resolution, 151
  in medical imaging, 750–752
  in microscopy, 1136
  in overhead surveillance systems, 789–790
  in scanning capacitance microscope (SCM), 19
  in X-ray fluorescence imaging, 1482–84
Spatial response (SR), 50, 54–56, 68, 78, 79, 80
Spatial uniformity, in ultrasonography, 1424
Spatial visual processing, human vision, 558–570
Special effect, 1034, 1053
Specimens, microscopy, 1108–1109
Speckle, in ultrasonography, 1420–1421
Spectra, microscopy, 1108
Spectra instant film, 847
Spectral filters, 55, 872
Spectral imaging, feature recognition and object classification, 356–357
Spectral lines, velocity analysis using, 685–686
Spectral luminosity function (SLF), 554–555
Spectral power density/distribution (SPD), 100, 524–527, 618
Spectral purity, radar and over-the-horizon (OTH) radar, 1147
Spectral radiant exitance, 222
Spectral radiant flux density, 222
Spectral response, in feature recognition and object classification, 358
Spectral sensitization, silver halide, 1294–1299
Spectrometer, 239, 244, 481–482, 650, 787, 970, 1403
Spectroradiometer, 185–186, 524
Spectroscopy, 571
  Constellation X mission, 1509
  electron paramagnetic resonance (EPR) imaging for, 289
  in endoscopy, 338
  high-resolution secondary ion mass, 477–491
  quantum terahertz biocavity spectroscopy, 1403
  terahertz electric field imaging and, 1393–1404
Spectrum, 57, 1053
Specular surfaces, 234
Speed, 1053
Speed of film, 124, 139, 1023
Speed of light, 224–225
Spherical aberration, 92, 98, 1088–1089, 1117, 1123
Spherical waves, 214
Spherics, 890
Spider stop, microscopy, 1128
Spin angular momentum, 217
Spin density, in magnetic resonance imaging (MRI), 983–984
Spin echo
  in magnetic resonance imaging (MRI), 981, 992
  in RF magnetic field mapping, 1125–1126
Spin states, atomic, 217
Spindt emitters, 385
Spindt technique, 385
Splice, 1053
Splicer, 1053
Splicing tape, 1053
Spline fits, thresholding and segmentation, 644–645
Split and merge techniques, thresholding and segmentation, 645–646
Splitters, holography, 509–509
Spontaneous emission, 253–254
Spontaneous Raman scattering, 411
Spool, 1053
Sporadic E, 1144
Sports medicine, force imaging, 422
SPOT high-resolution visible (HRV) imaging systems, 649, 655

Spotlight, 1053
Spotlights, radar, 1146
Sprockets, in projectors, 1036, 1053
Sputtering, 383, 478
Square root integral (SQRI), quality metrics, 606
Squarilium, 1179
SQUID sensors, analog and digital, 9–15
Stabilization of dyes, 831, 841–842
Staircase patterns, 75
Standard definition TV (SDTV), compression, 157
Standard illuminants, CIE, 103–104
Standard Observer, 618
Standards converters, television, 1373
Stanford Research Institute, 375
Stanford, Leland, 1022
Static electricity, 1053
Stationarity, 1085–1086
Statistical redundancy and compression, 150–156
Steadicam, 1031
Steering, 4–5
Stefan–Boltzmann law/constant, 222, 804
Steganography, digital watermarking and vs., 160
Step response function (SRF), 402
Stereo display technologies, 1327–1344
Stereo pairs, in three-dimensional imaging, 1329–1330
Stereo window, 1329
StereoGraphics three-dimensional imaging, 1333
StereoJet three-dimensional imaging, 1332
Stereolithography three-dimensional imaging, 1327
Stereomicroscope, 1106
Stereophonic, 1053
Stereoscopic vision, 566
Stiffness matrix, 627
Stiles–Crawford effect, 58, 513, 542, 552–554
Stiles–Holladay approximation, 548
Still photography, 491–494, 1344–58
Stimulated echo, in magnetic resonance imaging (MRI), 983
Stimulated emission, 223, 253, 254–255
Stock, 1053
Stop, 1053
Stop motion, 1053
Stops, 1354
Storage and retrieval systems
  art conservation and analysis using, 661–682
  in forensic and criminology research, 716–717
  Methodology for Art Reproduction in Color (MARC), 664
  in motion pictures, 1038–39
  Visual Arts System for Archiving and Retrieval of Images (VASARI), 663–664
Storage systems, secondary ion mass spectroscopy (SIMS), 482–484
Storm tracking, 655–656, 765–769
Storyboard, 1053
Stratospheric Observatory for Infrared Astronomy (SOFIA), 692–693
Streak cameras, 499–500
Strehl ratio, 543, 544, 1090
Strike filtering, gravity imaging, 452
Strip cameras, 500
Strip scintillation detectors, particle detector imaging, 1165–67
Stripe, magnetic, 1053
Stroboscopic photography, 492, 493–494
Structural dilation, 433
Structural erosion, 432–433
Subband/wavelet coding, compression, 154
Subcarriers, television, 1362, 1365
Subclustering, in feature recognition and object classification, 367–370
Subjective quality metrics, 602, 606–610
Sublimation printers, 189–190, 194–195
Subtraction, image processing, 590
Subtraction, Minkowski, 432, 612
Subtractive color, 833–841, 1053
Subtractive color matching, 102
Subtractive color mixing, 127–128, 139
Sulfite developer, color photography, 130
Sulfonamidonaphthol, 837, 838
Sulfonamidophenol, 838
Sulfoselenide, 301
Sulfur plus gold sensitization, silver halide, 1292–1293
Sulfur sensitization, silver halide, 1289–1292
Super xxx films, 1025
Super Panavision, 1053
Superadditivity, silver halide, 1303
Superconducting quantum interference devices (SQUIDs), 9–15, 976
Superconductors, 9–15, 484, 486–487, 1106
Superposition, 239–240
Superscope, 1053
Supersensitization, silver halide, 1298–1299
Supersonic flow, flow imaging, 409
Supertwisted nematic (STN) liquid crystal displays, 961–962
Surface acoustic waves (SAW), in scanning acoustic microscopy (SAM), 1236–1243
Surface stabilized ferroelectric LCD (SSFLC), 965
Surface wave radar, 1141
Surround function, 77, 79
Surround speakers, 1054
Surveillance et Alerte Foudre par Interferometrie Radioelectriquie (*See* SAFIR)
Surveillance imaging
  in forensic and criminology research, 709, 714–715
  overhead, 773–802
  radar and over-the-horizon (OTH) radar, 1141–1153
SVGA video, in field emission displays (FED), 382
Swan, J.W., 455
Sweetening, 1054
Swell, 1054
Swiss PdbViewer, 708
SX70 instant film, 844–847
Symmetry, in compression, 152
Sync pulse, 1054, 1360
Sync sound, in motion pictures, 1033
Synchroballistic photography, 500–501
Synchronization
  high-speed photography and, 493
  in motion pictures, 1054
  in television, 1360, 1375
Synchronizer, 1054
Synchrotron radiation (SR), 221, 1476
Synthetic aperture radar (SAR), 356, 648, 789
Systeme Probatoire d'Observation de la Terre (SPOT), 779–780

T

T-grain emulsion, 1054
T1/T2 relaxation, MRI, 983–984, 988–991
Tail ends, 1054
Take, 1054
Tamper detection, digital watermarking, 159
Tandem color printing, 328
Tape splice, 1054
Tapetum, 513
Taylor procedure, 3
Taylor series, 227
Technicolor, 1024
Technirama, 1031, 1054

Techniscope, 1054
Telecine, 1054
Telemacro lens, 1354–1355
Telephotography, 59
Telephoto lens, 1347, 1354
Telescopes (*See also* Astronomy), 210, 1072
  Apollo Telescope Mount, 1507
  astronomy science and, 682–693
  Atacama Large Millimeter Array (ALMA), 693
  Chandra Observatory, 1508
  Constellation X Observatory, 693, 1509
  Einstein Observatory Telescope, 1507
  Giant Segmented Mirror Telescope, 693
  Kirkpatrick Baez telescopes, 1502–1503
  lidar and, 871–872
  limitations on, 688, 690–691
  liquid mirror telescopes (LMT), 872
  mirrors for, 691
  multilayer telescopes, 1503
  Next Generation Space Telescope (NGST), 693
  in overhead surveillance systems, 783
  ROSAT telescopes, 1507
  Space Infrared Telescope Facility (SIRTF), 690, 691–692
  Stratospheric Observatory for Infrared Astronomy (SOFIA), 692–693
  Terrestrial Planet Finder, 693
  thin mirror telescopes, 1501–1502
  TRACE telescopes, 1507–1508
  Very Large Array Radio Telescope, 693
  Wolter, 1499–1501
  X-ray Evolving Universe Satellite, 1509
  X-ray interferometric telescopes, 1503–1504
  X-ray telescope, 1495–1509
  XMM Newton telescope, 1508
Television (*See also* Motion pictures; Video), 59, 1021
  ATSC Digital Television Standard for, 1359, 1382–1389
  black-and-white, 1359
  broadcast transmission standards, 1359–1393
  cathode ray tube (CRT) using, 47
  chromaticity in, 148
  component systems in, 148–150
  compression in, 150–157
  digital watermarking and, 146–148
  digitized video and, 149–150
  high-definition (HDTV), 41, 42, 47, 147, 151, 153, 157, 1039, 1047, 1382, 1390
  image aspect ratio in, 147
  image intensity in, 147–148
  interlaced scanning in, 146–147
  luminance in, 148
  National Television System Committee (NTSC) standards for, 146–149
  National Television Systems Committee (NTSC), 1359–1393
  PAL standard for, 146–149, 1359–1393
  progressive scanning in, 146–147
  red green blue (RGB) system in, 147–148
  scanning in, 146–148
  SECAM standard for, 146–148, 1359–1393
  standard definition TV (SDTV), 157
  trichromatic color systems in, 147–148
Television and Infrared Observational Satellite (TIROS), 757, 777
Temperature
  photodetectors and, 1188–1190
  in scanning acoustic microscopy (SAM), 1230
Temperature calculation, in infrared imaging, 814–815
Temperature effects, flow imaging, 411–416
Temperature mapping, in infrared imaging, 812–815
Temperature measurement, planar laser-induced fluorescence (PLIF), 863
Temperature, color, 103, 525
Tempone, 289
Temporal dependencies, 180, 184
Temporal homogeneity in cathode ray tube (CRT), 182
Temporal lobe, 569
Temporal resolution, 151, 1424
Terahertz electric field imaging, 1393–1404
Terminator, radar, 1145
TERRA, 659
Terrain correction, gravity imaging, 448
Terrestrial Planet Finder, 693
Test patterns, quality metrics, 603
Tetramethyl ammonium hydroxide (TMAH), 384
Texas Instruments, 376–377
Text tag information, in search and retrieval systems, 617
Textile presses, 456
Textural gradient, 1328
Texture, in search and retrieval systems, 622–625
Texture processing, image processing, 583–584
Theatres, 1038
Thematic Mapper, 648, 653, 654, 657, 779
Thermal emission, 1176, 1184–1187
Thermal field emitter (TFE), 90
Thermal head, in dye transfer printing, 191–193
Thermal imaging, 810–811
Thermal Infrared Multispectral Scanner (TIMS), 650
Thermal radiation, 356
Thermal signatures, 803
Thermal sources, 222
Thermal transfer process, 189, 853
Thermally assisted fluorescence (THAF), 863
Thermionic emission, 223
Thermofax, 299
Thermograms, 802–817
Thermographic imaging, 851–854
Thermography, 802–817, 864–867
Thermoplastics, 509
Thiazolidine, 840, 841
Thickness extinction contours, 282
Thin-film technology, 383–384
  in field emission displays (FED), 377, 379, 383–384
  in liquid crystal displays, 957
  in scanning acoustic microscopy (SAM) analysis for, 1228
Thin lens conjugate equation, 1078
Thin mirror telescopes, 1501–1502
Thin objects, in transmission electron microscopes (TEM), 270–271
Think Laboratories, 461
Thinker ImageBase, 617
Thiols, 135
Thiopyrilium, 1179
Thiosulfate bleaches, 139
35 mm film, 1022, 1024, 1054
Thomson scattering, 249–250, 256
Thread, 1054
Three-dimensional imaging, 1054, 1072, 1327–1344
  in biochemical research, 694–708
  Doppler radar and, 1465–1468
  flow imaging and, 416–417
  force imaging and, 424
  ground penetrating radar and, 472–475, 476
  human vision and, 566

Three-dimensional imaging, (*continued*)
in meteorological research, 772
in ultrasonography, 1433
Thresholding, 590, 584–589, 637–638
Thresholds, quality metrics, 609–610
Throw, 1054
Thunderstorm Sensor Series (TSS), 907, 922
Thunderstorms (*See* Lightning locators)
Thyristor flash systems, 1348–1349
Tidal correction, gravity imaging, 447
Tight wind, 1054
Tilting, 4–5
Time delay and integration (TDI), 785, 1018
Time domain waveform analysis, 912–914
Time–energy uncertainty principle, 259
Time lapse, 1054
Time of arrival (TOA) lightning locators, 906–907, 935, 941–945
Time of flight imaging, 989–991, 1015–1016
Time parallel techniques, in three-dimensional imaging, 1331
Time projection chamber (TPC), particle detector imaging, 1160
Time sequence maps, electroencephalogram (EEG), 201–204
Time slice imaging, ground penetrating radar, 472–475
Time Zero film, 846–847
Timing, 184, 1054
Timing layer, in instant films, 832
Titanyl phthalocyanine (TiOPc), 1180–1181
Todd AO, 1031, 1054
Toe, 1054
Tomography, 1404–1411
flow imaging and, 416
ground penetrating radar and, 475–476
image formation in, 571
low resolution electromagnetic tomography (LORETA), 204–208
in medical imaging, 743
in radiographic imaging, 1068
single photon emission computed tomography (SPECT), 1310–1327
terahertz electric field imaging and, 1399–1400
Tone, 598–616, 1054
Tone burst wave mode, in scanning acoustic microscopy (SAM), 1231, 1233
Toner, in electrophotography, 301, 312, 313–315, 325–329
Top hat transforms, 430
Topographic imaging technology, 199–201
TOPS software, 708
Toroidal coils, deflection yoke, 41–42
Total internal reflectance (TIR), 238–239
Total scattering cross section, 250
Tournachon, Gaspard Felix, 773
TRACE telescopes, 1507–1508
TRACKERR, 1158
Tracking, radar and over-the-horizon (OTH) radar, 1148, 1152
Trailer, 1055
Trajectories, particle detector imaging, 1157
Trajectory effects, gravity imaging, 444
Tranceivers
in magnetic resonance imaging (MRI), 999
in terahertz electric field imaging and, 1399–1400
Transducers, 1, 1418–1419, 1424–1429
Transfer function, 264, 575
Transfer process, in electrophotography, 322–324, 322
Transverse electromagnetic modes (TEM), 392
Transform coding, compression, 153–154
Transformation, compression, 153
Transfusion, in electrophotography, 322
Transistors, scanning capacitance microscope (SCM) analysis, 23
Transition, 1055
Transitions, atomic, 215
Translation invariant operators, 431–436
Transmission, 527–529, 548–550, 561, 1404
Transmission electron microscopes (TEM), 23, 87, 93, 262–274
Transmission grating, X-ray telescopes, 1505
Transmission holograms, 507–508
Transmission line model (TLM), 937
Transmittance, 51, 54, 58, 236–237, 527–529, 783, 803, 1072, 1095
Transmitters, lidar, 871–872
Transparency views, in three-dimensional imaging, 1333
Transverse chromatic aberration, 545
Transverse electric or magnetic waves, 235
Transverse magnification, 1076
Transverse viewing, in three-dimensional imaging, 1336
Transverse waves, 212
Trapping, scanning capacitance microscope (SCM) analysis, 23
Traps, silver halide, 1273–75
Traveling ionspheric disturbance (TID), 1144
Traveling matte, 1055
Trellis coding, 1390
TREMBLE lightning locators, 909
Triangle, 1055
Triangulation, 571, 572–573, 908
Triarylmethyl radicals, 289
Triboelectrification, 944
Trichromatic color theory, television, 147–148
Trichromatic color vision, 567
Trichromatic receptors, human vision and color vision, 519
Tricolor image processing systems, 101–102
Triiodide, 1213
Trims, 1055
Trinitron electron gun, 40
Triphenylamine, 1179
Tripods, 1030
Tristimulus values, 102, 148, 531–534, 537
Tritanopia, 522–523
Tropical Rainfall Measuring Mission (TRMM), 660, 771–772, 890–904, 929, 932–935, 1473
Truck, 1055
Trucks, 1030
True color mode, cathode ray tube (CRT), 174
TSUPREM4 calibration, 28
Tube length, microscopy, 1115
Tungsten filaments, 222
Tungsten light, 1055
Tunics of human eye, 746
Turbulent flow, flow imaging, 405
Twinning, silver halide, 1266–1267
Twisted nematic (TN) liquid crystal displays, 959–961
Two-beam dynamic theory for crystals, 281–284, 281
Two-dimensional Fourier transforms, 1104–1105
Two-dimensional imaging
backlight systems for, 1339
ground penetrating radar and, 471–472
in infrared imaging, 809

in magnetic resonance imaging (MRI), 983, 987–988
in search and retrieval systems, 623, 628–630
terahertz electric field imaging and, 1398
Two-point resolution limit, 1087
Two-positive imaging, gravure printing, 461
Two-scale relations, 1446
Two-slit interference, 242–243
Type 500/600 instant films, 847
Type C videotape, 1055
Type K/T/U/Y or Z core, 1055

U

U space representation, 5–6
U2 aerial surveillance planes, 775–776
UC SIM, 478–484
Ultra high-frequency (UHF) television, 1362
Ultra Panavision, 1031
Ultramicroelectrodes (UME), 1248–1259
Ultrasonic cleaner, 1055
Ultrasonography, ultrasound, 1412–1435
in endoscopy, 338–340
image formation in, 571, 573
in magnetic resonance imaging (MRI) vs., 983
in medical imaging, 745
in scanning acoustic microscopy (SAM), 1228
Ultraviolet radiation, 218, 219, 239, 356, 1055
art conservation and analysis using, 661, 668–672
electron paramagnetic resonance (EPR) imaging for, 296
extreme ultraviolet imaging (EUV), 1005–06
far ultraviolet imaging of proton/electron auroras, 1016–1020
fluorescence microscopy, 1135–37
gravure printing, 461
photodetectors and, 1196
radar and over-the-horizon (OTH) radar, 1143
Ultraviolet catastrophe, 211
Uncertainty principle, 215, 259
Uncrossed viewing, in three-dimensional imaging, 1336
Undulator magnet, 221
Uniform Chromaticity Scale (UCS), 535
Universal leader, 1055
University of Chicago, secondary ion mass spectroscopy (SIMS) in (UC SIM), 478–479
Unmanned aerial vehicles (UAVs), 780
Unsharp masks, in forensic and criminology research, 725
Unsqueezed print, 1055
Upatnieks, J., 504
Upward continuation, gravity imaging, 451
Useful yield, secondary ion mass spectroscopy (SIMS), 477
UVW and U*V*W* coordinate systems, 108

V

V number (*See also* Abbe number), 234
Vacuum and wave equation, 212
Value, 103
Valve rollers, 1055
Van Allen belts, energetic neutral atom (ENA) imaging, 1006–1010
Van Dyke Company, 456
Variable area sound track, 1055
Variable density sound track, 1055
Variable length coding, 1388
Varifocal mirrors, in three-dimensional imaging, 1342–1343
Vectograph three-dimensional imaging, 1331
Vector quantization, compression, 154, 633
Vegetation, geologic imaging, 653
Velocimetry, 413–416
Velocity effects, 228, 764
flow imaging and, 413–416
planar laser-induced fluorescence (PLIF) in, 863
spectral line analysis of, 685–686
Verifax, 299
Versatec, 299
Vertical derivatives, gravity imaging, 452
Vertical disparity, in three-dimensional imaging, 1329
Vertical interval time code (VITC), 1055
Very high-frequency (VHF) television, 1362
Very Large Array Radio Telescope, 693
Vesicular films, art conservation and analysis using, 662
Vestigial sideband (VSB) television, 1362, 1389
VGA video, in field emission displays (FED), 382
Vibration, molecular, 216
Vibrational relaxation, 255
Video (*See also* Motion pictures; Television), 1385–1388
authentication techniques, 740
cameras for, 1029–31
component video standards, 1380–82
compressed video, 1385–86
digital (*See* Digital video)
Digital Video Broadcast (DVB), 1392
in forensic and criminology research, 709–714
format conversion in, 720–722
group of pictures (GOP) in, 1386
high-speed photography and, 498–499
I, P, and B frames in, 1387
photoconductors cameras, 1174
Polachrome, 848
Polavision, 848
surveillance imaging using, 714–715
Video assist, 1031
VideoDisc, 16
Videophone, 156–157, 156
Videotape editing, 1035
Viewer, 1055
Viewfinders, 1029, 1346
Viewing angle, 387, 967
Viewing distance, 1347
ViewMaster, 1330, 1331
Vignetting, 1055
Virtual image, 1040, 1328, 1330
Virtual phase CCD (VPCCD), 1198
Virtual states, 259
Virtual transitions, 258–259
Visible light, 356, 1072
Visible Human Project, search and retrieval systems, 616–617
Visible light, 218, 219, 665–666, 782, 803, 1393
Vision tests, 747
VisionDome, 1335–1336
VistaVision, 1031
Visual angle, human vision, 747
Visual areas, 516, 518, 563, 565, 569
Visual Arts System for Archiving and Retrieval of Images (VASARI), 663–664
Visual cortex, 65, 72, 563–570
Visual field, human vision, 566
Visual field mapping, in magnetic field imaging, 975
Visual information rate, 50, 73–74
Visual magnification, 1077
Visual quality, 74–75
Visualization technology, 773, 1327

VisualSEEk, 618, 621–622, 624, 627, 629, 630
Vitascope, 1022
Voice over, in motion pictures, 1033, 1055
Voids, 604
Volcanic activity, geologic imaging, 651
Volume grating, holography, 508
Volume imaging, holography, 507–508
Volume Imaging Lidar, in meteorological research, 769
Volumetric displays, in three-dimensional imaging, 1341–1343
von Ardenne, Manfred, 262
von Laue interference function, 279
von Uchatius, Franz, 1022
VORTEX radar, 1471
VREX micropolarizers, in three-dimensional imaging, 1334–1335

W

Wall eyed, in three-dimensional imaging, 1330
Warm up, liquid crystal displays (LCDs), 184
Waste management, geologic imaging, 656–659
Water Cerenkov counters, particle detector imaging, 1162
Watermarking, digital (*See* Digital watermarking)
Watershed transforms, 430, 587, 646
Watts, 524
Wave aberration, 542–544
Wave equation, in transmission electron microscopes (TEM), 212
Wave fronts, 212, 243, 1083, 1084, 1086, 1090
Wave number, 213
Wave propagation, 220–231
  in ground penetrating radar, 466, 467–469
Wave vector transfer (Q), 251
Wave vs. particle behavior of light, 210–211
Waveform monitors, television, 1361
Waveform repetition frequency (WRF), radar and over-the-horizon (OTH) radar, 1147, 1150
Waveforms, 212, 285, 1147, 1150–1151
Waveguides, radar, 1452
Wavelength, 212, 1072, 1109
Wavelength analysis, 448
Wavelet coding, compression, 154
Wavelet transforms, 1444–1450
Wavelets, 243, 622
Wax melt printers, 190–191, 195
Weak beam images, in transmission electron microscopes (TEM), 270
Weather radar, 1450–74
Weave, 1055
Web-fed presses, gravure printing, 459–460, 459
Weber–Fechner law, 747
Weber fractions, quality metrics, 611
Weber's law, 611
Wedgewood, Josiah, 1345
Weighting, amplitude, 3
Wet-gate printer, 1055
WHAT IF software, 708
Whisk broom scanners, 806
White, 219
White balance, 520
White field response, flow imaging, 397
White light, 219
White point normalization, 533
White uniformity, cathode ray tube (CRT), 35
White, reference, 102
Whole field imaging, 1072
Wide-angle lens, 1347, 1354
Wide-screen, 1031, 1055
Wien displacement law, 222, 803
Wiener filters, 49, 69, 71, 73, 76, 80, 167, 1322
Wiener matrix, 60
Wiener restoration, 68, 70–71, 74–75, 82
Wiggler magnet, 221
Wild, 1055
Wind profiling radar, 1149, 1469–1471
Winding, 1055
Window thermal testing, using infrared imaging, 815
Wipe, 1055
Wire chamber scintillation tracking, 1159–1162, 1168
WKB approximation, in charged particle optics, 89
Wold components, in search and retrieval systems, 623
Wollaston prisms, 1106
Wolter telescopes, 1499–1501
Work print, 1056
Working distance, microscopy, 1116
WorkWall three-dimensional imaging, 1335
World Geodetic System, 445
Wright, Wilbur, 773
Write black, in electrophotography, 317
Write gates, SQUID sensors, 13–14
Write white, in electrophotography, 317
Wynne, Klass, 1400–01

X

X-ray analysis (EDX), 262, 478
X-ray astronomy, 219
X-ray crystallography, 696–699
X-ray diffractometers, 244
X-ray Evolving Universe Satellite, 1509
X-ray fluorescence (XRF), 676–677, 1475–1495
X-ray interferometric telescopes, 1503–04
X-ray telescope, 239, 1495–1509
X-ray telescopes, 239
X-ray transform, 1404
X-rays, 210–211, 218, 219, 221, 224, 239, 242, 249, 256–260, 272, 350, 590, 803, 1067, 1393
  Array of Low Energy X Ray Imaging Sensors (ALEXIS), 905, 929
  art conservation and analysis using, 672–680
  astronomy science and, 683, 688
  in biochemical research, 694, 696–699, 705
  Bragg reflection in, 244
  image formation in, 572
  in medical imaging, 743, 745, 748, 752–753, 756
  non-silver output in, 676
  phosphor thermography, 865
  photodetectors and, 1197
  radar and over-the-horizon (OTH) radar, 1143
  sources of, 1477
  in tomography, 1404
  X-ray Evolving Universe Satellite, 1509
  X-ray fluorescence imaging, 1475–1495
  X-ray interferometric telescopes, 1503–1504
  X-ray telescope, 1495–1509
Xenon arc, 1056
Xenon lamps, 1037
Xerography (*See also* Electrophotography), 299, 1174
Xeroradiography, 312
Xerox copiers, 574
Xerox Corporation, 299
Xerox Docu Tech, 300, 302, 303
XMM Newton telescope, 1508
XYZ coordinate system, 107, 532, 537, 619

Y
Yellow, 1056
Yellow couplers, color photography, 134
YIQ coordinate system, 106–107, 149
Young, Thomas, 242–243
Yttrium aluminum garnet (YAG) laser, 391
YUV coordinate system, 106–107, 149

Z
Z contrast imaging, 277
Z dependence, 478
Zeeman effect, 217, 218
Zeeman states, 217
Zeiss, Carl, 1106
Zernicke moments, 627
Zernike polynomials, 542–543
Zernike, Frits, 1106, 1128
Zero frame reference mark, 1056
Zero padding, 68
Zero power (afocal) systems, 1074, 1079
Zinc oxide, 381
Zoetropic effect, 1022
Zoom in/out, 1056
Zoom lens, 1029
Zoylacetanilides, 134
Zweiton, 1365
Zwitterions, 124–125