Three-Dimensional Holographic Imaging. Edited by Chung J. Kuo, Meng Hua Tsai
Copyright © 2002 John Wiley & Sons, Inc.
ISBNs: 0-471-35894-0 (Hardback); 0-471-22454-5 (Electronic)

THREE-DIMENSIONAL HOLOGRAPHIC IMAGING

WILEY SERIES IN LASERS AND APPLICATIONS

D. R. VIJ, Editor
Kurukshetra University

OPTICS OF NANOSTRUCTURED MATERIALS • Vadim Markel

LASER REMOTE SENSING OF THE OCEAN: METHODS AND APPLICATIONS • Alexey B. Bunkin and Konstantin Voliak

COHERENCE AND STATISTICS OF PHOTONS AND ATOMS • Jan Peřina, *Editor*

METHODS FOR COMPUTER DESIGN OF DIFFRACTIVE OPTICAL ELEMENTS • Victor A. Soifer

THREE-DIMENSIONAL HOLOGRAPHIC IMAGING • Chung J. Kuo and Meng Hua Tsai

THREE-DIMENSIONAL HOLOGRAPHIC IMAGING

Edited by

Chung J. Kuo
Meng Hua Tsai

A WILEY-INTERSCIENCE PUBLICATION

JOHN WILEY & SONS, INC.

This publication is designed to provide accurate and authoritative information in regard to the subject matter covered. It is sold with the understanding that the publisher is not engaged in rendering professional services. If professional advice or other expert assistance is required, the services of a competent professional person should be sought.

ISBN 0-471-22454-5

This title is also available in print as ISBN 0-471-35894-0.

For more information about Wiley products, visit our web site at www.Wiley.com.

To my father, Ming-Fu Kuo, my mother, Yin-Chiao Chao, Kuo, and my wife, Chih-Jung Hsu

CHUNG J. KUO

To my husband, Chu Yu Chen

MENG HUA TSAI

CONTRIBUTORS

Benton, Stephen Media Laboratory, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139

Caulfield, H. John Department of Physics, Fisk University, 1000 18th Avenue North, Nashville, Tennessee 37208

Cescato, Lucila Laboratório de Óptica, Instituto de Física Gleb Wataghin, UNICAMP, Cx. P. 6165, 13083-970 Campinas, SP, Brasil

Chang, Hsuan T. Department of Electrical Engineering, National Yunlin University of Science and Technology, Touliu Yunlin, 64002 Taiwan

Chang, Ni Y. Department of Electrical Engineering, National Chung Cheng University, Chia-Yi, 62107 Taiwan

Chen, Oscal T.-C. Department of Electrical Engineering, National Chung Cheng University, Chia-Yi, 621 Taiwan

Dai, Li-Kuo Solid-State Devices Materials Section, Materials and Electro-Optics Research Division, Chung-Shan Institute of Science and Technology, Tao-Yuan, 325 Taiwan

Frejlich, Jaime Laboratório de Óptica, Instituto de Física Gleb Wataghin, UNICAMP, Cx. P. 6165, 13083-970 Campinas, SP, Brasil

Huang, Kaung-Hsin Solid-State Devices Materials Section, Materials and Electro-Optics Research Division, Chung-Shan Institute of Science and Technology, Tao-Yuan, 325 Taiwan

Hwang, Jen-Shang Department of Electrical Engineering, National Chung Cheng University, Chia-Yi, 621 Taiwan

Jannson, Tomasz Physical Optics Corporation, 2545 West 237th Street, Torrance, California 90505

Jih, Far-Wen Solid-State Devices Materials Section, Materials and Electro-Optics Research Division, Chung-Shan Institute of Science and Technology, Tao-Yuan, 325 Taiwan

Kuo, Chung J. Institute of Communication Engineering, National Chung Cheng University, Chia-Yi, 62107 Taiwan

Liu, Wei-Jean Department of Electrical Engineering, National Chung Cheng University, Chia-Yi, 621 Taiwan

Pappu, Ravikanth Media Laboratory, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139

Plesniak, Wendy Media Laboratory, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139

Poon, Ting-Chung Optical Image Processing Laboratory, Bradley Department of Electrical and Computer Engineering, Virginia Polytechnic Institute and State University, Blacksburg, Virginia 24061

Schilling, Bradley W. U.S. Army CECOM RDEC, Night Vision and Electronic Sensors Directorate, Fort Belvoir, Virginia 22060

Shamir, Joseph Department of Electrical Engineering, Technion, Israel Institute of Technology, Haifa 32000, Israel

Sheen, Robin Department of Electrical Engineering, National Chung Cheng University, Chia-Yi, 621 Taiwan

Tang, Shiang-Feng Solid-State Devices Materials Section, Materials and Electro-Optics Research Division, Chung-Shan Institute of Science and Technology, Tao-Yuan, 325 Taiwan

Ternovskiy, Igor Physical Optics Corporation, 2545 West 237th Street, Torrance, California 90505

Tsai, Meng Hua Department of Information Technology, Toko University, Chia-Yi, 613 Taiwan

Weng, Ping-Kuo Solid-State Devices Materials Section, Materials and Electro-Optics Research Division, Chung-Shan Institute of Science and Technology, Tao-Yuan, 325 Taiwan

CONTENTS

PREFACE

Holography has been extensively studied for the past 50 years. With the advent of electronic devices such as image-capturing devices [charge-coupled device (CCD) and complementary metal–oxide–semiconductor (CMOS) sensor] and the spatial light modulator (SLM), it is now possible to capture the interference pattern in real time and then display it on a SLM. In other words, a three-dimensional holographic pattern can be captured by an image-capturing device or calculated by a computer and the three-dimensional object can then be reconstructed by using electro-, acousto-, or magnetooptic SLM or computer peripherals. Moreover, many holographic techniques were also invented recently and attracted the attentions of researchers in photonics area.

The chapters in this book were contributed by different research groups around the world and introduce the reader to the general concepts and fundamental research issues of holographic techniques. Although each chapter is self-contained, the chapters are organized in the following order: The first part (Chapters 1–5) of the book deals with holographic techniques and related issues. The applications and components of holographic techniques are covered in the second part (Chapters 6–9). Finally, the stereovision technique and its analysis are presented (Chapter 10). Due to the extensive coverage of topics in holographic technique, this book can be used as a graduate textbook for three-dimensional real-time holography or a reference book for researchers and students who are working at holographic techniques. Since each chapter is self-contained, readers can study only the chapters that are of interest to them.

We are indebted to S. A. Benton at MIT Media Laboratory, whose encouragement during the preparation of this book is very much appreciated. One of us (Chung J. Kuo) is indebted to his Ph.D. students Chia H. Yeh and Yi C. Tsai for their support. Finally, secretarial support from Meei-Jy Shyong, Yi-Jing Li, and Avon Ning is very much appreciated.

Chung J. Kuo
Meng Hua Tsai

National Chung Chang University
September 2001

ABOUT THE EDITORS

Chung J. Kuo received B.S. and M.S. degrees in Power Mechanical Engineering from the National Tsing Hua University, Taiwan, in 1982 and 1984, respectively, and a Ph.D. degree in Electrical Engineering from Michigan State University (MSU) in 1990. He joined the Electrical Engineering Department of the National Chung Cheng University (NCCU) in 1990 as an associate professor and then became a full professor in 1996. He is now the chairman of the Graduate Institute of Communications Engineering of the NCCU. Dr. Kuo was a visiting scientist at the Opto-Electronics and System Lab, Industrial Technology Research Institute, in 1991 and at IBM T. J. Watson Research Center from 1997 to 1998 and a consultant to several international/local companies. He is also an adjunct professor at the National Cheng Kung University.

Dr. Kuo's interests are in image/video signal processing, very large scale integrated circuit signal processing, and photonics and is the codirector of the Signal and Media (SAM) Laboratory at NCCU. He has received the Distinguished Research Award of the NCCU (1998), the Overseas Research Fellowship of the National Science Council (NSC) (1997), the Outstanding Research Award of the College of Engineering, NCCU (1997), the Medal of Honor of the NCCU (1995), the Research Award of the NSC (every year since 1991), the Best Engineering Paper Award of Taiwan's Computer Society (1991), the Electrical Engineering Fellowship of MSU (1989), and the Outstanding Academic Achievement Award of MSU (1987). He was a guest editor for two special sections of *Optical Engineering* and an invited speaker and program committee chairman and member for several international/local conferences. He also serves as an associate editor of the *IEEE Signal Processing Magazine* and president of the SPIE Taiwan Chapter (1998–2000). Dr. Kuo is a member of Phi Kappa Phi, Phi Beta Delta, IEEE, OSA, and SPIE and is listed in *Who's Who in the World.*

Meng Hua Tsai received her B.S. degree in Engineering Science from the National Cheng Kung University, Tainan, Taiwan, in 1991, and M.S. and Ph.D. degrees in Electrical Engineering from Michigan State University (MSU), East Lansing, MI, in 1995 and 1999, respectively. During her stay at MSU, she received both the Teaching Assistantship from the Department of

Electrical Engineering and Research Assistantship from the College of Engineering. She also worked as a research assistant in the Electronic and Surface Properties of Materials Center at MSU from 1998 to 1999.

In 1999, she joined the Graduate Institute of Communications Engineering of the National Chung Cheng University (NCCU), Chia-Yi, Taiwan, for postdoctoral research and is involved in the design and optimization of free-space optical communication system for chip-to-chip interconnection. She is now an assistant professor in the Department of Information Technology, Toko University, Chia-Yi, Taiwan, and also an adjunct assistant professor at the NCCU since 2000. Her current research interests include modeling and characterization of plasma sources for semiconductor processing, optoelectronic devices, and fiber-optic communication systems.

Three-Dimensional Holographic Imaging. Edited by Chung J. Kuo, Meng Hua Tsai
Copyright © 2002 John Wiley & Sons, Inc.
ISBNs: 0-471-35894-0 (Hardback); 0-471-22454-5 (Electronic)

CHAPTER 1

Introduction

MENG HUA TSAI
Department of Information Technology
Toko University
Chia-Yi, 613 Taiwan

CHUNG J. KUO
Institute of Communication Engineering
National Chung Cheng University
Chia-Yi, 62107 Taiwan

The concept of holography was first introduced by Gabor in 1948 [1]. A holograph is a recording of the interference patterns formed between two beams of coherent light coming from a laser on a light-sensitive media such as a photographic film. A brief introduction of holography follows. The light beam coming from a laser is first divided into two beams by a beamsplitter and then expanded. One beam, called the reference beam, goes directly to the photographic plate. The other beam is directed onto a three-dimensional (3D) object under observation. The scattered light waves from the object combines with the light waves from the reference beam at the photographic plate. Because of their high degree of coherence, the two sets of waves form an interference pattern on the plate. These interference fringes are recorded on the plate and form a *hologram.* A vivid three-dimensional virtual image of the observing object is then reconstructed by illuminating the hologram with a plane-parallel light beam from the laser. For more details of holographic recording technology and its application, the reader should refer to Refs. 2 and 3.

In addition to the holographic technique stated above, many other types of holographic techniques are available. This book will give the reader a more complete aspect of holographic recording techniques. It covers the concept of holographic recording of a moving point, the stability problem in holographic recording, and the real-time holographic recording technique: optical scanning holography (OSH), holographic information-processing technique in electronic holography, the application of OSH to laser radar systems, the principles and design considerations of the optoelectronic devices frequently used in holographic recording, and the design of a computer-generated hologram (CGH). Finally, a theory called analysis of catastrophe is introduced, which may be an indication of the basis for visual perception.

A brief description of the book follows. Chapter 2 discusses the theory of holographic recording of moving point trajectories followed by a comparison of optical and electronic recording methods. The results are also compared with other ways of recording a line in 3D space, which include recording an actual luminous line, sequential recording of points, and computer generation of lines. When taking a holographic recording, it is important to have a stable system to ensure reproducibility of the hologram and subsequently a clear outcome image of the object. This is usually difficult to achieve because of the changes in the optical path between the arms of the interferometers. These changes may be caused by the mechanical vibrations of the optical components, external perturbations transmitted to the setup, or thermal drifts in the air between the interfering beams. These factors can cause the movement of the fringes and result in low reproducibility of the holographic recording. To correct this problem, an active fringe stabilization system that detects the fringe perturbation and produces a phase correction feedback signal to compensate the perturbations is needed.

Chapter 3 presents a self-stabilized holographic recording system that uses the hologram being recorded as a reference to stabilize the exposure. The fringe stabilization system is composed of a detection system that uses the wave mixing to amplify the holographic fringes, an electronic system that provides the synchronous detection and the amplification, and a phase modulator device, which realizes the correction feedback in the phase. Later in the chapter, applications of this system to self-stabilize holographic recording in different photosensitive materials is discussed.

In Chapter 4, a real-time holographic recording technique called optical scanning holography is proposed in which holographic information of an object can be recorded using two-dimensional (2D) optical heterodyne scanning. Upon scanning, the scattered or reflected light is detected by photodetectors. The instantaneous electrical signal from the photodetectors thus contains the holographic information of the scanned object. This holographic recording technique is commonly known as electronic holography since it uses electronic processing instead of photographic films. Some important applications of OSH are presented in this chapter, such as 3D holographic microscopy, 3D image recognition, and 3D preprocessing and coding.

Chapter 5 describes the experiments with tangible, dynamic holographic images using a prototype system called the holo–haptic system, which comprises a sizeable arsenal of computers and both commercial and custom hardware. By combining a force model with the spatial visual image, it allows fingertips to apply a “reality test” to the images and provides the most intimate way of interacting with them.

In the process of electronic holography, computers are generally used. To handle the process with computers, the original analog holographic information is converted into digital format using intensity-recording devices such as charge-coupled devices (CCDs). To have high resolution, the digitalized holographic information will become very huge. Therefore, it should be compressed for ease of storage and transmission. A high compression ratio is expected to reduce the huge data amount of holographic information. As a result, the distortion introduced by compression occurs and the performance based on different compression methods should be examined. Chapter 6 investigates the characteristics of the interference pattern and proposes a novel method to enhance the light efficiency of the holography. Sampling and quantization effects on digitized holographic information are briefly introduced. A nonlinear quantization model is used to reduce the quantization noise. Finally, a Joint Picture Expert Group (JPEG)–based technique is used for compressing the interference pattern that has been transferred to a gray-scale image.

Chapter 7 describes the application of OSH (addressed earlier in Chapter 4) to the laser radar problem. The technique is similar in operation to a standard laser radar system in which the image is built pixel by pixel as the laser pattern is scanned over the object. The main difference between a standard laser radar and a holographic laser radar is in the scanning field. A typical laser radar system employs a spot scan while a holographic laser radar is based on scanning with an optically heterodyned Fresnel zone pattern (FZP). The electronic nature of the system offers the spectral flexibility advantage over traditional holographic recording systems. Theoretically, it is possible to record holograms by this technique in any spectral band where a coherent (laser) source and detector combination exists. Therefore, the technique is particularly well suited to multicolor holography and offers the possibility of holographic recording at infrared wavelengths. The scanning aspect of the technique offers another advantage by relaxing the size constraints of the object or scene to be recorded. With this system, it is feasible to record holograms of large-scale objects, or scenes.

Chapter 8 presents the theory and applications of optoelectronic devices commonly used in electronic holography as light sources or recording devices. It includes both light-emitting components such as laser diodes and light-receiving components such as photodiodes. In optical sensor design, the complementary metal–oxide–semiconductor technologies are compared in order to analyze their features, performances, and applications. Passive and active pixel sensors and the fill factors, quantum efficiencies, and fixed pattern noises of these pixel sensors are presented. The driving circuit design of these devices is discussed and, finally, a prototype chip with a die size of 1.8 mm × 1.5 mm is implemented for chip-to-chip optical interconnection to integrate four photodiodes, photoreceivers, laser diode drivers, and laser diode pads. The performance of this system is analyzed and a conclusion is made.

Traditionally, a hologram is obtained by recording the interference pattern of two different light beams on a high-resolution photographic film. A recent development in hologram making is applying the semiconductor fabrication technique to a material with its surface profile designed by a computer program. This is the so-called computer-generated hologram. The concept of CGH can also be used to design an optical element to diffract an incoming light beam to any desired position. In such an application, a more appropriate name, instead of CGH, would be a diffractive optical element (DOE). In Chapter 9, the design and implementation of CGH/DOE are discussed followed by a description of the fabrication processes for CGH/DOE.

The last chapter introduces a means to recover and use the 3D information from the 2D scene, called analysis-by-catastrophes (ABC). The process describes the scene in terms of only two primitives, the fold and cusp catastrophes, along with their particular locations, scales, and orientations. In this way it dramatically reduces the processing load on the visual cortex. The results of this work indicate that ABC is a possible model for visual perception since it agrees with the many features of visual cortex: local matching of corresponding parts of two stereoscopic retina images, local 3D features of monoscopic images, and modular and modestly parallel visual cortex architecture, to name a few.

REFERENCES

1. D. Gabor, "A new microscopic principle", *Nature* **4098**, 777 (1948).
2. H. M. Smith, *Principles of Holography*, Wiley-Interscience, New York, 1975.
3. R. H. Collier, C. B. Burkhardt, and L. H. Lin, *Optical Holography*, Academic, New York, 1972.

Three-Dimensional Holographic Imaging. Edited by Chung J. Kuo, Meng Hua Tsai
Copyright © 2002 John Wiley & Sons, Inc.
ISBNs: 0-471-35894-0 (Hardback); 0-471-22454-5 (Electronic)

CHAPTER 2

Holograms of Real and Virtual Point Trajectories

H. JOHN CAULFIELD

Department of Physics, Fisk University
1000 18th Avenue North
Nashville, Tennessee 37208

JOSEPH SHAMIR

Department of Electrical Engineering
Technion, Israel Institute of Technology
Haifa 32000, Israel

2.1 INTRODUCTION

In relativity, the orbit of a point event through space–time is called its world line. The world line itself is timeless, because it contains time as one of its dimensions. Over a period of years, we have been fascinated by the prospect of recording world lines of moving points of light holographically. Of course, these will have their three-dimensional (3D) spatial (the 3D trajectory) pattern and be timeless. There will be no way to give a direction of time and all we know is what events (3D positions) are the time neighbors of others.

Does this multidecade effort shed light on relativity or make it easier to understand? Probably not. Holography can help us understand relativity, but that work is due to Abramson [1], not us. Surprisingly, our efforts have caused us to understand holography better. In this work we discuss holographic recording of moving points and compare the results with various aspects of other ways of recording a line in 3D space, such as recording an actual luminous line, sequential recording of points, and computer generation of lines.

2.2 EARLY WORK

Our interest began with our efforts to generate 3D holographic images of synthetic scenes. Why not draw the scene with a moving point source using holography with a fixed reference beam to record the 3D object? Figure 2.1 shows the geometry. We moved the point continuously parallel to the recording plate. Our results were wonderful, both theoretically and experimentally [2].

Theoretically, we showed that coherently time averaging an Airy pattern (the far-field complex wave front of a point source) leads to a $\sin x/x$ pattern (the far-field complex wave front that would have been produced had the whole line been present at once). This seemed quite profound at the time. The coherent integration obliterated the time dimension. It may still be profound. We know that physics based on instants and infinitesimal points fails profoundly at the quantum level. It lacks the coherent integration into the whole. Experimentally, we found that the image of a clean bright line was produced. Without that success, we would not have persisted through the dark decades of disappointments and partial successes that followed.

Physicists progress by jumping to unwarranted generalizations and then examining the results. This is not so much a method as a predisposition. The obvious thing to do after the first success was to move to more complex space–time patterns. We expected, naively it now appears, no problem in recording arbitrarily complex scenes in this way. Instead, we encountered two major problems. One problem we understood almost immediately and later were able to work around to some extent. The other problem we did not even understand, although we immediately invented a way to work around it. We address those two problems below.

2.2.1 Brightness Problem

As we all should have known, there is a communication-theoretic limitation on the information content of the image and how much information we actually see depends on the encryption method. All of the great holographers (e.g.,

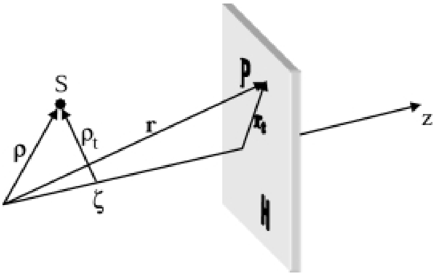


Figure 2.1 Schematics of the optical configuration. *S* is a point source and *H* is the hologram. The other parameters are referred to in the mathematical analysis.

Gabor, Leith, and Denisyuk) knew that. We did too, but it is easy to forget. The information storage density (that is bits per square centimeters for thin holograms and bits per cubic centimeters for thick holograms) is very material dependent. Resolution and noise are the primary determinants. If we use all of that capacity coherently to record a single point, the image may have tremendous signal-to-noise ratio (SNR). On the other hand, if we record and reconstruct N distinct, equally bright points, then each can have *at most* $1/N$ of the available light and $1/N$ of the single-point SNR. We emphasized the words "at most." Only if each point comes from a hologram with unit contrast can we achieve the $1/N$ brightness condition. This would be the case if we recorded the hologram of N coherent points simultaneously. However, in the case as was done in our first holograms, we are talking about recording the N points sequentially. Thus we have holograms from N essentially independent points fully overlapping and then each will use only $1/N$ of the shared dynamic range. The brightness and SNR of each point can be at most $1/N$ of the values achievable for a single point. So, whichever way we choose to record the N points, the brightness and SNR cannot be better than $1/N$ that of a single point and, usually, it will be much lower.

Returning to our special interest here of a continuously moving point, one should ask the question: How big is N? This is a question we did not even begin to answer in the middle period of this multidecade effort.

We now know that the above discussion is oversimplified and that there are ways, depending on the recording material and recording condition, to improve the situation. In fact, at a quite early stage we did conceive of and demonstrate a way to improve the brightness and SNR. We simply moved the points close to the recording medium. Because of the limited angular divergence of the point source, the area on the recording medium illuminated at any instant was small. Thus there was no need for a reference beam where there was no object beam, so we could block that part of the reference beam. Using a complicated optomechanical system, we scanned a point in 3D space near the recording plate and tracked it with the corresponding part of the reference beam. All of the time, most of the recording material received light only near the image of the reference point. The rest of the recording medium was shielded and, therefore, not degraded. Thus no point suffered the full $1/N$ penalty, and very bright images were obtained [3].

2.2.2 Longitudinal Motion Problem

Initially, we did not call the problem by this name. All we observed was that when we moved the point in a 3D orbit (rather than in the 2D plane, parallel to the recording medium), we did not get very good images. In fact, the images were terrible. We did not know why, but we did find a satisfactory experimental way to fix the problem. We chopped (binary time modulated) both beams. For reasons we did not understand at the time, this allowed us to record beautiful 3D images [4].

This review of the history of a small part of holography allows us to introduce the current state of the art. We now know what the longitudinal motion problem was and why chopping "cured" it. We will show below that all parts of the Airy pattern are "blurred out" during any substantial longitudinal motion. Chopping reduced the blurring effects by recording just a very short light segment for each chopping cycle. A general mathematical analysis of the phenomena involved in holographic recording of moving sources follows below. The general consequences will then be represented with some demonstrative examples of special interesting cases.

2.3 MATHEMATICAL ANALYSIS

Assume a point source of strength A_0 positioned at a point represented by $\boldsymbol{\rho} = \hat{x}\xi + \hat{y}\eta + \hat{z}\zeta$ on the left of a recording plane with coordinates $\mathbf{r} = \hat{x}x + \hat{y}y + \hat{z}$, where the "hat" denotes a unit vector. The complex amplitude distribution on the recording plane (Fig. 2.1) is a spherical wave given by (see, e.g., Ref. 5),

$$u_0(\mathbf{r}) = \frac{A_0}{jk|\mathbf{r} - \boldsymbol{\rho}|} \exp(jk|r - \boldsymbol{\rho}|) \tag{2.1}$$

where $k = 2\pi/\lambda$ is the wave number and λ is the wavelength. If the point source moves, the vector $\boldsymbol{\rho}$ is a function of time, which makes the complex amplitude on the observation plane also a function of time. If we wish to record a hologram, we need a reference beam, u_r, and expose a recording medium for a time T. That is, the recorded intensity pattern will be given by

$$I(\mathbf{r}) = \int_0^T |u_r + u_o|^2 \, dt = \int_0^T (|u_r|^2 + |u_o|^2 + u_r^* u_o + u_r u_o^*) \, dt \tag{2.2}$$

The first two terms constitute the so-called zero-order term, which is of no interest at this point, while the last two terms are responsible for the reconstruction of the recorded object. The fourth term reconstructs a phase-conjugate image that, if properly recorded, is spatially separated from the third term which represents the true image. Therefore, we shall concentrate now on the third term, which has the form

$$I_t(\mathbf{r}) = \int_0^T u_r^*(\mathbf{r}) u_o(\mathbf{r}, t) \, dt \tag{2.3}$$

where we noted explicitly that the object wave depends on time while the reference wave is constant in time.

To simplify calculation, we assume the validity of the paraxial approximation,

$$|\mathbf{r}-\boldsymbol{\rho}| = \sqrt{(x-\xi)^2+(y-\eta)^2+(z-\zeta)^2} \approx (z-\zeta)\left[1+\frac{(x-\xi)^2+(y-\eta)^2}{2(z-\zeta)^2}\right] \tag{2.4}$$

Without losing generality we may take a plane reference wave propagating in the positive z direction,

$$u_r(\mathbf{r}) = A_r \exp(jkz) \tag{2.5}$$

This special choice of reference wave is not ideal in practice, but the physical conclusions are similar for other choices, as long as the reference wave is constant in time. Substituting into Eq. 2.3, one exponential factor cancels and we obtain the interesting term of the recorded distribution as

$$I_t(\mathbf{r}) = \int_0^T \frac{A_r A_o}{jk|\mathbf{r}-\boldsymbol{\rho}|} \exp\left\{-jk\zeta(t) + \frac{jk}{2[z-\zeta(t)]}\left([x-\xi(t)]^2+[y-\eta(t)]^2\right)\right\} dt \tag{2.6}$$

where we noted explicitly the dependence of the source coordinates on time. Since we already assumed the paraxial approximation, we may also assume that the variation of $\boldsymbol{\rho}$ does not have a significant effect on the denominator of the first factor, and we may approximate the whole amplitude factor by a constant, C. In fact, this constant has no physical significance because it will be modified by the recording process. Thus, whenever of interest, we shall use this final value, which determines the image brightness or, as it is now usually referred to, the *diffraction efficiency* of the hologram. With these considerations taken into account we may write the term responsible for reconstructing the image in the form

$$I_t(\mathbf{r}) = C \int_0^T \exp\left\{-jk\zeta(t) + \frac{jk}{2[z-\zeta(t)]}\left([x-\xi(t)]^2+[y-\eta(t)]^2\right)\right\} dt \tag{2.7}$$

To further simplify the expression, we choose the origin such that $z = 0$ over the recording plane and define the transverse vectors

$$\mathbf{r}_t = \hat{x}x + \hat{y}y \qquad \boldsymbol{\rho}_t = \hat{x}\xi + \hat{y}\eta \tag{2.8}$$

Evaluating the squares in the exponent, we obtain

$$I_t(\mathbf{r}) = C\int_0^T \exp[-jk\zeta(t)]\exp\left[\frac{-jk|\mathbf{r}_t|^2}{2\zeta(t)}\right]\exp\left[-jk\frac{|\boldsymbol{\rho}(t)|^2}{2\zeta(t)}\right]\exp\left[jk\,\frac{\mathbf{r}_t\cdot\boldsymbol{\rho}_t}{\zeta(t)}\right]dt \tag{2.9}$$

where the dot represents the scalar product of the two transverse vectors.

In the general case, if the source moves substantially more than a wavelength during the exposure, the integration will average out to zero and most of the information recorded will be lost. That is the reason for our frustration during our earlier investigations. However, our initial success was caused by the fact that certain trajectories are partially immune to this averaging effect and, luckily, we chose one of them. Moreover, a residual signal may still be observed even for other cases due to the normalization effect (maximum transmission cannot exceed unity) and some possible nonlinearity of the recording. For illustrative purposes we shall investigate a few special cases below.

2.3.1 Longitudinal Translation with Constant Velocity

Since the transverse coordinate of the source remains constant in this case, we may choose the z axis along the trajectory such that $\boldsymbol{\rho}_t = 0$. Accordingly, Eq. 2.9 reduces to

$$I_t(\mathbf{r}) = C\int_0^T \exp[-jk\zeta(t)]\exp\left[\frac{-jk|\mathbf{r}_t|^2}{2\zeta(t)}\right]dt \tag{2.10}$$

Motion with constant velocity along the z axis can be written as

$$\zeta = \zeta_0 + v_z t \tag{2.11}$$

where v_z is the velocity of the source and ζ_0 is the starting point. Maintaining the paraxial approximation, we may assume $\zeta_0 \gg v_z t$ during the integration time, and then we may write

$$\frac{1}{\zeta(t)} = \frac{1}{\zeta_0 + v_z t} \approx \frac{1}{\zeta_0}\left(1 - \frac{v_z t}{\zeta_0}\right) \tag{2.12}$$

Returning to Eq. 2.10, we obtain

$$I_t(\mathbf{r}) \approx C\exp[-jk\zeta_0]\exp\left[\frac{-jk|\mathbf{r}_t|^2}{2\zeta_0}\right]\int_0^T \exp\left[-jkv_z t\left(1 - \frac{|\mathbf{r}_t|^2}{2\zeta_0^2}\right)\right]dt \tag{2.13}$$

Evaluating the integral, we obtain

$$I_t(\mathbf{r}) \approx \frac{C}{jkv_z(1 - |\mathbf{r}_t|^2/2\zeta_0^2)} \exp[-jk\zeta_0] \\ \times \exp\left[\frac{-jk|\mathbf{r}_t|^2}{2\zeta_0}\right]\left\{1 - \exp\left[-jkv_zT\left(1 - \frac{|\mathbf{r}_t|^2}{2\zeta_0^2}\right)\right]\right\} \tag{2.14}$$

With some rearrangement of factors, this can be written in the form

$$I_t(\mathbf{r}) \approx \frac{C}{jkv_z(1 - |\mathbf{r}_t|^2/2\zeta_0^2)} \left\{\exp[-jk\zeta_0]\exp\left[\frac{-jk|\mathbf{r}_t|^2}{2\zeta_0}\right]\right. \\ \left. -\exp[-jk(\zeta_0 + v_zT)]\exp\left[-\frac{jk|\mathbf{r}_t|^2}{2\zeta_0}\left(1 - \frac{v_zT}{\zeta_0}\right)\right]\right\} \tag{2.15}$$

Apart from the constant amplitude and phase factors, this expression contains two quadratic phase factors and thus represents two spherical wave fronts. One of these originates at the initial position of the source while the second has a radius of curvature modified to $R = \zeta_0^2/(\zeta_0 - v_zT)$. This is an equivalent point source at some intermediate position between the starting point and the ending point of the trajectory. There is also an exposure–time and velocity-dependent phase difference between the two sources leading to interference effects that are also exposure and velocity dependent. As a result, the final reconstructed pattern cannot be uniquely predicted from a practical point of view. In any case, the source trajectory cannot be reconstructed unless the total displacement does not exceed a wavelength by much, where “much” is not too well defined.

2.3.2 Longitudinal Vibration

While a constant longitudinal motion cannot be practically recorded holographically, the situation may change significantly when the translation is not uniform. Like many other discoveries in physics, this fact was discovered accidentally at the early stages of holography when, for no apparent reason, holograms were obtained with ugly dark lines across them. After some deliberation this phenomenon was traced back to the inadequate vibration isolation of the optical table. Eventually, this discovery gave birth to the widespread use of *holographic interferometry* [6] practiced now for various technological and scientific applications such as the evaluation of the vibration modes of mechanical structures. To see how this effect ties to the subject of our discussion here, we return to our point source and let it vibrate longitudinally. We may then describe its z coordinate by the relation

$$\zeta = \zeta_0 + a\cos\Omega t \tag{2.16}$$

where a is the vibration amplitude about ζ_0 and Ω is the circular frequency. Again, maintaining the paraxial approximation for relatively small displacements, we require the vibration amplitude to be small ($a \ll \zeta_0$) and then we may write

$$\frac{1}{\zeta(t)} \approx \frac{1}{\zeta_0}\left(1 - \frac{a\cos\Omega t}{\zeta_0}\right) \tag{2.17}$$

Substituting into Eq. 2.10, we obtain

$$I_t(\mathbf{r}) \approx C\exp[-jk\zeta_0]\int_0^T \exp(-jka\cos\Omega t)\exp\left[-\frac{jk|\mathbf{r}_t|^2}{2\zeta_0}\left(1 - \frac{a\cos\Omega t}{\zeta_0}\right)\right]dt \tag{2.18}$$

Now there are two factors under the integration that are temporally modulated. First, the quadratic phase factor is modified by the time-varying factor on the right. However, by our assumption of small vibrations, the time-varying term is negligible as compared to unity; thus it may be ignored within our approximations. This is not true for the second factor, which was the main reason for the destruction of the holographic recording when the source possessed a uniform motion. To obtain a simple expression, we ignore the modification of the quadratic phase factor and take an integration time much larger than the oscillation period. With these assumptions we obtain

$$I_t(\mathbf{r}) \approx C\exp[-jk\zeta_0]\exp\left[\frac{-jk|\mathbf{r}_t|^2}{2\zeta_0}\right]J_0(ka) \tag{2.19}$$

where J_0 is the zero-order Bessel function of the first kind. We essentially recorded the quadratic phase factor representing the reconstruction of the point source but the resulting diffraction efficiency depends on the vibration amplitude through the Bessel function. Obviously nothing will be recorded when the vibration amplitude factor corresponds to a root of the Bessel function. This was the origin of the dark lines in those historic experiments.

2.3.3 Transverse Motion with Constant Velocity

Motion with a constant velocity in a transverse plane can be described by defining $\zeta = \zeta_0$ and taking the x axis along the direction of motion with its origin at the starting point. Thus the position of the source at a time t is given by $\xi = v_x t$. Substitution into Eq. 2.9 leads to

$$I_t(\mathbf{r}) = C\exp[-jk\zeta_0]\exp\left[\frac{-jk|\mathbf{r}_t|^2}{2\zeta_0}\right]\int_{-T}^{T}\exp\left[-jk\frac{(v_x t)^2}{2\zeta_0}\right]\exp\left[jk\frac{xv_x t}{\zeta_0}\right]dt \tag{2.20}$$

where, for later mathematical convenience, we took the exposure time as $2T$ starting at a time $t = -T$. Changing variables to $\tau = v_x t/\lambda\zeta_0$ and absorbing all constants into C, we obtain

$$I_t(\mathbf{r}) = C\exp\left[\frac{-jk|\mathbf{r}_t|^2}{2\zeta_0}\right]\int_{-v_xT/\lambda\zeta_0}^{v_xT/\lambda\zeta_0} \exp[-jk\lambda^2\zeta_0\tau^2]\exp[j2\pi x\tau]\,d\tau \tag{2.21}$$

Using the rectangular (rect) function, this can also be written in the form

$$I_t(\mathbf{r}) = C\exp\left[\frac{-jk|\mathbf{r}_t|^2}{2\zeta_0}\right]\int_{-\infty}^{\infty} \operatorname{rect}\left(\frac{\lambda\zeta_0}{v_xT}\tau\right)\exp\left[-j\tfrac{1}{2}k\lambda^2\zeta_0\tau^2\right]\exp[j2\pi x\tau]\,d\tau \tag{2.22}$$

The integral is a Fourier transformation from the τ domain to the x domain, resulting in

$$I_r(\mathbf{r}) = C\exp\left[\frac{-jk|\mathbf{r}_t|^2}{2\zeta_0}\right]\left[\operatorname{sinc}\left(\frac{v_xT}{\lambda\zeta_0}x\right)\exp\left(\frac{jkx^2}{2\zeta_0}\right)\right] \tag{2.23}$$

where we again absorbed some constant factors into the constant C. For a long exposure the rect function is long, leading to a sinc function that approaches a delta function. As a result, the 1D quadratic phase factor compensates one dimension of the 2D quadratic phase factor, leading to a cylindrical wave that converges to a line, the trajectory of the source. If the exposure is short, this line is modified by the sinc function. Actually since the propagation is essentially related to a Fourier transformation, this sinc function is transformed, approximately, back to a rect function in the space domain, which delimits the finite trajectory of the source. Unlike in the case of longitudinal motion, we see that a recording of uniform motion along a transverse line reconstructs the trajectory, as observed in those early experiments.

2.3.4 Circular Motion in a Transverse Plane

Here we take again $\zeta = \zeta_0$ and now we have $|\boldsymbol{\rho}_t|^2 = R^2$, where R is the radius of the trajectory and $\mathbf{r}_t \cdot \boldsymbol{\rho}_t = R|\mathbf{r}|\cos(\phi_0 + \Omega t)$, where Ω is the angular velocity of the source and ϕ_0 is some initial position. Substitution into Eq. 2.9 yields

$$\begin{aligned} I_r(\mathbf{r}) = C\int_0^T &\exp[-jk\zeta_0]\exp\left[\frac{-jk|\mathbf{r}_t|^2}{2\zeta_0}\right] \\ &\times\exp\left[-jk\frac{R^2}{2\zeta_0}\right]\exp\left[jk\frac{R|\mathbf{r}_t|\cos(\phi_0+\Omega t)}{\zeta_0}\right]dt \end{aligned} \tag{2.24}$$

Again absorbing all constant factors in the constant C, we may simplify this relation to

$$I_r(\mathbf{r}) = C \exp\left[\frac{-jk|\mathbf{r}_t|^2}{2\zeta_0}\right] \int_0^T \exp\left[jk\frac{R|\mathbf{r}_t|\cos(\phi_0 + \Omega t)}{\zeta_0}\right] dt \tag{2.25}$$

For long exposures the initial phase is irrelevant and the integral reduces to a Bessel function:

$$I_r(\mathbf{r}) = C \exp\left[\frac{-jk|\mathbf{r}_t|^2}{2\zeta_0}\right] J_0(kR|\mathbf{r}_t|) \tag{2.26}$$

The quadratic phase factor that multiplies this Bessel function is equivalent to a lens of focal length ζ_0. Thus, illumination with a plane wave generates the Fourier transform of the Bessel function, which is the original circular trajectory.

2.4 ANALOGIES TO CODED APERTURE IMAGING

The use of a reference beam coherent with each object beam allowed the coherent summation of those object beams. This is what converted the moving Airy pattern into a sinc pattern, for example. We now wish to remove some of those restrictions. Suppose there were no reference beam at all. Then, obviously, the recorded pattern would be a blur from each point. This is of no interest whatever. By adding a local reference beam to the object beam from a point near the hologram, we were able to prevent exposure of much of the hologram at any instant. This meant that we lost coherence between points. Thus, we have called this the incoherent case. It is the nature of a world line to lose information on the temporal dimension. The hologram is the same regardless of the order in which the points were recorded. In fact, all of them may have been recorded at once so far as we can tell from the hologram. These observations allow us to identify these incoherent world line holograms with a different type of hologram — a coded aperture imaging record.

Coded aperture imaging began in the early 1960s with the work of Mertz [7]. It was intended to form images of mutually incoherent object beams. The idea was to let each object cast a shadow onto the recording plane and the mask used before the shadow casting was a Fresnel zone plate. Consider a single point source. As the point moves laterally, so does its shadow. As the point gets closer to the recording plane, its shadow gets bigger. Therefore, the three spatial coordinates of the point are encrypted in the recording. In addition, shining monochromatic light at the transparent record will decrypt that information automatically. Early on, we called this holography by shadow casting [8].

Now we understand that coded aperture imaging is a kind of world line hologram. The only difference from our near-hologram recording is that in coded aperture imaging all of the points are present at once.

Making that identification allows us to benefit from roughly 35 years of research on coded aperture imaging. This is especially true when we think of computer-generated world line holograms. Each point can be replaced by its own characteristic pattern, properly scaled and translated. One of the enabling "tricks" of coded aperture imaging is the use of off-axis Fresnel zone plate with a special linear (Ronchi) grating near the object [9]. This allows off-axis reconstruction. The analogy between this and our chopping experiment is obvious. Alternatively, we can use the negative of the traditional zone plate to allow some cancellation of the undiffracted light in on-axis reconstruction [10].

Quite independently of either our optical work or coded aperture imaging, the digital computer hologram group in Essen has designed point-by-point holograms for display [11]. They even used the technique of limiting the extent of the reference beam to reduce the demands on dynamic range of the recording material for near-hologram-plane points. Just as we did, they got very satisfactory displays of the patterns so recorded.

2.5 SYNTHETIC RECORDING

Point-by-point holograms can be recorded sequentially as suggested above. Alternatively, the set of points can be used as a whole to generate a pattern that can be recorded as a whole. Obviously, then it will be a coherent recording of a line. However, if we take an imaginary point source to calculate its diffraction pattern and store this electronically, we can display it on a spatial light modulator (SLM) or write the information onto a diffractive optical element. Illuminating this element by coherent light will reconstruct the point source. Now, as before, we can assemble in the computer memory the diffraction patterns of the whole sequence of points in the trajectory and then we essentially have an electronic hologram of the world line. That is, we throw the time information away before recording rather than during recording. The now-traditional approach to doing this is to compute a Fresnel transform of the world line and then use some computer hologram method to record it. But there may be a way to record such holograms on line electronically. We review one way of doing this briefly here.

For historical reasons not of interest here, we began to work on the electronic evolution of holograms on SLMs. Given a way to measure and evaluate the holographically produced image, we can use optimization algorithms, such as genetic algorithms or projections-onto-constraint-sets algorithms to adjust the pixel values of the SLM to achieve optimum results [12, 13]. That is, we would evolve a hologram pattern electronically with the figure of merit being the closeness of correspondence of the reconstructed wave-front intensity to the target world line [14].

Again the information capacity problem reappeared. However, using proper optimization algorithms, the storage capabilities were efficiently exploited and we could achieve quite nice images.

In some respects, the electronic and other computer holograms are superior. They do not suffer the 1/N loss from mutual incoherence of the hologram. The image is evaluated as a whole even though points may be measured one at a time. Of course this transduces the 3D image into a 2D SLM pattern, which is easily stored and transmitted. The light efficiency is very high. The directly optically recorded holograms tend to have low light efficiency, but by copying onto a second hologram that can be corrected.

2.6 DISCUSSION

In principle, this chapter reviews an extremely narrow aspect of holography. However, looking at this issue in a broader context, the developments described here are significantly related to many other aspects of holography and may have an appreciable impact on holography in the third millenium. We may recall that holography was invented for microscopic purposes [15]. Very little was achieved in that field until now, but holography made the big strides forward when it was realized that real-looking 3D images can be recorded and reconstructed [16, 17]. With these developments beautiful 3D display holograms of real objects could be made. This is where our first experiments came in: Can we record a drawing that physically does not exist? As described in this chapter, our success was quite limited. Much of the phenomena we observed at those early stages we did not exactly understand, and we even reinvented average-time holographic interferometry without realizing it.

The further development toward achieving our aim came when it was realized that a holographic recording is a "drawing" of interference fringes and, in principle, one may use a computer to calculate these fringes and plot them on a transparency [18]. Illuminating this transparency with the calculated reference wave will regenerate the object even if that object existed only in the computer memory. This was the beginning of what is referred to as *computer-generated holography* (CGH).

The initial idea behind CGH was the design of objects for comparison in a production line or for decorative displays. It did, however, not take long to find other applications in a diverse list of areas. The reason is that, if we generalize the notion of the CGH, it can be designed to generate any desired complex amplitude distribution as long as it does not contradict physical principles and technological limitations. This is really the basis for the more general field of *diffractive optical elements* (DOEs). Most DOEs are digitally designed, but from various aspects they function like optically recorded holograms. As indicated earlier, DOEs can now be designed for displaying line segments in 3D space [19–22] as well as much more complicated structures [23–25]. A specially interesting structure is an intensity distribution that rotates during

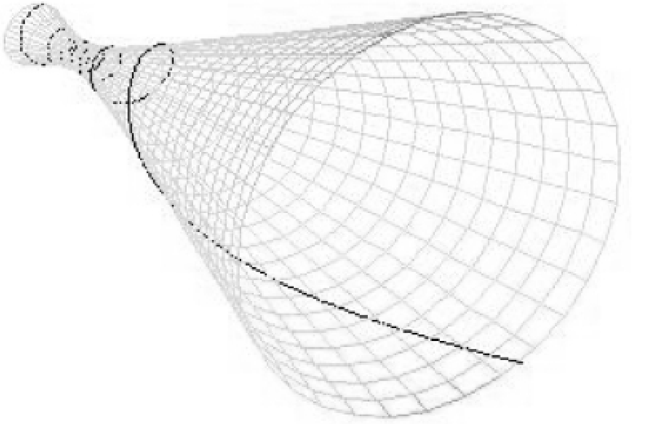


Figure 2.2 Schematic representation of a light structure designed to rotate as it propagates. Each point within a transverse distribution follows a helical trajectory during propagation.

propagation [26, 27]. In this structure, light rays describe a helical trajectory (Fig. 2.2); they are helical rays. Obviously, as shown in this chapter, such a continuous trajectory cannot be recorded as a continuous-time exposure of a moving point source.

2.7 CONCLUSIONS

A point moving through space–time describes an orbit that can be recorded holographically in a variety of ways. The resulting image has lost its temporal information. The consequence is that the same hologram can be recorded with points occurring in any order or even no order all or points present simultaneously. Such holograms are being studied for displays and for projection of patterns onto 3D surfaces. Their study is also a very useful means to understand and teach holography.

REFERENCES

1. N. Abramson, *Light in Flight or the Holodiagram: The Columbi Egg of Optics*, SPIE Press, Bellingham, 1996.
2. H. J. Caulfield, S. Lu, and H. W. Hemstreet, Jr., "Holography of moving objects," *Phys. Lett.* **25A**, 294 (1967).
3. E. S. Gaynor, W. T. Rhodes, and H. J. Caulfield, "Exposure compensation for sequential superposition holographic display," *Appl. Opt.* **26**, 4373–4376 (1987).

4. H. J. Caulfield, S. Lu, and J. L. Harris, "Biasing for single-exposure and multiple-exposure holography," *J. Opt. Soc. Am.* **58**, 1003 (1968).
5. J. Shamir, *Optical Systems and Processes*, SPIE Press, Bellingham, 1999.
6. R. L. Powell and K. A. Stetson, "Interferometric vibration analysis by wavefront reconstruction," *J. Opt. Soc. Am.* **55**, 1593 (1965).
7. L. Mertz, "A dilute image transform with application to an x-ray star camera," *Proc. Symp. Mod. Opt.* (1968).
8. H. J. Caulfield and A. D. Williams, "An introduction to holography by shadow casting," *Opt. Eng.* **12**, 3 (1973).
9. H. H. Barrett and F.A. Horrigan, "Fresnel zone plate imaging of gamma rays: theory," *Appl. Opt.* **12**, 2686 (1973).
10. M. D. Tipton, J. E. Dowdey and H. J. Caulfield, "Coded aperture imaging with an on-axis fresnel zone plates," *Radiology*, **112**, 155 (1974).
11. A. Jendral and O. Bryngdahl, "Synthetic near-field holograms with localized information," *Opt. Lett.* **20**, 1204–1206 (1995).
12. Mahlab, J. Shamir, and H. J. Caulfield, "Genetic algorithm for optical pattern recognition," *Opt. Lett.* **16**, pp. 648–650 (1991).
13. T. Kotzer, J. Rosen, and J. Shamir, "Application of serial and parallel projection methods to correlation filter design," *Appl. Opt.* **34**, 3883–3895 (1995).
14. R. Piestun and J. Shamir, "Control of wavefront propagation with diffractive elements," *Opt. Lett.* **19**, 771–773 (1994).
15. D. Gabor, "Microscopy by reconstruction of wavefronts," *Nature* **161**, 777 (1948); *Proc. Roy. Soc. A*, **197**, 454 (1949); *Proc. Roy. Soc. B*, **64**, 449 (1951).
16. E. N. Leith and J. Upatnieks, "Reconstructed wavefronts and communication theory," *J. Opt. Soc. Am.* **52**, 1123 (1962); E. N. Leith and J. Upatnieks, "Wavefront reconstruction with continuous tone transparencies," *J. Opt. Soc. Am.* **53**, 522 (1963).
17. Y. N. Denisyuk, *Sov. Phys. Dok.* **7**, 543, (1962).
18. A. W. Lohmann and D. P. Paris, "Binary Fraunhofer hologram generated by computer," *Appl. Opt.* **6**, 1739–1748 (1967).
19. J. Durnin, "Exact solutions for nondiffracting beams," *J. Opt. Soc. Am. A* **4**, 651–654 (1987).
20. G. Indebetouw, "Nondiffracting optical fields: some remarks on their analysis and synthesis," *J. Opt. Soc. Am. A* **6**, 150–152 (1989).
21. R. Piestun, and J. Shamir, "Control of wavefront propagation with diffractive elements," *Opt. Lett.* **19**, 771–773 (1994)
22. J. Rosen and A. Yariv, "Snake beams: A paraxial arbitrary focal line," *Opt. Lett.* **20**, 2042–2044 (1995).
23. B. Spektor, R. Piestun, and J. Shamir, "Dark beams with a constant notch," *Opt. Lett.* **21**, 456–458 (also p. 911) (1996)
24. R. Piestun, B. Spektor, and J. Shamir, "Wave fields in three dimensions: Analysis and synthesis" *J. Opt. Soc. Am. A* **13**, 1837–1848 (1996).
25. R. Piestun, B. Spektor, and J. Shamir, "Unconventional light distributions in 3-D domains," *J. Mod. Opt.* **43**, 1495–1507 (1996).

26. Y. Y. Schechner, R. Piestun, and J. Shamir, "Wave propagation with rotating intensity distributions," *Phys. Rev. E* **54**, R50–R53 (1996).

27. R. Piestun, Y. Y. Schechner, and J. Shamir, "Propagation invariant wavefields with finite energy," *J. Opt. Soc. Am. A* (in press).

Three-Dimensional Holographic Imaging. Edited by Chung J. Kuo, Meng Hua Tsai
Copyright © 2002 John Wiley & Sons, Inc.
ISBNs: 0-471-35894-0 (Hardback); 0-471-22454-5 (Electronic)

CHAPTER 3

Self-Stabilized Real-Time Holographic Recording

LUCILA CESCATO and JAIME FREJLICH

Laboratório de Óptica, Instituto de Física Gleb Wataghin
UNICAMP, Cx. P. 6165
13083-970 Campinas, SP, Brasil

3.1 INTRODUCTION

Holographic exposures are used in a wide variety of applications from the three-dimensional imaging of objects to optical memories. By using the interference between two laser beams, it is possible to record holographic optical components, to perform phase conjugation experiments, and to precisely measure distances, displacements, or wave-front distortions. A simple sinusoidal interference pattern projected into a photosensitive material may also be used to study the optical changes induced in the material by light.

The major problem of holography is the low reproducibility or blurring due to the movement of the fringes, caused by changes in the optical path between the arms of the interferometers. These changes may be generated by mechanical vibrations of the optical components, external perturbations transmitted to the setup or thermal drifts in the air between the interfering beams.

Rapid exposures and interferometers with small arms reduce some of these effects, but many applications require illumination of large areas, resulting in low light intensities or long arm interferometers. In such cases, or even to obtain good reproducibility of the exposures, the only way is the use of an active fringe stabilization system that detects the fringe perturbation and produces a phase correction feedback signal to compensate for the perturbations.

Many systems proposed to correct fringe perturbations are now commercially available. Neumann and Rose [1] proposed the first stabilization system for holography. In their system the holographic pattern was amplified using a microscope objective and a photodetector was placed directly in the amplified fringe pattern for measurement of the fringe shifts. The electrical signal generated by the photodetector was amplified and used to feed back a phase-shifting device placed in one of the arms of the interferometer to compensate for the perturbations of the interference fringe pattern.

In 1976, Johanson et al. [2] proposed the use of a grating previously recorded in the holographic setup to amplify the fringe pattern instead of the objective. The superimposition of the actual interference fringe pattern with a previously recorded hologram results in a Moire-like fringe pattern that arises from the interference of one of the waves with its holographic reconstruction by the other wave. Any phase shift in the holographic microscopic interference pattern corresponds to the same fraction of the period shift in the Moire-like fringe pattern. A factor of 10^4 may be easily obtained, facilitating visual observation and instrumental detection of the interference pattern displacements.

In 1977, MacQuigg [3] improved upon the ideas of Neumann and Rose [1] and Johanson et al. [2] by introducing a small dither signal into the phase-shifting device of the feedback system, thus allowing the use of a tuned lock-in amplifier in the feedback loop. In this way it is not the intensity of the interference pattern that is measured but its first derivative. The system may be locked in at a bright or black fringe (zero derivative) without being perturbed by the background or continuous light level oscillations.

Another possibility of amplification of the microscopic holographic patterns used in the commercial system [4] is the use of a reference glass plate that interferes one part of the transmitted beam with the reflection of the other beam. If the angle between the transmitted beam and the reflected beam is very small, a large interference pattern will be formed behind the plate, resulting in high spatial amplification of the microscopic interference pattern.

Both detection methods, however, have the disadvantage that the detection is carried out at the reference hologram or at the glass plate and not at the point where the hologram is being recorded. In 1988 [5], the closed-loop phase stabilization system proposed by MacQuigg [3] was analyzed using concepts of wave mixing to explain the Moire-like fringe pattern. The wave-mixing analysis allowed the establishment of a relation between the Moire-like fringe pattern, the microscopic holographic fringes, and the reference hologram. The knowledge of this relation brought new possibilities for the application of this system. The most important of these possibilities is the use of the hologram that is being recorded as a reference to stabilize the exposure. We call this process self-stabilized holographic recording.

The performance achieved in this feedback system is so high that even small changes in the optical constants of the photosensitive material induced by light during the exposure may be used as a reference hologram to operate the stabilization system [6].

This fringe stabilization system will be described in detail in the next section, and in the Section 3.3 it will be applied to self-stabilize holographic recording in different photosensitive materials.

3.2 FRINGE STABILIZATION SYSTEM

The fringe stabilization system is composed of a detection system that uses wave mixing to amplify the holographic fringes, an electronic system that provides the synchronous detection and the amplification, and a phase modulator that realizes the correction feedback in the phase. The phase modulator may also produce the phase dither for the synchronous detection. A block diagram of the system is shown in Figure 3.1.

3.2.1 Holographic Setup

There are many types of interferometers for the generation of holographic fringes. The only requirement for using a fringe stabilization system is the presence of one active phase shift element in one of the arms of the interferometer. This element provides the reference signal for the synchronous detection and the phase feedback control. In our case this element is a piezoelectric-supported mirror. By changing the high voltage applied on the piezoelectric crystals, the mirror moves an amount linearly proportional to the voltage. Other active phase modulators, as for example electro-optic crystals, may be used, but the phase changes produced by this device usually do not provide the large phase shift required to compensate for the external

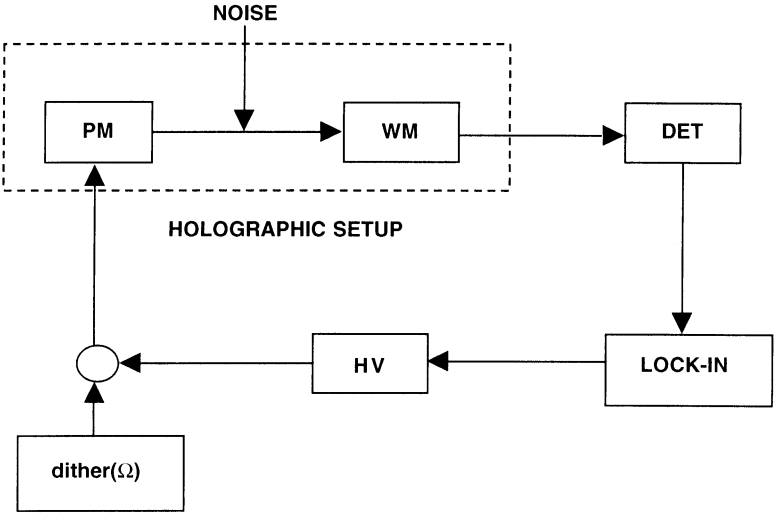


Figure 3.1 Block diagram of fringe stabilization system. Holographic setup generates fringe pattern; wave mixing (WM) amplifies fringe pattern; photodetector (DET) and lock-in amplifier detect perturbations and feed back active phase modulator (PM) through its high-voltage source (HV). Same phase modulator is fed by external oscillator that furnishes dither or reference signal (Ω) for synchronous detection.

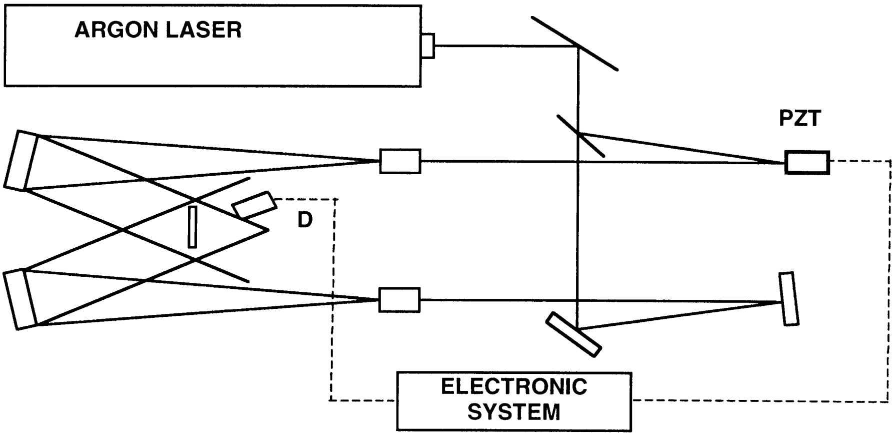


Figure 3.2 Schema of holographic setup and fringe stabilization system. Laser beam is divided, expanded, and collimated, generating interference pattern. Holographic material is placed in interference fringe region. Detector, placed behind holographic plate, measures wave-mixing signal by electronic system. Electric signal is amplified and fed back to holographic setup through piezoelectric-supported mirror (PZT).

perturbations. Figure 3.2 shows a scheme of one holographic setup together with the detection and feedback system. The system uses the Moire-like pattern formed between the transmitted beam and the diffraction of the second beam in the hologram to amplify the microscopic fringe pattern.

An analysis of the wave mixing using the reflected waves in the hologram instead of the transmitted waves has been described recently [7]. This system allows self-stabilization in photosensitive films on nontransparent substrates such as semiconductors or metals.

3.2.2 Wave Mixing

In the region of the intersection between two interfering beams, plane-parallel fringes will be formed in the direction of the bisector of the interfering beams. Figure 3.3 shows a cross section of such fringes together with a hologram (grating) recorded in the same fringe pattern.

Assuming that Σ_1 and Σ_2 are the interfering waves, the hologram generates two more waves Σ'_1 and Σ'_2 that can be thought of as the reconstructed wave fronts or diffracted waves. Behind the hologram, in the direction of Σ_1 we have two waves: the transmitted wave Σ_1 and the reconstructed wave Σ'_1, where Σ'_1 is the holographic reconstruction of the wave Σ_1 realized by the reference wave Σ_2. The same occurs in the direction of Σ_2.

The wave mixing between each pair of transmitted and diffracted (reconstructed) waves generates a Moire-like pattern, as can be seen in Figure 3.4.

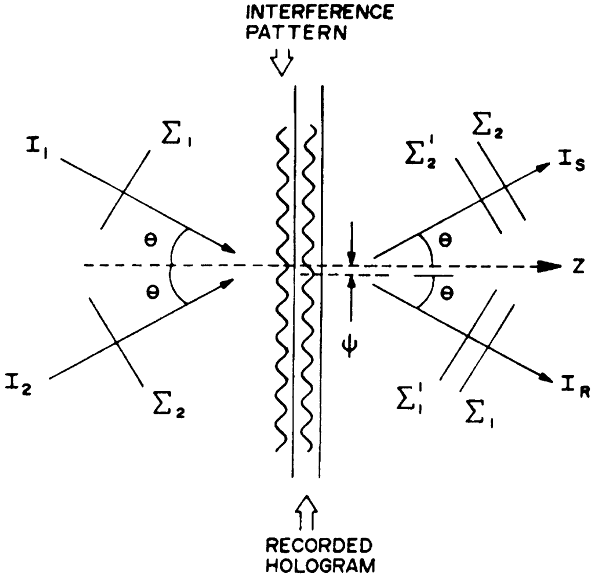


Figure 3.3 Schema of the wave mixing between the transmitted wave and the reconstructed (diffracted) wave.

The misalignment between the interference fringe and the recorded grating produces a Moire interference pattern whose period increases with the matching of the gratings. Any phase perturbation ψ in the optical path between the arms of the interferometer shifts the interference pattern of the same phase ψ in relation to the recorded hologram. The same phase difference appears between the waves Σ_1 and Σ_1' and between the waves Σ_2 and Σ_2'.

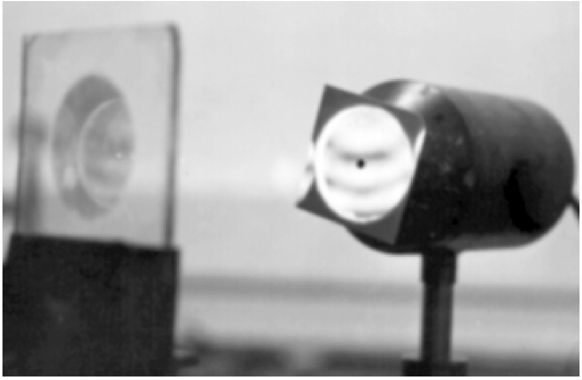


Figure 3.4 Moire-like pattern projected on the detector.

Let the hologram be described by a complex refractive index modulation:

$$n = n_0 + n_1 \cos(Kx) \tag{3.1}$$

where $K = 2\pi/d$, x is the coordinate perpendicular to the pattern of fringes, n_0 is the average complex refractive index, and n_1 is the amplitude of the complex refractive index modulation. Assuming also that the interference pattern is ψ-phase shifted in relation to the hologram (grating) and that I_1 and I_2 are the intensities of the two parallel polarized coherent interfering waves Σ_1 and Σ_2 respectively, the interference pattern, in relation to the reference hologram, may be represented by

$$I = I_1 + I_2 + 2\sqrt{I_1 I_2}\cos(Kx - \psi) \tag{3.2}$$

In this case, the intensities I_R and I_S in the directions of the interfering beams behind the hologram may be described by [5]

$$I_R = \eta_0 I_1 + \eta_1 I_2 - 2\sqrt{\eta_0 \eta_1 I_1 I_2}\cos(\varphi) \tag{3.3}$$

$$I_S = \eta_1 I_1 + \eta_0 I_2 + 2\sqrt{\eta_0 \eta_1 I_1 I_2}\cos(\varphi) \tag{3.4}$$

where η_1 is the diffraction efficiency of the first-diffracted order by transmission and η_0 is the diffraction efficiency of the zero-diffracted order; φ is the phase difference between the interfering waves Σ_1 and Σ_1' (first- and the zero-diffracted waves) in the direction of I_R or between Σ_2 and Σ_2' in the direction of I_S. Note that φ may be different from ψ by a factor that depends on the mechanism of recording and of the photosensitive material. For amplitude-recording materials $\varphi \equiv \psi$, while for phase-recording materials there is an additional phase shift of $\frac{1}{2}\pi$ between the zero- and first-diffracted waves; thus $\varphi \equiv \psi + \frac{1}{2}\pi$.

3.2.3 Synchronous Detection

If a low-amplitude, high-frequency phase dither is injected in the system through the phase modulator [piezoelectric (PZT) supported mirror], by adding to the high-voltage supply of the PZT an alternating current (ac) voltage of frequency Ω and amplitude v_d, the phase φ in Eqs. 3.3 and 3.4 may be replaced by

$$\varphi + \varphi_d \sin(\Omega t) \tag{3.5}$$

with

$$\varphi_d = K_{\mathrm{PZT}}^{\Omega} v_d \ll 1 \tag{3.6}$$

where $K_{\mathrm{PZT}}^{\Omega}$ represents the overall PZT-driven mirror voltage-to-phase conversion factor at the frequency Ω. Thus Eq. 3.3 becomes

$$I_R = \eta_0 I_1 + \eta_1 I_2 - 2\sqrt{\eta_0 \eta_1 I_1 I_2}\cos[\varphi + \varphi_d \sin(\Omega t)] \tag{3.7}$$

Using the equalities for $\sin[\varphi_d \sin(\Omega t)]$ and $\cos[\varphi_d \sin(\Omega t)]$ in terms of Bessel functions, we may develop

$$\cos[\varphi + \varphi_d \sin(\Omega t)] = \cos(\varphi)\left[J_0(\varphi_d) + 2\sum_{n=1}^{\infty} J_{2n}(\varphi_d)\cos(2n\Omega t)\right] - \sin(\varphi)\left[2\sum_{n=0}^{\infty} J_{2n+1}(\varphi_d)\sin[(2n+1)\Omega t]\right] \tag{3.8}$$

where $J_i(\varphi_d)$ is the Bessel function of order i.

Thus Eq. (3.7) may be developed in a harmonic series of fundamental frequency Ω:

$$I_R = I_R^{\mathrm{dc}} + I_\Omega + I_{2\Omega} + I_{3\Omega} + \cdots \tag{3.9}$$

where

$$I_R^{\mathrm{dc}} = \eta_0 I_1 + \eta_1 I_2 - 2\sqrt{\eta_0\eta_1 I_1 I_2}\cos(\varphi)J_0(\varphi_d) \tag{3.10}$$

$$I_\Omega(t) = -4J_1(\varphi_d)\sqrt{\eta_0\eta_1 I_1 I_2}\sin(\varphi)\sin(\Omega t) \tag{3.11}$$

$$I_{2\Omega}(t) = 4J_2(\varphi_d)\sqrt{\eta_0\eta_1 I_1 I_2}\cos(\varphi)\cos(2\Omega t) \tag{3.12}$$

These light intensity harmonics of the dither signal Ω may be better understood with the aid of Figure 3.5. The cosine curve represents the Moire-like pattern in each of the directions I_R or I_S. If the detector is set in a small part of the fringe pattern (as in Fig. 3.4), the effect of the dither phase signal is to produce small movements of the fringe pattern, generating harmonics in the light intensity. The first-harmonic signal has maximum amplitude at the linear part of the fringe pattern (region of maximum derivative as a function of φ) while the maximum of the second-harmonic signal will be in the dark or bright fringes of the interference pattern.

The signal from the photodetector, being proportional to the light intensity, contains all the harmonic terms of the dither signal. If this signal is measured through a lock-in amplifier, we can select a voltage signal proportional to the amplitude of the first or second harmonic of the light intensity (V_Ω and $V_{2\Omega}$, respectively).

3.2.4 Feedback Optoelectronic Loop and Fringe Stabilization

Both harmonic signals described by Eqs. 3.11 and 3.12 carried information about the phase shift φ between the interfering beams behind the hologram. This phase shift is directly related to the phase shift ψ (between the interference pattern and the hologram), which represents the phase perturbations in the holographic setup. Thus either harmonic signal or a combination of the signals may be used as an error signal for operating the feedback loop.

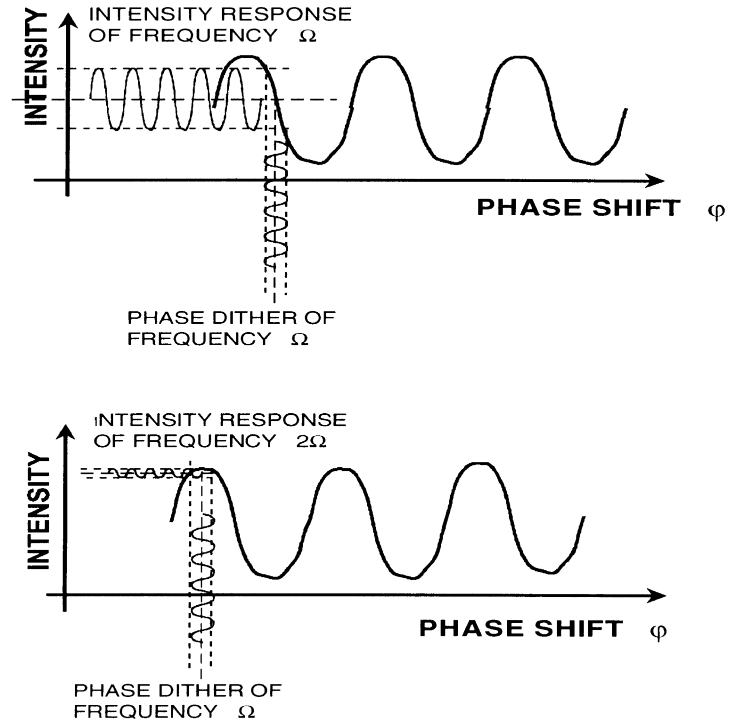

INTENSITY RESPONSE OF FREQUENCY Ω
INTENSITY
PHASE SHIFT φ
PHASE DITHER OF FREQUENCY Ω
INTENSITY RESPONSE OF FREQUENCY 2Ω
INTENSITY
PHASE SHIFT φ
PHASE DITHER OF FREQUENCY Ω

Figure 3.5 Generation of harmonics in the holographic fringe pattern by the phase dither signal: (*a*) position of maximum amplitude of the first-harmonic signal; (*b*) position of maximum amplitude of second-harmonic signal.

3.2.4.1 Proportional Single-Harmonic Feedback The operation of the actively stabilized setup using either the first or second harmonic is schematically described in the block diagram of Figure 3.1. The output voltage of the lock-in amplifier tuned into the first harmonic (V_Ω) or second harmonic ($V_{2\Omega}$) may be amplified and fed back to the holographic system through the source of the phase modulator (PZT-supported mirror).

First-Harmonic Feedback. If the first-harmonic signal (V_Ω) is used, and accounting that the PZT also needs a constant electric bias V_0, the total phase shift furnished by the PZT φ_{PZT} (neglecting the phase dither) may be described by

$$\varphi_{\mathrm{PZT}} = K_{\mathrm{PZT}}[V_0 + A\sin(\varphi)] \tag{3.13}$$

where K^0_{PZT} is the same overall voltage-to-phase K^Ω_{PZT} but for $\Omega = 0$, and

$$A = K_0 K_l K_p[-4J_1(\varphi_d)\sqrt{\eta_0 \eta_1 I_1 I_2}] \tag{3.14}$$

with K_0 and K_l being the amplification factors for the high-voltage supply of the PZT and lock-in amplifier, respectively, and K_p the photodetector response or conversion light–voltage factor.

Assuming that the phase shift between the interfering waves φ (neglecting the dither) is now given by

$$\varphi = \varphi_N + \varphi_{\text{PZT}} \tag{3.15}$$

with φ_N being any phase shift or noise caused by phase perturbations in the holographic setup and φ_{PZT} the phase shift furnished by the PZT, substituting φ_{PZT} from Eq. 3.13 into Eq. 3.15, the equilibrium condition is found by

$$\frac{\varphi - \varphi_N - K^0_{\text{PZT}} V_0}{K^0_{\text{PZT}} A} = \sin(\varphi) \tag{3.16}$$

For large amplification (large A factor) this occurs for values of $\sin(\varphi) \cong 0$, resulting in $\varphi \cong N\pi$. Substituting these values of φ in Eq. 3.3, I_R will be a maximum (bright fringe) or minimum (dark fringe) depending on N being odd or even. This condition may be chosen by changing the signal of the amplification factor (or easily by inverting the phase 180° in the lock-in amplifier). If I_R is in a bright fringe on the detector, the equivalent fringe position in the Moire pattern in the other beam direction (I_S) will be a dark fringe. This can be seen from Eq. 3.4 and satisfies the energy conservation principle. This situation corresponds to the maximal wave coupling between the interfering beams.

Second Harmonic Feedback. If the PZT-supported mirror is fed back with the second-harmonic voltage ($V_{2\Omega}$) output of the lock-in amplifier, Eq. 3.13 will be substituted by

$$\varphi_{\text{PZT}} = K^0_{\text{PZT}}[V_0 + B\cos(\varphi)] \tag{3.17}$$

where

$$B = K_0 K_l K_p [4J_2(\varphi_d)\sqrt{\eta_0 \eta_1 I_1 I_2}] \tag{3.18}$$

Then, the equilibrium in Eq. 3.16 will be substituted by

$$\frac{\varphi - \varphi_N - \text{K}^0_{\text{PZT}} V_0}{K^0_{\text{PZT}} B} = \cos(\varphi) \tag{3.19}$$

As $\psi_d \ll 1$, $B \ll A$ and a poorer performance of the stabilization system should be expected. In practice, this does not occur because the lower amplitude of the signal is somewhat compensated by the lower noise from the second-harmonic detection.

In this case the equilibrium will be reached when $\cos(\varphi) \cong 0$ or $\varphi \cong \frac{1}{2}(2N + 1)\pi$. This corresponds to null energy wave coupling or no energy exchange between the wave beams.

3.2.4.2 Integral Harmonic Feedback

In first- or second-harmonic feedback, described above, the correction signal is linearly proportional to the amplitude of the first or second harmonic with amplification constants A and B, respectively. If there is a continuous phase shift drift during the exposition (φ_N increases continuously during the exposure), the PZT-supported mirror must be shifted continuously to compensate for the fringe drift. In this case, as the amplification is not infinity, φ will shift further away continuously from its initial value of zero or $\frac{1}{2}\pi$ to provide the necessary voltage for the PZT-supported mirror. To correct this phase shift, V_0 must be manually changed during the exposure to compensate for the drift.

In this case, more efficient performance of the stabilization system may be obtained if an integral feedback is used by introducing a simple integrator device between the lock-in amplifier output and the PZT power supply. The effect of this integrator in the feedback loop may be understood by substituting the linear time-independent terms $A\sin(\varphi)$ and $B\cos(\varphi)$ in Eqs. 3.13 and 3.17, respectively, by the integral voltages

$$\frac{A}{\tau_i}\int_0^t \sin(\varphi)\,dt \tag{3.20}$$

$$\frac{B}{\tau_i}\int_0^t \cos(\varphi)\,dt \tag{3.21}$$

with τ_i being the time constant of the integrator device and the t the time elapsed since the instant that the loop is closed ($t = 0$).

In the integral feedback the correction signal may be much greater than the error signal, keeping the phase φ close to its initial value.

The necessary amplification may be achieved by increasing the amplification factors A or B or by decreasing the integration time τ_i.

This integral feedback is particularly interesting for compensating large and slowly varying perturbations like temperature drift and air current drafts, allowing the stabilization of nonstationary holograms.

3.2.4.3 Self-Stabilized Recording

The possibility of choosing the conditions of operation of the feedback loop (first- or second-harmonic detection) allows the choice of the phase φ for the values 0, π, or $\pm\frac{1}{2}\pi$. Although it seems somewhat limited, it allows the operation of most practical situations of self-stabilized recording, as will be described below.

To use the grating being recorded in a permanent material as a reference for the fringe stabilization (self-stabilized recording), the interference pattern must be in phase with the optical modulation being formed (ψ must be zero). If this condition is not satisfied, the grating will be moving continuously until complete homogeneous exposition of the material. For phase modulation materials as photoresist films or photothermoplastics, there is an additional phase shift of $\frac{1}{2}\pi$ between the first and the zero diffracted ($\varphi = \psi + \frac{1}{2}\pi$). Thus, if ψ must be zero, φ must be $\frac{1}{2}\pi$ requiring the use of the second-harmonic signal to feed back the stabilized real-time recording. Photocromic materials present no phase difference between the diffracted and transmitted waves, and thus $\varphi = \psi$ (and ψ must be zero), requiring the use of the first harmonic to feed back the stabilization system.

To record holograms in real-time materials, as for example photorefractive crystals, self-stabilized recording may be used in any situation because a nonstationary grating may be recorded. The hologram will be stationary, however, only for certain phase shifts, which depend on the mechanism of the phase modulation of the crystal. For electro-optic crystals in the diffusion regime, there is a phase shift of $\frac{1}{2}\pi$ between the interference pattern and the refractive index modulation generated in the crystal [8]; then a first-harmonic feedback is necessary to obtain a stationary hologram. In the presence of external electrical fields, in the drift regime this phase shift depends on the amplitude of the external electrical field [8], so stationary holograms may not be achieved with either first- or second-harmonic feedback.

3.2.4.4 Fringe Lock with Arbitrary Phase φ

To record self-stabilized stationary holograms in photorefractive crystals in the presence of an external electrical field [8], or even to self-stabilize holograms using the reflected waves [7], feedback using single-harmonic signals does not work because the phase shift φ between the interfering waves must be different from 0, π, or $\pm\frac{1}{2}\pi$.

By adequately processing and combining the first- and second-harmonic signals, before feedback of the PZT-supported mirror, it is possible also to lock the fringe pattern with an arbitrary phase shift φ. A discussion of this processing, described in detail in a recent paper [9], is presented below.

The electrical signal measured directly by the photodetector contains all harmonics of the dither frequency Ω. The second-harmonic electrical signal may be represented by

$$V_{2\Omega}(t) = 4k_2 J_2(\varphi_d)\sqrt{\eta_0\eta_1 I_1 I_2}\cos(\varphi)\cos(2\Omega t) \tag{3.22}$$

which is directly proportional to the light intensity $I_{2\Omega}(t)$ of Eq. 3.12.

The first-harmonic signal, proportional to light intensity $I_{\Omega}(t)$ of Eq. 3.11, is separated from I_{R} (Eq. 3.9) using a bandpass filter. After the bandpass filter, the first-harmonic signal is frequency doubled, phase shifted in relation to the second-harmonic signal (Eq. 3.22), and amplified to generate a new electrical second-harmonic signal:

$$V_{\Omega 2}(t) = -4k_1 J_1(\varphi_d)\sqrt{\eta_0\eta_1 I_1 I_2}\sin(\varphi)\sin(2\Omega t + \delta) \tag{3.23}$$

The time phase δ and the amplification are chosen to generate a second-harmonic signal $V_{\Omega 2}$ with the same amplitude and $\frac{1}{2}\pi$ shifted in relation to the original second-harmonic signal $V_{2\Omega}$. The signals $V_{2\Omega}$ and $V_{\Omega 2}$ are then added, generating a unique second-harmonic signal:

$$V_{+}(t) = V_0 \cos(2\Omega t + \varphi) \tag{3.24}$$

If the $V_{+}(t)$ signal is now measured using a phase-sensitive lock-in amplifier tuned in 2Ω, the V_X and V_Y output of the lock-in amplifier will be given by

$$V_X = V_0 \cos(\varphi + \theta_S) \tag{3.25}$$

$$V_Y = V_0 \sin(\varphi + \theta_S) \tag{3.26}$$

The phase θ_S is an arbitrary value chosen for the reference signal coordinate system in the lock-in amplifier. If the V_Y signal is used to feed back the loop, $\sin(\varphi + \theta_S)$ will be kept at zero; thus, φ will be kept at the value $\varphi = -\theta_S$, which can be chosen arbitrarily in the lock-in amplifier.

3.2.5 Simultaneous Stabilization and Monitoring

Observing Eqs. 3.11 and 3.12, it can be seen that both first- and second-harmonic signals are just $\frac{1}{2}\pi$ shifted in φ. If one harmonic is used to stabilize the interference pattern, the other harmonic will be kept at maximal. If both harmonics of the light signal are measured simultaneously by using two lock-in amplifiers, one can be used to feed back the stabilization system while the other may be used to monitor the stabilization. If we use a previously recorded reference grating, the harmonic used to stabilize the pattern will be kept null while the other (monitoring signal) will stay fixed at a certain value.

If the reference grating is being formed during the recording, both harmonics will increase because the grating efficiency (η_1) will increase with exposure time (energy). Assuming that the optical modulation of the grating that is forming is very low, $\eta_1 \ll 1$ and the diffraction efficiency of the zero transmitted order η_0 may be considered approximately ≈ 1 in both Eqs. 3.11 and 3.12. The signals V_Ω and $V_{2\Omega}$ measured by two lock-in amplifiers tuned at the first (Ω) and second (2Ω) harmonic respectively will be given by

$$V_\Omega = K_p 4 J_1(\varphi_d) \sqrt{I_1 I_2} \sqrt{\eta_1} \sin(\varphi) \tag{3.27}$$

$$V_{2\Omega} = -K_p 4 J_2(\varphi_d) \sqrt{I_1 I_2} \sqrt{\eta_1} \cos(\varphi) \tag{3.28}$$

If one of the harmonics is used to feed back the system, the other will follow the evolution of the diffraction efficiency of the grating being formed. Note that for very small diffraction efficiencies, the square root can be much greater than the diffraction efficiency itself, making the measurement very sensitive. If the optical recording mechanism is known, it is possible to model this evolution and to calculate parameters of the material or the reaction that is induced by the light.

For high diffraction efficiencies $\sqrt{\eta_1\eta_0}$ cannot be approached by $\sqrt{\eta_1}$ in Eqs. 3.27 and 3.28, but the signal evolution also may be used to study the recording mechanism of the material. If the period of the grating is small enough to have only first- and zero-diffracted orders, $\sqrt{\eta_1\eta_0}$ may be approached by $\sqrt{\eta(1-\eta)}$, where η is the diffraction efficiency of the first-diffracted order and $1-\eta$ is the diffraction efficiency of the zero-diffracted order.

3.3 APPLICATIONS

Self-stabilized holographic recording is a very effective technique for realizing high contrast and reproducible holographic exposures, and it may be used in different photosensitive materials, even real-time or permanent materials that need postexposure processing. The unique requirement is the presence of any weak real-time spatial modulation induced by light. Real-time diffraction efficiencies as small as 10^{-6} are enough to self-stabilize the whole holographic exposure. In this section we will describe self-stabilized recording in different photosensitive materials to illustrate the possibilities and perspectives of this new technique.

3.3.1 Self-Stabilized Holographic Recording in Photoresist Films

Photoresist films are materials widely employed in microelectronic processes. After exposure in an appropriate light pattern, the photoresist is developed in a wet solution to convert the light pattern into a relief mask for the etching of the substrate [10]. Photoresist is used in holography to record relief holograms for embossing and replication [11] or as masks for lithography of holographic optical elements.

Residual real-time refractive index modulation has been observed in both negative and positive photoresist [6, 12]. Diffraction efficiencies of about 10^{-4} can be measured, even without development, in positive photoresist films of 1μm thickness exposed at the line $\lambda = 458$ nm of an Ar laser [13]. Although this index modulation could be very low at the beginning of exposure, the stabilization system is so sensitive that it can be used as a reference hologram to stabilize the entire holographic exposition of the photoresist.

In photoresist films, real-time modulation in the refractive index generates a phase hologram [13]. Thus, as discussed in Section 3.2.2, $\varphi \equiv \psi + \frac{1}{2}\pi$, and to keep $\psi = 0$, φ must be fixed at $\frac{1}{2}\pi$. In this case, the second-harmonic signal $V_{2\Omega}$ must be used to feed back the piezoelectric-supported mirror for self-stabilization of the holographic exposure, imposing $\cos(\varphi) \cong 0$. Then, the

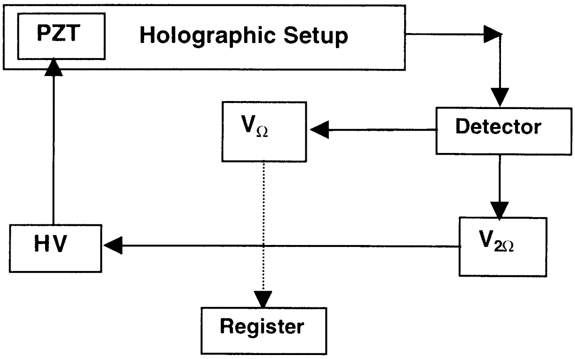


Figure 3.6 Flow chart of the stabilization and monitoring system of self-stabilized holographic recording in photoresist film.

first-harmonic signal will be given by

$$V_\Omega = K_p 4J_1(\varphi_d)\sqrt{I_1 I_2}(\sqrt{\eta_1} \qquad 3.29)$$

The diffraction efficiency η_1 depends on the refractive index modulation of the photoresist grating that is forming and thus will increase during the exposure. So the signal V_Ω should increase continuously during the exposure and could be used to monitor the exposure. Figure 3.6 shows this procedure in a flow chart.

For comparison, Figure 3.7 shows the time evolution of the V_Ω signals of self-stabilized and nonstabilized holographic exposures in AZ 1400 photoresist films of 1 μm thickness on glass substrates. Both films were exposed to the same holographic pattern of period 0.83 μm using the 458-nm line Ar laser and beam irradiances of 50 μW/cm^2 for each beam. The peaks in the nonstabilized V_Ω evolution represent phase perturbations in the holographic setup. Note that, as a consequence of these perturbations, the contrast of the fringe pattern is reduced, resulting in a lower refractive index modulations and lower maximum value of V_Ω.

The effect of these phase perturbations on the profile recorded in the photoresist after development depends on several parameters, for example, photoresist type and development process [14, 15]. The general effect, however, is to reduce the depth in the photoresist recorded grating and to increase the roughness of the developed photoresist surfaces. If the development time is increased to compensate for the contrast reduction of the fringe pattern, a strong narrowing in the peaks of the structures will be observed.

This effect can be seen in scanning electron microscopy (SEM) photographs of the cross section of photoresist gratings (see Fig. 3.8).

This narrowing of the peaks is caused by the isotropy of the wet development and may lead to complete collapse of the structure.

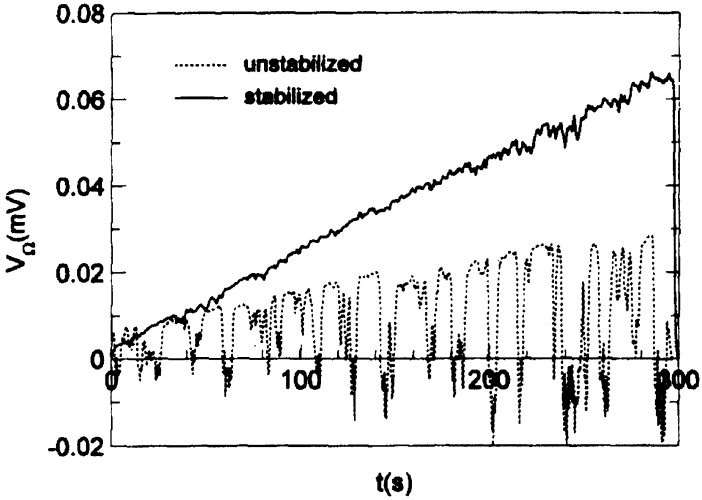


Figure 3.7 First-harmonic V_Ω time evolution during holographic recording of two photoresist gratings with and without stabilization. The AZ 1400 photoresist films have 1 μm thickness and exposures were performed in 458-nm line Ar laser with irradiance of about 50 μW/cm^2.

These results show that self-stabilization during the recording using residual real-time index modulation is very effective and fundamental to produce deep relief photoresist structures. Figure 3.9 shows a high-aspect-ratio mask profile recorded in a photoresist film using this technique, with its respective V_Ω evolution.

In addition to the good results obtained in the self-stabilized recording of photoresist masks, the V_Ω signal is related to the diffraction efficiency of the grating that is forming. Thus, by using a proper diffraction theory, it is possible to relate the diffraction efficiency with the refractive index chances of the material. By modeling the dependence of the refractive index changes with the light, it is possible to study the photoreaction mechanism and to measure parameters of the photoreaction [13].

3.3.2 Self-Stabilized Photoelectrochemical Etching of *n*-InP(100) Substrates

Photoelectrochemical (PEC) etching is an interesting technique to direct engraving a light pattern in semiconductors [16]. In PEC etching, the photogenerated minority carriers in a semiconductor are drawn by the reverse applied potential to the semiconductor–eletrolyte interface, where they catalyze the anodic dissolution of the semiconductor. Thus the regions of the semiconductor surface that were illuminated will be etched, converting a light pattern into a relief pattern engraved in the semiconductor surface.

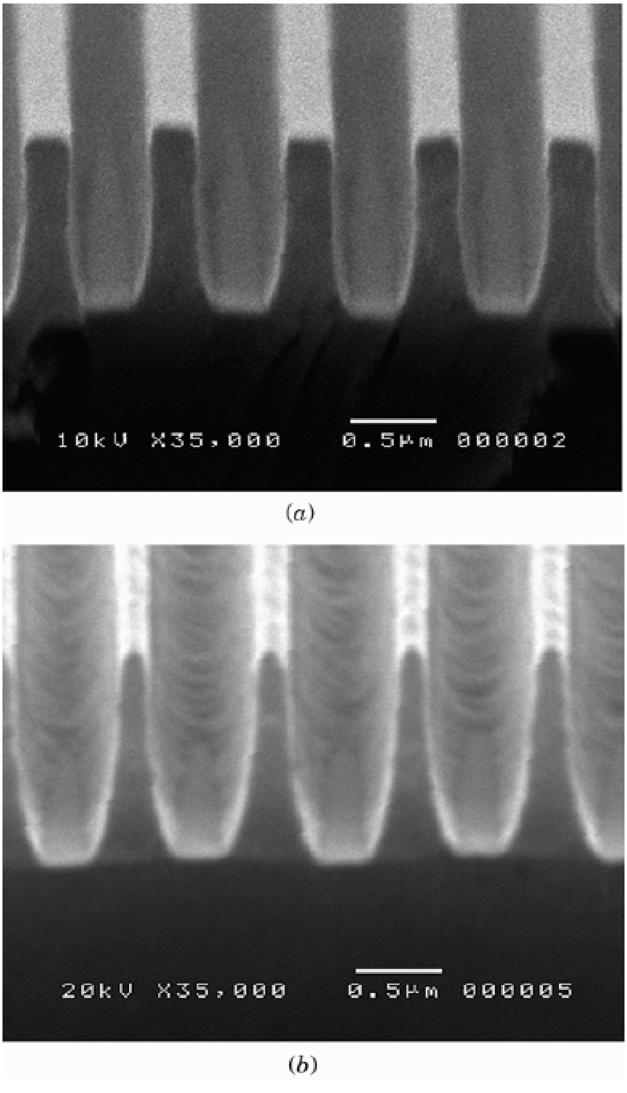
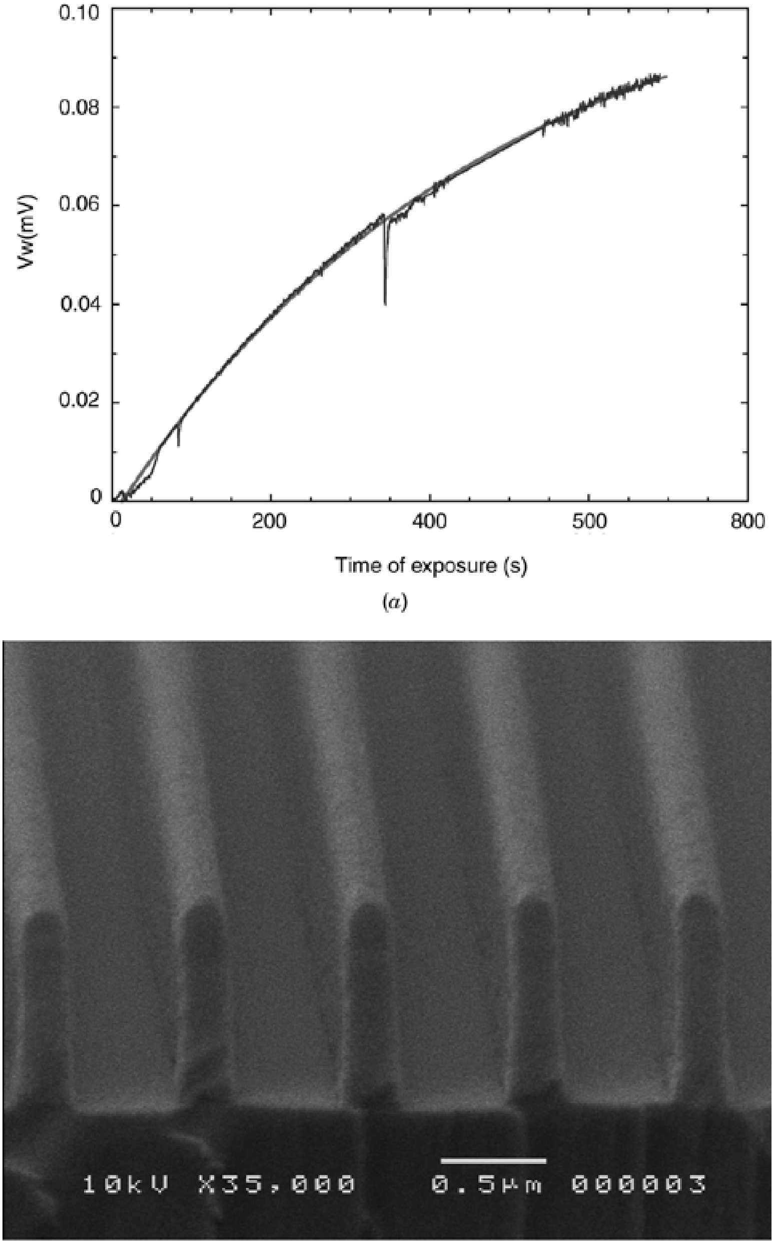


Figure 3.8 SEM photographs of cross section of photoresist gratings recorded holographically in AZ 1400-17 with same exposure energy (600 mJ/cm^2) developed in AZ 400K 1 : 4: (*a*) without phase perturbations developed during 40 s; (*b*) in presence of phase perturbations, developed in 55 s.

The major problem that limits the application of this technique to holographic recordings is the low contrast of the interference pattern due to vibrations and thermal and concentration gradients, because the holographic pattern must be projected inside a liquid cell (electrolyte). Figure 3.10 shows the PEC etching of a holographic grating in *n*-InP semiconductor substrate.



(*a*)

(*b*)

Figure 3.9 Mask recorded in AZ 1400-17 photoresist exposed to energy of 600 mJ/cm^2 and developed in AZ400K 1:4 during 45 s. (*a*) first-harmonic evolution during holographic exposure; (*b*) SEM photograph of cross section of resulting mask profile.

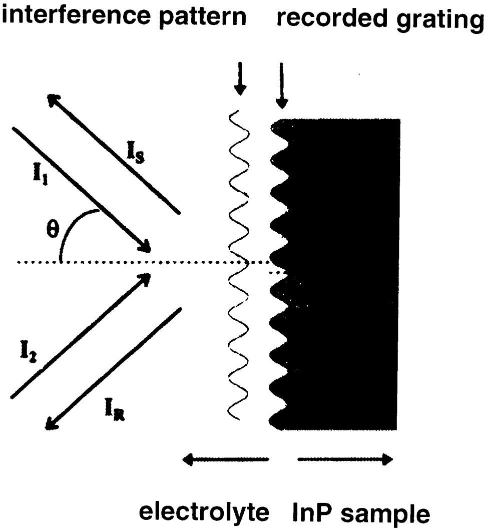


Figure 3.10 Schema of PEC etching of *n*-InP sample. Holograhic interference pattern is projected in liquid electrolyte.

To solve this problem, the grating that is being etched in relief on the surface of the *n*-InP sample can be used as a reference hologram to self-stabilize the holographic recording [17]. As the semiconductor crystal is nontransparent for the visible wavelengths, the wave mixing between the reflected waves must be used. In this case, in each of the directions I_R and I_S of the reflected beams the same relations 3.3 and 3.4 may be applied.

As the relief grating produces a phase modulation in the diffracted waves, $\varphi \equiv \psi + \frac{1}{2}\pi$, and the second harmonic must be used to self-stabilize the recording to keep the interference pattern matched with the grating ($\psi = 0$). In this case, similar to the self-stabilized recording in photoresist films, the second harmonic can be used to monitor the recording or to measure the etching evolution.

Figure 3.11 shows examples of the V_Ω evolution curve as a function of time for stabilized and nonstabilized PEC etching realized under similar conditions.

As can be seen, the self-stabilization is very effective for the PEC etching. This fact is confirmed by SEM views of the samples. The grating corresponding to the nonstabilized recording is even difficult to be observed.

The fundamental difference between the self-stabilized PEC etching and the self-stabilizing recording in photoresist films is that in the former the diffraction efficiencies achieved are much higher. Then the $\sqrt{\eta_0\eta_1}$ may not be approached by $\sqrt{\eta_1}$ in Eqs. 3.27 and 3.28 and the evolution of the V_Ω signal is different, as can be seen in Figure 3.12.

The maximum in the V_Ω signal corresponds to the maximum at $\sqrt{\eta_0\eta_1}$ because when the diffraction efficiency of the first-diffracted order (η_1) in-

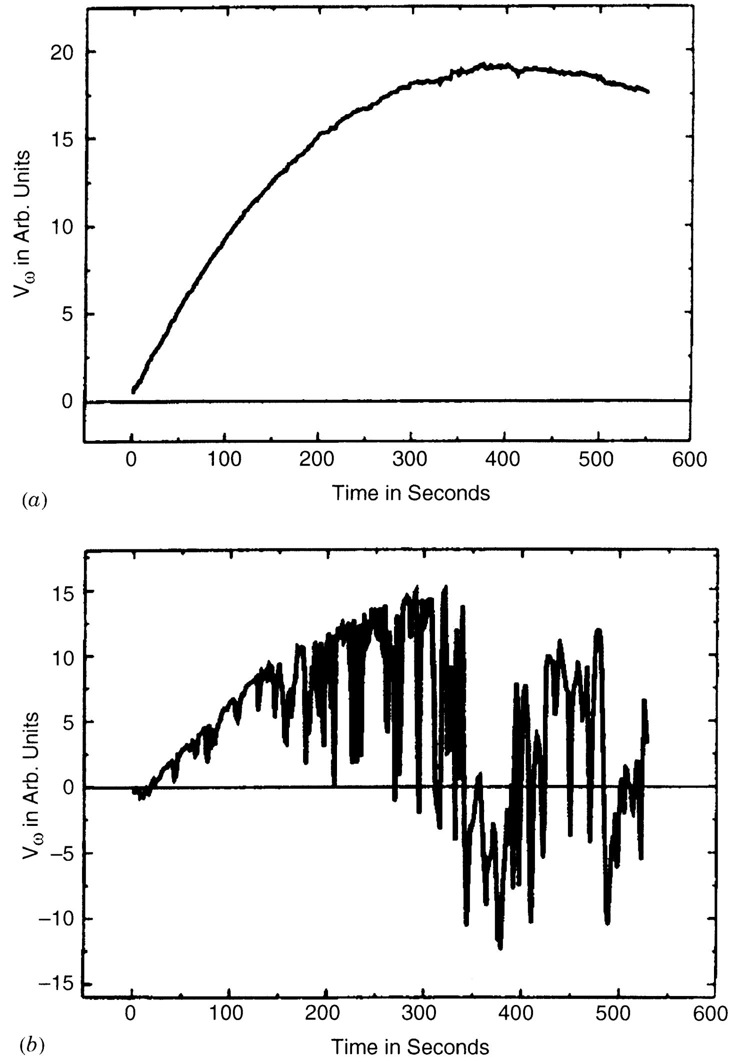


Figure 3.11 A V_Ω (evolution curve as function of the time for gratings with period of $0.4\,\mu$m etched at current of $800\,\mu$A/cm^2 and for same final removed charge density of 100 mC/cm^2: (*a*) stabilized and (*b*) without stabilization. Both samples are aligned with ⟨0-11⟩ crystal axis parallel to lines of interference pattern.

creases, the diffraction efficiency of the zero-reflected order (η_0) must decrease to keep constant the total energy. When η_0 approaches zero, both signals, V_Ω and $V_{2\Omega}$ approach zero. When this occurs, the feedback loop must be turned off and its signal must be changed to compensate for the phase change between the interfering waves [17].

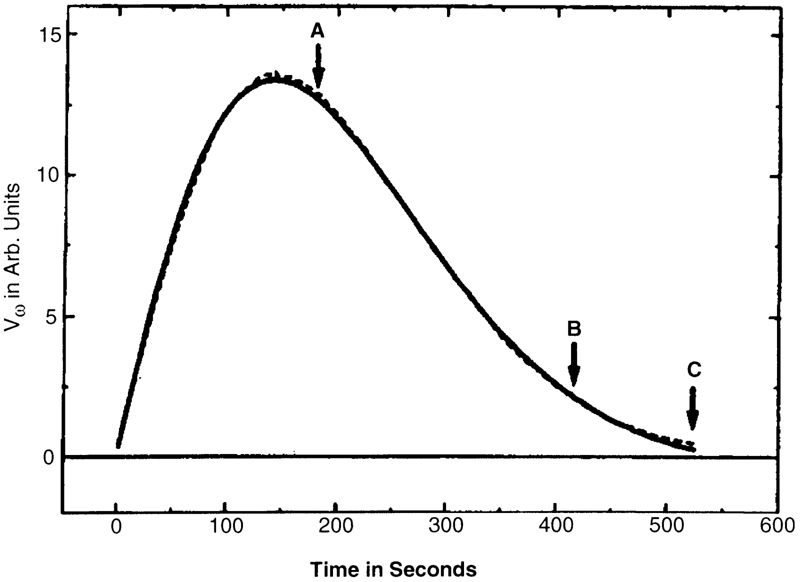


Figure 3.12 A V_Ω signal time for PEC etching of surface (100) of *n*-InP with grating lines aligned parallel to axis ⟨0-11⟩.

Figure 3.13 shows SEM photographs of the cross sections of two gratings of 0.4 μm of period recorded in the surface (100) of *n*-InP by PEC etching using two different orientations of the sample in relation to the interference pattern.

Note that, for the etching along the interference fringes aligned in the direction $\langle 0\bar{1}1 \rangle$, the indium stop plans are revealed giving the grooves a triangular form (Fig. 3.13*a*). For the orthogonal direction (Fig. 3.13*b*) the profile is rounded. As can be seen from the same figures, the etched depth is limited for the triangular profile.

3.3.3 Self-Stabilized Holographic Recording in Photorefractive Crystals

Holographic recording in real-time reversible photosensitive materials has particular features and potential applications in many scientific and technological fields.

Photorefractives [8, 18] are known to exhibit real-time reversible recording properties with the additional advantage of unlimited number of recording/ erasure cycles. Lithium niobate, barium titanate, bismuth and silicium oxides, gallium arsenide, and indium phosphide, among many others, are well-known photorefractive crystals exhibiting widely different properties in terms of wavelength and intensity sensitivities, recording time constant, darkness storage possibilities, maximum diffraction efficiency, and associated effects such as

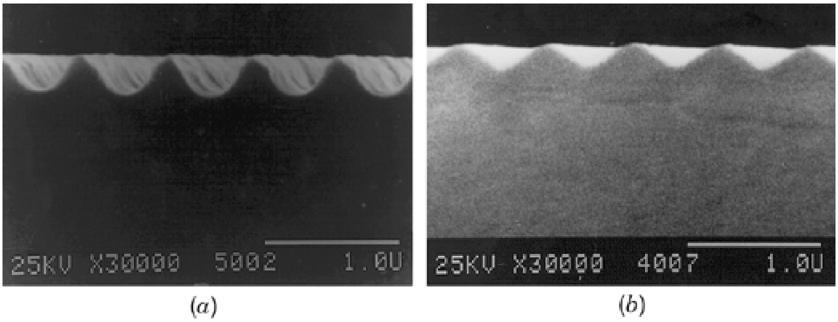


Figure 3.13 SEM photographs of cross section of gratings recorded in (100) *n*-InP samples by PEC etching: (*a*) grating lines parallel to direction $\langle 0\overline{1}1\rangle$; (*b*) grating lines parallel to direction $\langle 011\rangle$.

photochromic and light-induced photochromic effects. Organic polymers are now also available that exhibit photorefractive effects, with the advantage that their properties can be tailored to specific requirements. Table 3.1 lists the properties of some photorefractive crystals.

Photorefractives are photoconductive and electro-optic materials [8]: Under the action of light of adequate wavelength, charge carriers (electrons and/or holes) are excited from photoactive centers somewhere inside the material band gap into the conduction and/or valence band. These carriers move by diffusion and/or by the action of an externally applied electric field until they are retrapped somewhere else into accepting centers. If a spatially modulated pattern of light shines on the sample, the charge carriers are progressively accumulated in the darker areas, where the photoexcitation rate is lower. An electric charge modulation in the crystal volume results in a corresponding space-charge electric field (E_{sc}) modulation ruled by Poisson's law:

$$\nabla \cdot (\varepsilon \varepsilon_0 \mathbf{E}_{sc}) = \rho \tag{3.30}$$

Where ε_0 is the vacuum electric permittivity, ε is the dielectric constant, and ρ

TABLE 3.1 Properties of Selected Photorefractive Crystals

Crystal	Diffraction Efficiency	Speed	Spectral Range
$LiNbO_3$	Up to 100%	Slow	Red to green
$BaTiO_3$	Up to 60–80%	Moderate	Red to green
GaAs	<1%	Very fast	Near infrared
$Bi_{12}SiO_{20}$	Up to 15%	Fast	Green to blue
$Bi_{12}TiO_{20}$	Up to 15%	Fast	Red to green

is the electric charge volume density. Because of the electro-optic (also known as the "Pockels") effect, a refractive index modulation Δn arises in the crystal volume. This modulation is proportional to the space-charge field defined in Eq. 3.30:

$$\Delta n = -\tfrac{1}{2} n^3 r_{\mathrm{eff}} R E \tag{3.31}$$

where n is the average refractive index of the material and r_{eff} is the effective electro-optic coefficient that is a tensorial parameter. In this way, any information that encoded as a spatial modulation of the light projected onto the photorefractive crystal will be stored in it as a refractive index modulation in the sample volume, that is, a volume hologram will be formed. The process is schematically represented in Figure 3.14.

If the refractive index modulation in the photorefractive crystal arises by a purely diffusion controlled mechanism of the charge carriers, as is the case for $Bi_{12}SiO_{20}$ without an externally applied field, the index-of-refraction hologram is $\frac{1}{2}\pi$ phase shift relative to the recording pattern of fringes onto the crystal (Fig. 3.14). In this case the phase shift φ between the interfering beams behind the crystal will be zero or π. The recording of a self-stabilized hologram in this material, under these conditions, will require the use of the first harmonic as error signal in the feedback loop, while the second harmonic can be used to follow the evolution of the diffraction efficiency during the recording process [19]. Note that for the case of photoresists and PEC etching in n-InP the second harmonic is used as an error signal instead.

In the presence of external electrical fields, however, the index-of-refraction modulation in the crystal will increase considerably [8]. The phase shift

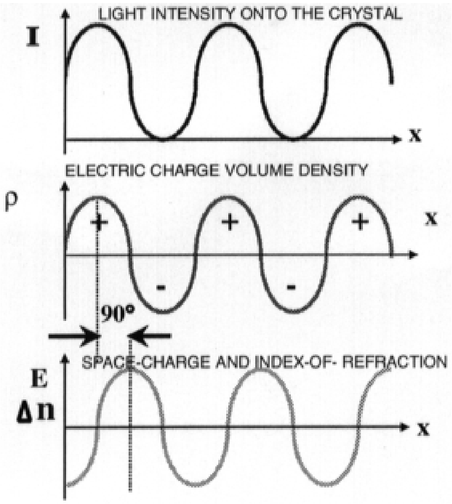

LIGHT INTENSITY ONTO THE CRYSTAL
I
x
ELECTRIC CHARGE VOLUME DENSITY
ρ
+
+
+
-
-
x
90°
SPACE-CHARGE AND INDEX-OF- REFRACTION
E
Δn
x

Figure 3.14 Recording of spatial light intensity modulation into refractive index modulation in volume of photorefractive crystal.

between the interference pattern and the refractive index grating will not be $\frac{1}{2}\pi$ any more but will depend on the value of the applied field and the characteristics of the crystal [20].

Some crystals that have strong photovoltaic effects, for example, lithium niobate crystals ($LiNbO_3$), exhibit a π-shifted holographic phase, in which case, as for the case of photoresists, the second-harmonic signal is used as error signal in the feedback loop [21, 22].

Independent of the feedback condition, measurement of the diffraction efficiency in the self-stabilized recording of a photorefractive crystal is a powerful technique to study the recording mechanism [23–25] and to measure fundamental parameters of the crystals [26–28].

To illustrate the practical applications of self-stabilized holographic recording in photorefractive crystals, we will describe its use for the measurement of mechanical vibrations in an experiment of holographic interferometry.

3.3.3.1 Holographic Interferometry The use of holography to measure the mechanical vibration of surfaces has long been known [29]. The possibility of using photorefractive materials in this field, however, is relatively new and has been in permanent development since the first paper by Huignard [30] on this subject. Self-stabilized real-time holography is completely new and was first shown to be feasible in 1986 [31].

The schematic setup for this experiment is shown in Figure 3.15. The input laser beam is divided into a reference and an object using a polarization beamsplitter (PBS) cube. The amplitude ratio between both beams is controlled using a half-wave retardation plate (HWP) at the PBS input. The polarizations of both beams exiting the PBS are made parallel [as required to produce interference fringes in the photorefractive crystal (PRC)] by the use of another HWP at the cube output. A low-power microscope objective lens is used to expand the object beam to illuminate the whole target surface. A device composed of a PBS, two HWPs and one quarter-wave retardation plate (QWP) is used to direct all the light onto the target surface and collect the back-scattered light through the PBS into the recording photorefractive crystal $Bi_{12}TiO_{20}$ (BTO) with minimum losses. Two good-quality photographic objective lenses are used to produce a reduced image of the target into the crystal and then to produce an enlarged picture of the latter image onto the charge-coupled device (CCD) camera for observation and image-processing purposes. The reference beam is also directed onto the BTO to interfere with the object beam and produce the required hologram, which will be recorded in the photorefractive crystal volume. The hologram is produced in real time in the BTO and at the same time the object wave is reconstructed from it by the reference beam. The latter reconstructed wave is actually the reference beam diffracted by the hologram in the BTO crystal that carries all the information needed.

Further details about the setup, the way the hologram should be read and the way the information about the surface vibration is obtained from the

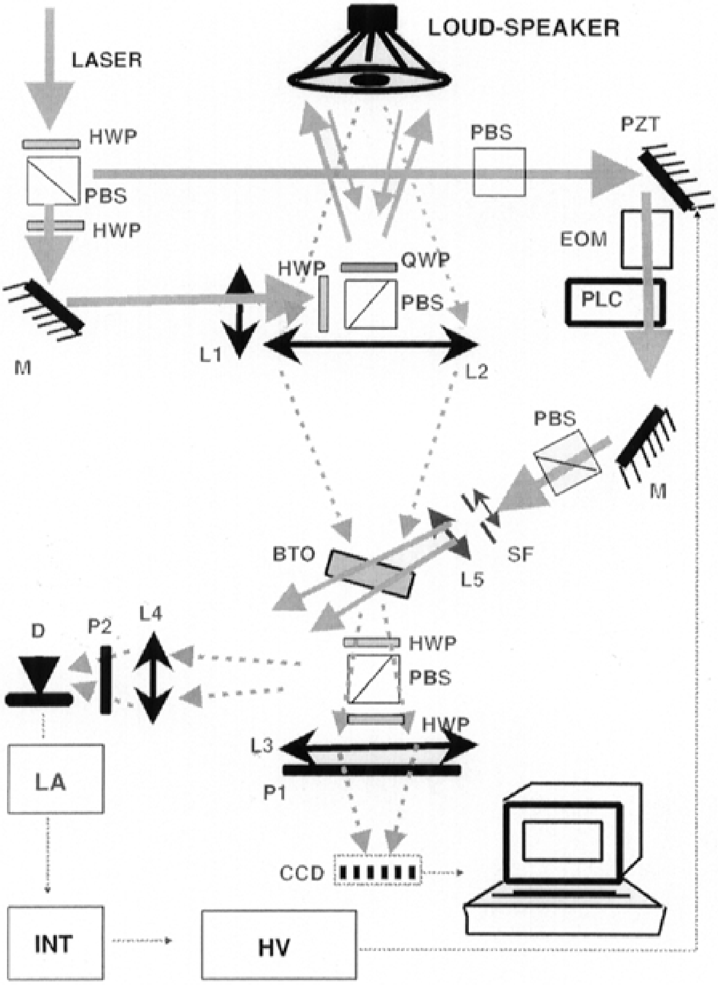


Figure 3.15 Schema to perform holographic interferometry using self-stabilized holographic recording in photorefractive BTO crystal.

hologram, as well as some information about the advantages of BTO-crystals for these applications, can be found in the literature [32, 33]. Here, we shall focus on the importance of self-stabilized holography for this particular application. In fact, the light that is scattered back from the target surface is usually rather weak. This means that the light shining on the BTO crystal to record the hologram is also quite weak. Then hologram buildup time (i.e., proportional to the average irradiance onto the crystal) will be proportionally large, and this means that the recording process will be very sensitive to external perturbations (e.g., air drafts, temperature variations). In certain

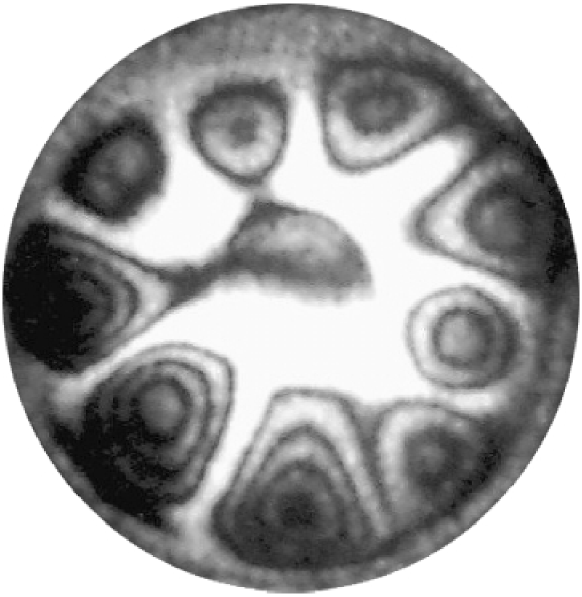


Figure 3.16 Photograph of real-time holographic pattern of vibrating membrane in self-stabilized holographic recording in BTO photorefractive crystal in pure diffusion regime. Bright regions correspond to vibration nodes.

circumstances the recording may be so perturbed that no hologram may be recorded at all in the crystal. To avoid this, the holographic setup is actively stabilized using the hologram recorded in the crystal, however weak it might be. For this purpose the beam along the transmitted reference wave behind the crystal in Figure 3.15 is used as an error signal to operate the stabilizing optoelectronic feedback loop.

Assuming that there are only two diffracted orders, the first and the zero, by energy conservation, we can substitute $\eta_1 = \eta$ and $\eta_0 = 1 - \eta$, in Eq. 3.3. The interference of the transmitted reference and the diffracted object beams may be therefore represented by

$$I_R = I_1\eta + I_2(1 - \eta) - 2\sqrt{I_1 I_2}\sqrt{\eta(1 - \eta)}\cos(\varphi) \tag{3.32}$$

where φ is the phase shift between the interfering beams and η is the diffraction efficiency of the hologram in the BTO. I_1 and I_2 is the irradiance of the light scattered back from the object onto the BTO and the irradiance of the reference beam shining on the BTO, respectively. The latter phase depends on the nature and characteristics of the recording material and the kind of hologram involved. It also depends on the phase shift between the pattern of fringes projected onto the crystal and the hologram produced by this pattern.

If the pattern of light is momentarily shifted due to an external perturbation, the value of φ will be consequently affected and the irradiance I_R will be accordingly modified. Photorefractive crystals operating in the pure diffusion regime, as is the present case, produce holograms that are $\frac{1}{2}\pi$ shifted from the recording pattern of light. Because of the index-of-refraction nature of the hologram recorded in these materials, an additional $\frac{1}{2}\pi$ phase is produced so that the equilibrium conditions require either $\varphi = 0$ or $\psi = \frac{1}{2}\pi$. To keep φ fixed at the required value, the small dither of frequency Ω is produced by the piezoelectric-supported mirror (PZT) in Figure 3.14, and the resultant first-harmonic term is selectively detected out from the irradiance I_R as explained in Section 3.2.4. A proportional electric error signal is thus generated:

$$V_\Omega \propto 4J_1(\varphi_d)\sqrt{I_1 I_2}\sqrt{\eta(1-\eta)}\sin(\varphi) \tag{3.33}$$

To operate the negative-feedback loop as reported in Section 3.2, so that the phase φ is actively fixed to zero or 180°, the setup is stabilized using the already recorded hologram as the reference. Using this technique, the recording process can be successfully carried out even for the case of extremely weak irradiances and very slow recording processes. Some pictures showing the vibrational modes of a thin metallic circular plate, excited by a loudspeaker placed behind the plate, are shown in Figure 3.16.

REFERENCES

1. D. B. Neumann and H. W. Rose, "Improvement of recorded holographic fringes by feedback control," *Appl. Opt.* **6**, 1097 (1967).
2. S. Johanson, L.-E. Nilsson, K. Biedermann, and K. Kleveby, "Holographic diffraction gratings with asymmetric groove profiles," in A. A. Friesem, E. Marom, and E. Wiener-Avnear (Eds.), *Applications of Holography and Optical Data Processing*, Proceedings of the International Conference on Holography and Optical Data Processing, Jerusalem, August 23–26, 1976, Pergamon, New York, 1976, p. 521.
3. D. R. MacQuigg, "Hologram fringe stabilization method," *Appl. Opt.*, **16**, 291–292 (1977).
4. http://www.xmission.com/~ralcon/whylock.html.
5. J. Frejlich, L. Cescato, and G. F. Mendes, "Analysis of an active stabilization system for a holographic setup," *Appl. Opt.* **27**, 1967–1976 (1988).
6. L. Cescato, G. F. Mendes, and J. Frejlich, "Stabilized holographic recording using the residual real-time effect in a positive photoresist," *Opt. Lett.* **12**, 982–983 (1987).
7. C. R. A. Lima and L. Cescato, "Mixing of the reflected waves to monitor and stabilize holographic exposures," *Opt. Eng.* **35**(10), 2804–2809 (1996).
8. P. Güenter and J. P. Huignard, *Photorefractive Materials and Their Applications I*, Topics in Applied Physics, Vol. 61, Springer, Berlin, 1988.
9. A. A. Freschi and J. Frejlich, "Adjustable phase control in stabilized interferometry," *Opt. Lett.* **20**(6), 635–637 (1995).

10. D. J. Elliott, *Integrated Circuit Fabrication Technology*, McGraw-Hill, New York, 1982, pp. 166–230.
11. J. W. Goodman, *Introduction to Fourier Optics*, 2nd ed., McGraw-Hill, Singapore, 1996, pp. 328–329.
12. J. Frejlich and L. Cescato, "Analysis of a phase modulating recording mechanism in negative photoresist," *J. Opt. Soc. Am.* **71**, 873–878 (1981).
13. L. Cescato and J. Frejlich, "Self-diffraction for intrinsic optical modulation evolution measurement in photoresist," *Appl. Opt.* **27**, 1984–1987 (1988).
14. B. M. Assunção, I. F. da Costa, C. R. A. de Lima, and L. Cescato, "Developed profile of holographically exposed gratings recorded in photoresists," *Appl. Opt.* **34**, 597–603 (1995).
15. L. Cescato, L. L. Soares, E. Luiz Rigon, M. A. R. Alves, and E. S. Braga, "Noise reduction in the recording of holographic masks in photoresist," in S. H. Lee and J. A. Cox (Eds.), *Micromachine Technology for Diffractive and Holographic Optics, Proceedings of SPIE*, Vol. 3879, 21–23 September 1999, Santa Clara, CA, pp. 214–222 (1999).
16. F. W. Ostermayer, Jr., P. A. Kohl, and R. H. Burton, *Appl. Phys. Lett.* **43** (7), 642–644 (1983).
17. D. Soltz, M. A. de Paoli, and L. Cescato, "Fringe stabilization and depth monitoring during the holographic photoelectro-chemical etching of n-InP (100) substrates," *J. Vac. Sci. Technol. B* **14**(3), 1784–1789 (1996).
18. J. Feinberg, "Photorefractive nonlinear optics," *Physics Today*, October 1988, pp. 46–52.
19. P. A. M. Santos, L. Cescato, and J. Frejlich, "Interference-term real-time measurement for self-stabilized two-wave mixing in photorefractive crystals," *Opt. Lett.* **13**, 1014–1016 (1988).
20. P. M. Garcia, L. Cescato, and J. Frejlich, "Phase-shift measurement in photorefractive holographic recording," *J. Appl. Phys.* **66**, 47–49 (1989).
21. A. A. Freschi and J. Frejlich, "Stabilized photorefractive modulation recording beyond 100% diffraction efficiency in $LiNbO_3$:Fe crystals," *J. Opt. Soc. Am. B* **11**, 1837–1841 (1994).
22. P. M. Garcia, K. Buse, D. Kip, and J. Frejlich, "Self-stabilized holographic recording in $LiNbO_3$:Fe crystals," *Opt. Commun.* **117**, 235–240 (1995).
23. J. Frejlich, A. A. Kamshilin, and P. M. Garcia, "Selective two-wave mixing in photorefractive crystals," *Opt. Lett.* **17**, 249–251 (1992).
24. S. Bian and J. Frejlich, "Photorefractive response time measurement in GaAs crystals using phase modulation in two-wave mixing," *Opt. Lett.* **19**, 1702–1704 (1994).
25. K. Buse, S. Kämper, J. Frejlich, R. Pankrath, and K. H. Ringhofer, "Tilting of holograms in photorefractive $Sr_{0.6}Ba_{0.39}Nb_2O_6$ crystals by self-diffraction," *Opt. Lett.* **20**, 2249–2251 (1995).
26. S. Bian and J. Frejlich, "Actively stabilized holographic recording for the measurement of photorefractive properties of a Ti-doped KNSBN crystal," *J. Opt. Soc. Am. B.* **12**, 2060–2065 (1995).
27. B. Sugg, K. V. Shcherbin, and J. Frejlich, "Determination of the time constant of fast photorefractive materials using the phase modulation technique," *Appl. Phys. Lett.* **66**, 3257–3259 (1995).

28. A. A. Freschi, P. M. Garcia, and J. Frejlich, “Charge-carrier diffusion length in photorefractive crystals computed from the initial hologram phase-shift,” *Appl. Phys. Lett.* **71**, 2427–2429 (1997).

29. R. J. Collier, C. B. Burckhardt, and L. H. Lin, *Optical Holography*, Academic, New York, 1971.

30. J. P. Huignard, “Time average holographic interferometry with photoconductive electro-optic $Bi_{12}SiO_{20}$ crystals,” *Appl. Opt.* **16**, 2796–2798 (1977).

31. A. A. Kamshilin, J. Frejlich, and L. Cescato, “Photorefractive crystals for the stabilization of a holographic setup,” *Appl. Opt.* **25**, 2375–2381 (1986).

32. E. A. Barbosa, J. Frejlich, V. V. Prokofiev, N. J. H. Gallo, and J. P. Andreeta, “Adaptive holographic interferometry for two-dimensional vibrational mode display,” *Opt. Eng.* **33**, 2659–2662 (1994).

33. J. Frejlich, E. de Carvalho, and A. A. Freschi, “Stabilized holographic setup for the real-time continuous measurement of surface vibrational mode patterns,” in *Proceedings of the Second International Conference on Vibration Measurements by Laser Techniques: Advances and Applications*, Ancona, Italy, 23–25 September, 1996, SPIE Vol. 2868, p. 205.

Three-Dimensional Holographic Imaging. Edited by Chung J. Kuo, Meng Hua Tsai
Copyright © 2002 John Wiley & Sons, Inc.
ISBNs: 0-471-35894-0 (Hardback); 0-471-22454-5 (Electronic)

CHAPTER 4

Optical Scanning Holography: Principles and Applications

TING-CHUNG POON

Optical Image Processing Laboratory
Bradley Department of Electrical and Computer Engineering
Virginia Polytechnic Institute and State University
Blacksburg, Virginia 24061

4.1 INTRODUCTION

Optical scanning holography (OSH) is a real-time holographic recording technique first suggested by Poon and Korpel [1] in which holographic information of an object can be recorded using two-dimensional (2D) optical heterodyne scanning. In other words, 3D information of an object can be extracted by 2D optical scanning of the object. Upon scanning, the scattered or reflected light is detected by a photodetector. The instantaneous electrical signal from the photodetector contains the holographic information of the scanned object. If the scanned electrical signal is stored in synchronization with the 2D scan signals of the scanning mechanism, what is stored as a 2D record is a hologram of the scanned 3D object. Since an electronic processing technique is used in the context of holographic recording, the technique is real time, bypassing the use of films for recording. Such a holographic recording technique nowadays is commonly known as electronic holography [2]. In Section 4.2, we provide a mathematical description of optical heterodyne scanning. In Section 4.3, we discuss the principles of OSH. Sections 4.4–4.6 present some important applications of OSH, such as 3D holographic microscopy, 3D image recognition, and 3D preprocessing and coding. Finally, in Section 4.4, we make some concluding remarks.

4.2 OPTICAL HETERODYNE SCANNING TECHNIQUE

A typical optical heterodyne scanning system is shown in Figure 4.1. Masks $m_1(x, y)$ and $m_2(x, y)$ are illuminated by uniform laser beams at frequencies ω_0

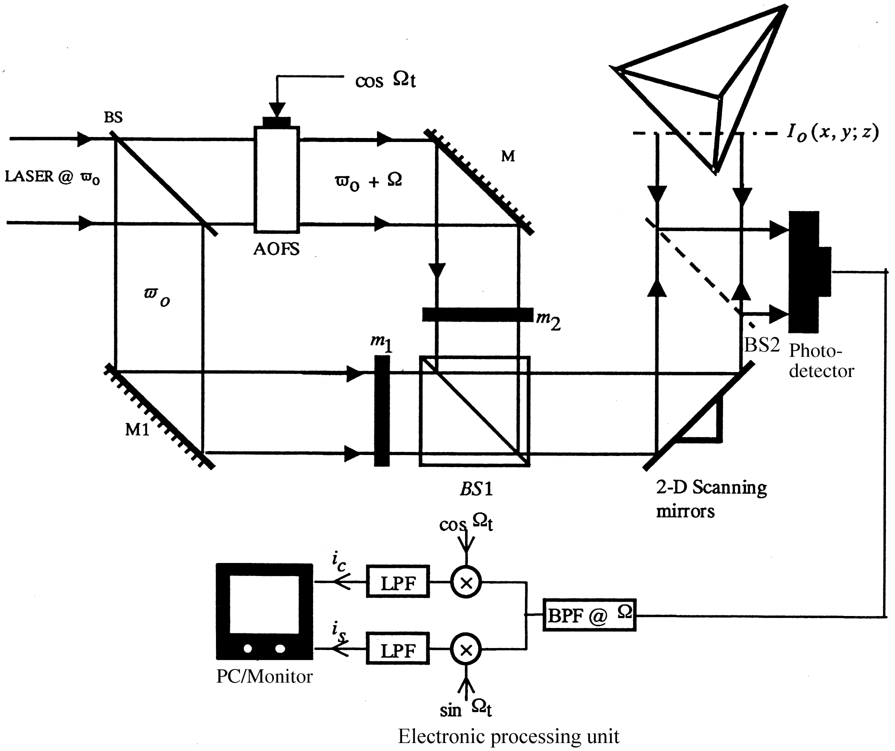


Figure 4.1 Optical heterodyne scanning image processor (AOFS, acousto-optic frequency shifter; BPF@Ω, bandpass filter tuned at frequency Ω, LPF, low-pass filter; ⊗, electronic multiplier).

and $\omega_0 + \Omega$, respectively. As shown in Figure 4.1, an acousto-optical frequency shifter (AOFS) has been inserted in the upper arm of the interferometer to provide a temporal frequency shift of Ω in the laser beam [3]. For simplicity, we write the two light fields after the two masks as $m_1(x, y)\exp(j\omega_0 t)$ and $m_2(x, y)\exp[j(\omega_0 + \Omega)t]$. Beamsplitter BS1 combines the two laser beams that diffract and give an intensity distribution $S(x, y; z)$ at a distance z away from the masks:

$$S(x, y; z) = |m_{1z}(x, y)\exp(j\omega_0 t) + m_{2z}(x, y)\exp[j(\omega_0 + \Omega)t]|^2, \qquad (4.1)$$

where $m_{1z}(x, y)$ and $m_{2z}(x, y)$ are Fresnel diffraction of $m_1(x, y)$ and $m_2(x, y)$ through a distance z, respectively. Here $m_i(x, y)$ and $m_{iz}(x, y)$ are related by a 2D convolution operation:

$$\begin{aligned} m_{iz}(x, y) &= \iint m_i(x', y')h(x - x', y - y')dx'\,dy' \\ &= m_i(x, y) \otimes h(x, y; z) \qquad i = 1, 2 \qquad (4.2) \end{aligned}$$

where $h(x, y; z)$ is the free-space impulse response and given by [4]

$$h(x, y; z) = \frac{jk_0}{2\pi z} \exp\left[-j \frac{k_0}{2z} (x^2 + y^2) \right] \tag{4.3}$$

with k_0 denoting the wave number of the light, and $\otimes$denotes 2D convolution operation [4]. The intensity distribution $S(x, y; z)$ is now used to 2D scan a 3D object. To simplify the discussion, let us assume that a planar intensity distribution of the 3D object, located z away from the masks, is $I_0(x, y; z)$. The photodetector collects the reflected light from I_0 through beamsplitter BS2, as shown in Figure 4.1. The scanned current, at the output of the photodetector, is thus given by

$$\begin{aligned} i(x, y; z) &= \iint_A I_0(x', y'; z) S(x - x', y - y'; z)\, dx'\, dy' \\ &= S^*(x, y; z) \odot I_0(x, y; z) \end{aligned} \tag{4.4}$$

Note that the integration is over the area A of the photodetector, the shifted coordinates of S represent the action of 2D scanning, $x = x(t)$ and $y = y(t)$, and finally $\odot$ denotes correlation operation [4]. The scanned current contains a baseband current and a heterodyne current at frequency Ω. When the scanned current is processed by the electronic processing unit as shown in Figure 4.1, two final processed currents, $i_c(x, y; z)$ and $i_s(x, y; z)$, apart from some inessential constants, are obtained [5]:

$$i_c(x, y; z) = \mathrm{Re}[m_{1z}(x, y) m_{2z}^*(x, y)] \odot I_0(x, y; z) \tag{4.5}$$

$$i_s(x, y; z) = \mathrm{Im}[m_{1z}(x, y) m_{2z}^*(x, y)] \odot I_0(x, y; z) \tag{4.6}$$

where $\mathrm{Re}[\cdot]$ and $\mathrm{Im}[\cdot]$ stand for the real and imaginary parts of the quantities being bracketed, respectively. Note that $i_c(x, y; z)$, and $i_s(x, y; z)$ represent the scanned and the processed information of $I_0(x, y; z)$ and can be displayed as a 2D processed image on a PC monitor. We want to point out that, in practice, the sine and cosine signals in the electronic processing unit may be derived from a signal generator used to drive the AOFS. By choosing masks $m_i(x, y)$ accordingly, we can achieve different processing operations on the object $I_0(x, y; z)$. In fact, heterodyning optical processors similar to the one shown in Figure 4.1 have been explored for bipolar incoherent image processing [1, 6].

A more generalized processing operation can be achieved when we combine Eq. 4.5 and Eq. 4.6 as follows:

$$i(x, y; z) = i_c(x, y; z) + j i_s(x, y; z) = [m_{1z}(x, y) m_{2z}^*(x, y)] \odot I_0(x, y; z) \tag{4.7}$$

The complex addition shown in Eq. 4.7 can be performed digitally or optically and, therefore, we can obtain complex operations on intensity distribution. To generalize Eq. 4.7 for 3D objects, we just need to integrate along the depth of the 3D object. We, therefore, have a 2D record $i(x, y)$, being processed by $m_{1z}(x, y)m_{2z}^*(x, y)$, of the 3D object:

$$i(x, y) = \int [m_{1z}(x, y)\mathrm{m}_{2z}^*(x, y)] \odot I_o(x, y; z)dz. \tag{4.8}$$

4.3 SCANNING HOLOGRAPHY

In this section, we show how the optical heterodyne scanning image processor shown in Figure 4.1 can be configured into a real-time optical holographic recording system. We let $m_1(x, y) = \delta(x, y)$ and $m_2(x, y) = 1$. Equation 4.1 becomes $S(x, y; z) \sim \sin[(k_0/2z)(x^2 + y^2) + \Omega t]$, that is, the scanning intensity pattern is of the form of a Fresnel zone pattern (FZP) with a time dependence. We shall call it a time-dependent FZP. Now, the two processed currents, from Eq. 4.5 and Eq. 4.6, become

$$i_c(x, y) \sim \int \sin \frac{k_0}{2z}(x^2 + y^2) \odot I_o(x, y; z)\, dz = H_{\sin}(x, y) \tag{4.9}$$

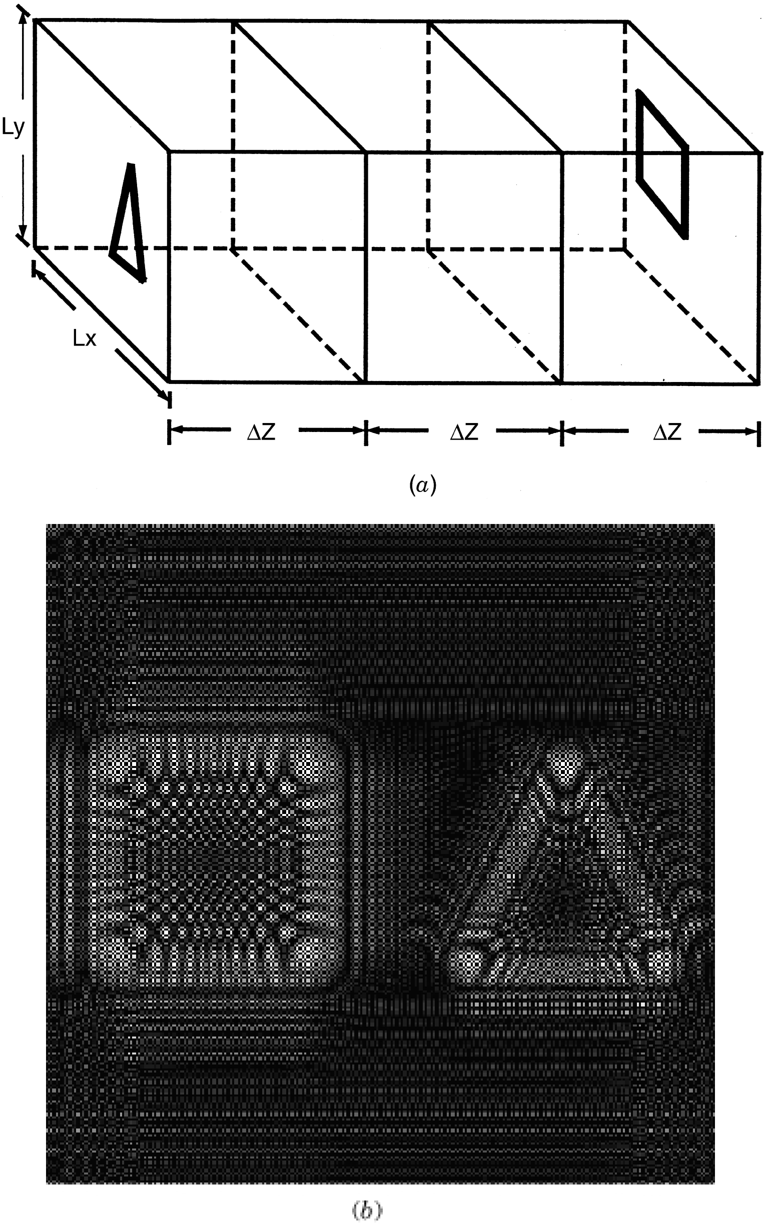
$$\mathrm{i}_s(x, y) \sim \int \cos \frac{k_0}{2z}(x^2 + y^2) \odot I_o(x, y; z)\, dz = H_{\cos}(x, y) \tag{4.10}$$

where $H_{\sin}(x, y)$ and $H_{\cos}(x, y)$ are real holograms and are the so-called sine–Fresnel zone pattern (sine–FZP) hologram and cosine–FZP hologram of object $I_o(x, y; z)$, respectively [5]. To produce a single-sided holographic record free from the zero-order term and the twin-image term [7], we add the two real holograms to form a complex hologram of I_o, $H_{I_o}(x, y)$, according to

$$\begin{aligned} H_{I_o}(x, y) &= H_{\cos}(x, y) - jH_{\sin}(x, y) \\ &= \int \exp\left[-j\frac{\pi}{\lambda z}(x^2 + y^2)\right] \odot I_o(x, y; z)\, dz. \end{aligned} \tag{4.11}$$

This 2D optical scanning technique used to generate holographic information of an object is called optical scanning holography [8–11]. As an example demonstrating the principles discussed, Figure 4.2*a* shows a 3D object consisting of two planar intensity distributions side by side and at different distances from the two masks. One distribution is a square pattern and the other a triangle pattern. The depth difference of the two slides is ~20.4 cm. This 3D object is modeled as $I_o(x, y; z) = I_1(x, y; z)\delta(z - z_1) + I_2(x, y; z)\delta(z - z_2)$. Its cosine hologram is given by, using Eq. 4.10,

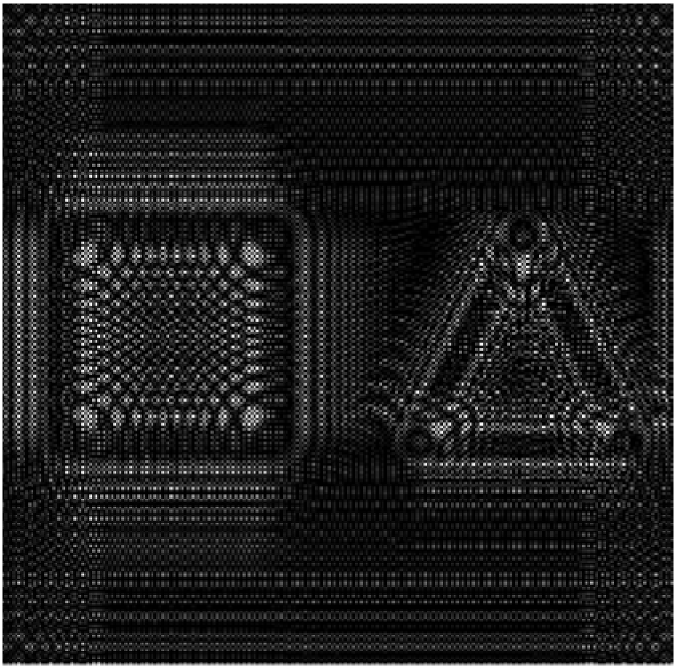
$$\cos\left[\frac{\pi}{\lambda z_1}(x^2 + y^2)\right] \odot I_1(x, y; z_1) + \cos\left[\frac{\pi}{\lambda z_2}(x^2 + y^2)\right] \odot I_2(x, y; z_2) \tag{4.12}$$



(*a*)

(*b*)

Figure 4.2 (*a*) Three-dimensional object, with $L_x = L_y = 1$ cm and $3\,\Delta z \sim 20.4$ cm; (*b*) corresponding cosine FZP hologram; (*c*) corresponding sine FZP hologram.



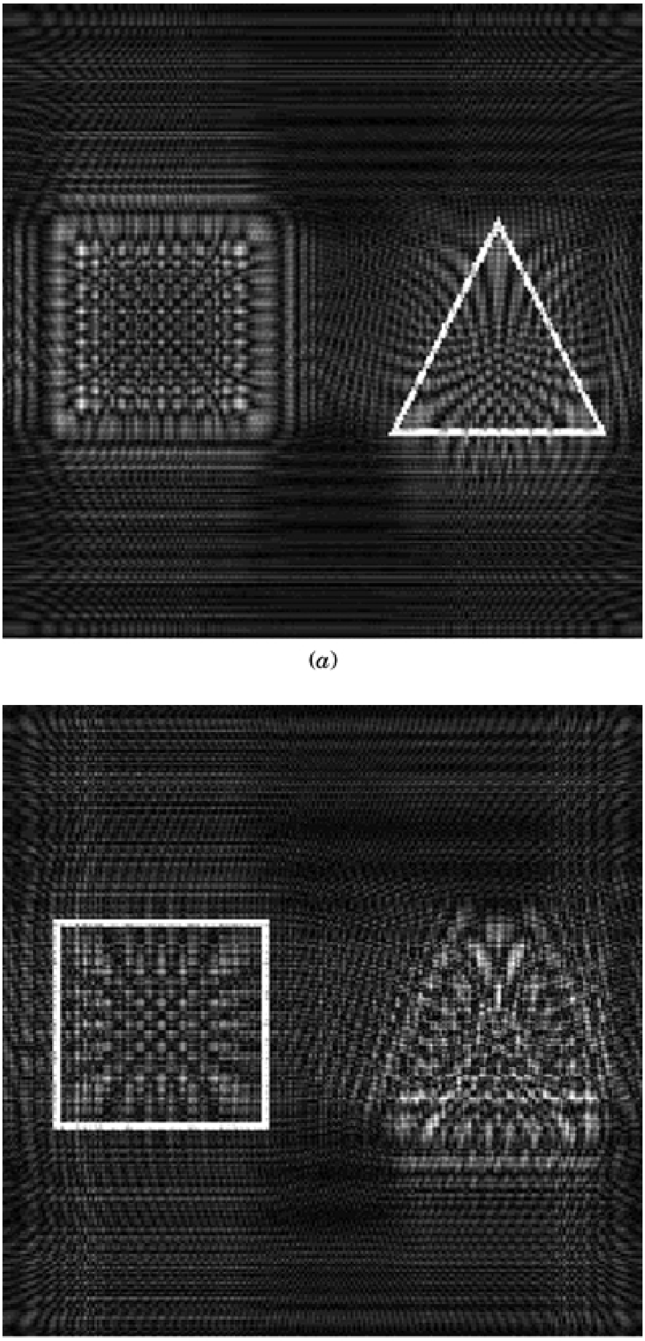
(*c*)

Figure 4.2 (*Continued*).

where I_1 and I_2 represent the planar intensity distributions of the "triangle" and the "square" at their perspective depth locations, z_1 and z_2. For the sine hologram, the cosine function is replaced by a sine function in Eq. 4.12. Figure 4.2*b* and *c* show the cosine and sine holograms for $z_1 \sim 20.4\,\text{cm}$, $z_2 = 2z_1 \sim 40.8\,\text{cm}$, and $\lambda \sim 0.6\,\mu\text{m}$ in the simulations. The complex hologram, using Eq. 4.11, is given by

$$\exp\left[-\frac{j\pi}{\lambda z_1}(x^2+y^2)\right] \odot I_1(x, y; z_1) + \exp\left[-\frac{j\pi}{\lambda z_2}(x^2+y^2)\right] \odot I_2(x, y; z_2) \tag{4.13}$$

For the reconstruction of holograms, this can be done by convolving the holograms given by either Eq. 4.12 or 4.13 with the free-space impulse response $h(x, y; z) = (j/\lambda z)\exp[-j\pi(x^2+y^2)/\lambda z]$. This corresponds to an optical reconstruction at a distance z from the hologram that would be obtained by illuminating the hologram with a plane wave. Figure 4.3*a* shows the reconstruction of the cosine hologram given by Eq. 4.12. The reconstruction is focused at $z = z_1$ as $h(x, y; z_1)$ was used for the convolution calculation. This



(*a*)

(*b*)

Figure 4.3 Reconstruction of cosine hologram at (*a*) $z = z_1$ and (*b*) $z = z_2$.

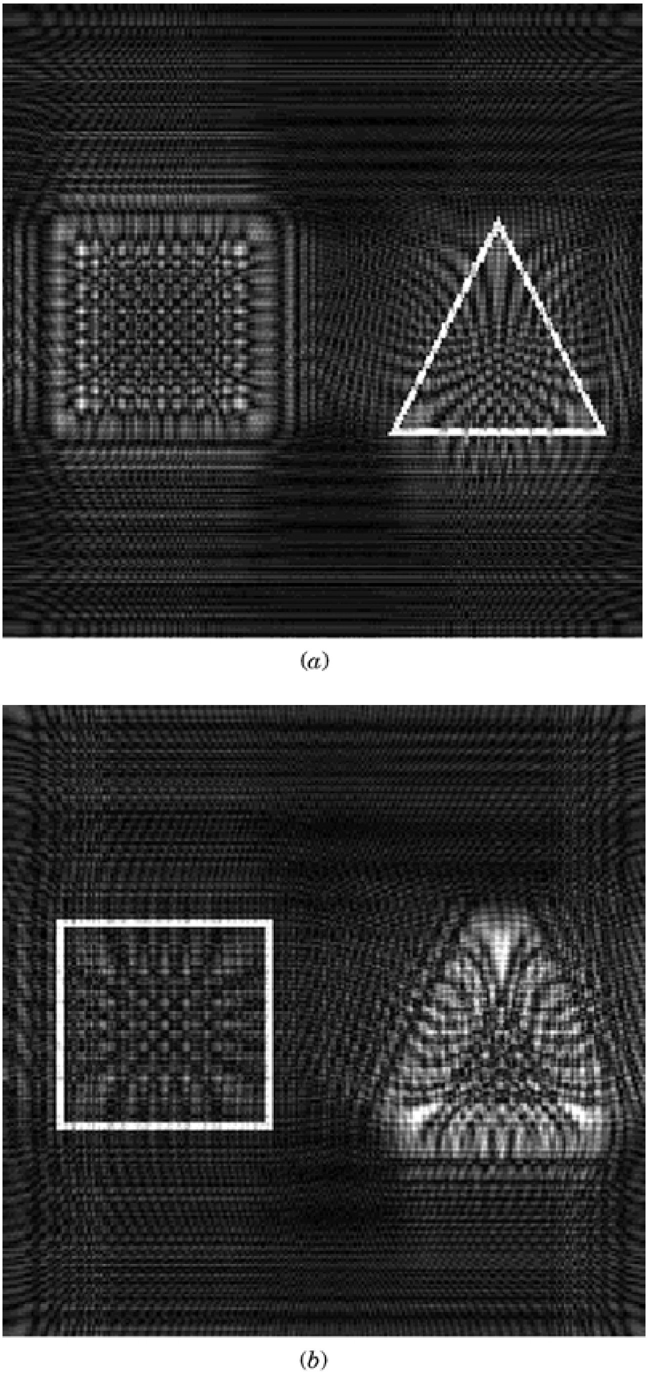
reconstruction is focused on the triangle object but is spoiled by twin-image noise. When $h(x, y; z_2)$ is used to reconstruct, the reconstruction as shown in Figure 4.3*b* is now focused at $z = z_2$, the location of the square. Again, twin-image noise is evident. Similar results are obtained when the sine hologram is used for reconstruction, as shown in Figures 4.4*a* and *b*, where the images are focused at the triangle and the square, respectively. However, when the complex hologram, given by Eq. 4.13, is reconstructed, we obtain

$$I_1(x, y; z_1) + \left\{ \exp\left[-\frac{j\pi}{\lambda z_2}(x^2 + y^2) \right] \odot I_2(x, y; z_2) \right\} \otimes h(x, y; z_1) \quad (4.14)$$

The first term corresponds to the triangle image properly focused and without twin-image noise. The second term represents the defocused image of the square. Figures 4.5*a* and *b* show the reconstruction of the complex hologram convolved with $h(x, y; z_1)$ and $h(x, y; z_2)$, respectively. In summary, the real holograms reconstruct focused images but spoiled by twin-image noise, while the complex hologram gives reconstructions without twin-image noise, as shown in Figure 4.5.

4.4 THREE-DIMENSIONAL HOLOGRAPHIC FLUORESCENCE MICROSCOPY

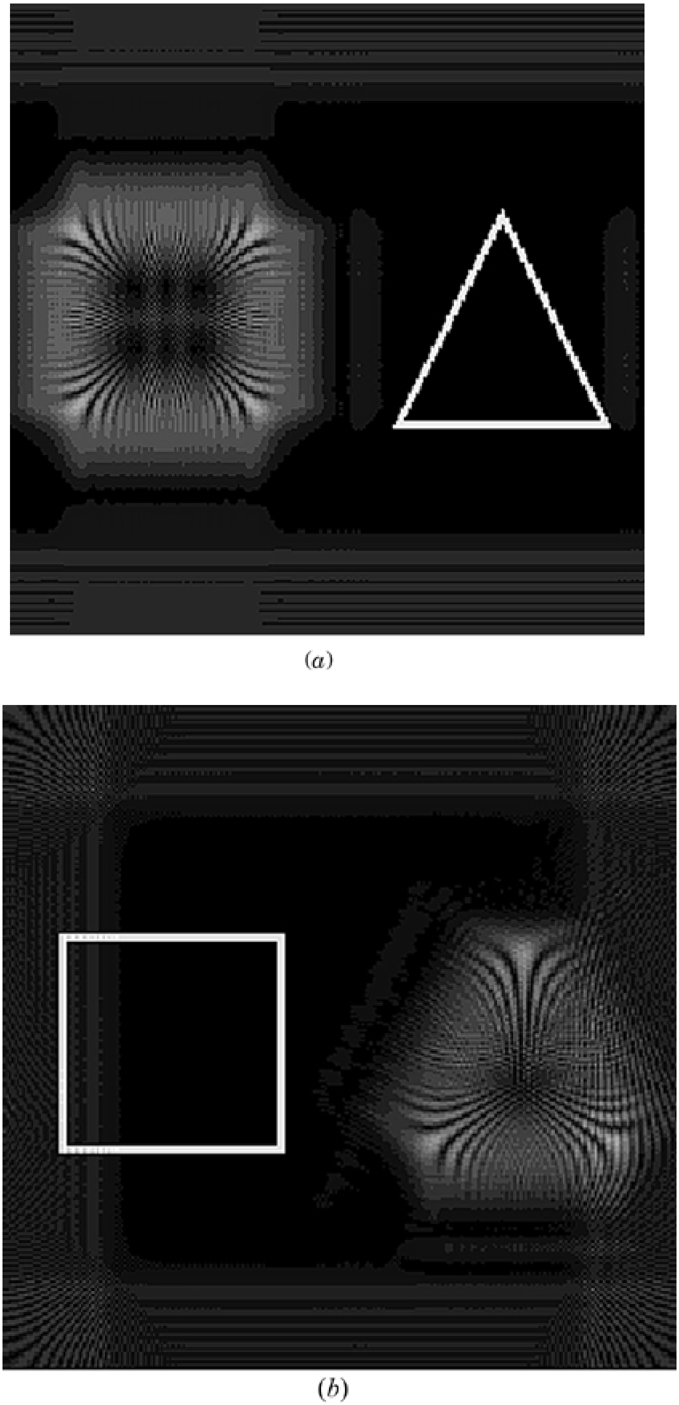
Three-dimensional imaging is a formidable task in optical microscopy, as it is well-known that the greater the lateral magnification, the smaller the depth of field. The past decade has witnessed an impressive emergence of 3D imaging techniques in microscopy. Specifically, a radically new microscope design, the scanning confocal microscope (SCM), has emerged [12, 13]. In the SCM, a doubly focused objective lens system and a pinhole aperture in front of a photomultiplier are used to image only a single point within the 3D specimen. The situation is shown in Figure 4.6. Light emitted by points outside the plane of focus (dashed line) is rejected by the pinhole aperture in front of the photodetector. Three-dimensional information is gathered by scanning the specimen in three dimensions while collecting the light transmitted through the specimen with the photodetector. Some of the main problems of the SCM are that the scanning instrumental tolerances required to achieve high-resolution imaging are very difficult to obtain, that 3D scanning is time consuming, and the working distance along the depth of the specimen is severely limited. In addition, for fluorescence microscopy applications, photobleaching is a concern and repeated 2D sectioning only exacerbates this problem by increasing exposure [14]. These drawbacks gave us the impetus to investigate holographic techniques [15]. An attractive quality of holography is the inherent ability to store 3D information in a 2D array without a depth scan. In this section, we discuss a new type of microscope that combines the 3D imaging capability of



(a)

(b)

Figure 4.4 Reconstruction of sine hologram at (*a*) $z = z_1$ and (*b*) $z = z_2$.


(*a*)

(*b*)

Figure 4.5 Reconstruction of complex hologram at (*a*) $z = z_1$ and (*b*) $z = z_2$.

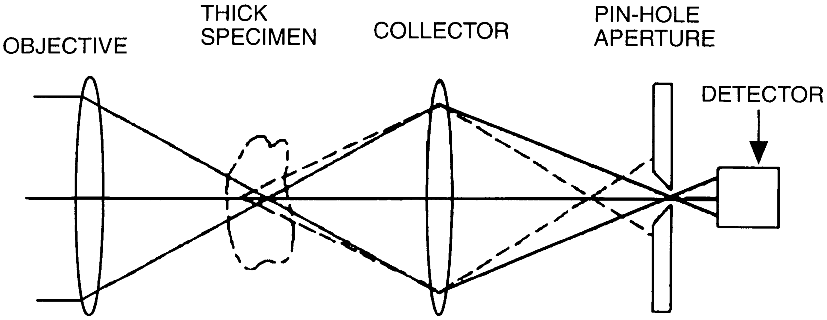


Figure 4.6 Principle of optical confocal scanning microscopy. (The thick specimen is 3D scanned.)

optical scanning holography with the advantages of fluorescence techniques, which leads to the first 3D fluorescence holographic microscope ever developed.

Figure 4.7 presents the holographic fluorescence microscope [15]. As discussed in Section 4.3, OSH is based on scanning the object with a temporally modulated FZP. To apply OSH to fluorescent specimens, the FZP is generated at a wavelength near the peak absorption of the specimen. The simplified block

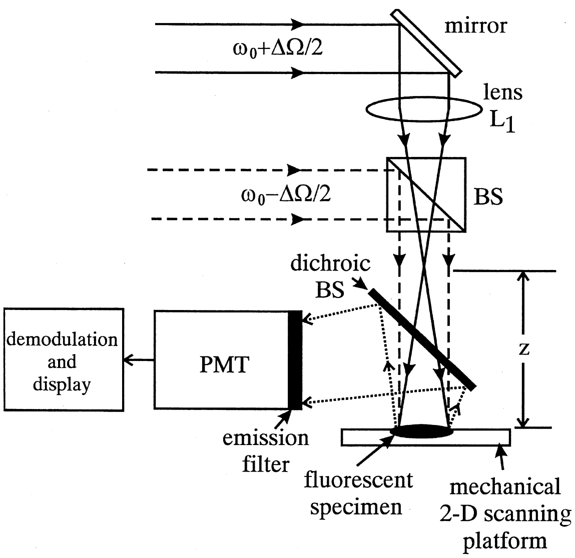


Figure 4.7 Experimental setup used to record hologram of fluorescent specimen by OSH (BS, beamsplitter; PMT, photomultiplier tube).

diagram of the experimental setup, sketched in Figure 4.7, shows two plane waves separated in frequency by $\Delta\Omega$. In the experiment, these plane waves originate from the 514-nm line of a multiline argon-ion laser ($\omega_0 = 3.667 \times 10^{15}$ rad/s). The frequency shift in each beam is achieved by an AOFS. The AOFS is used in a configuration that splits the laser light into two beams separated in frequency by $\Delta\Omega/2\pi = 10.7\,\text{MHz}$. The beams are then collimated and made parallel to each other, as shown in the figure. A lens is placed in one of the beams to form the spherical wave, which is then combined collinearly with the other beam at the dichroic BS. The resulting interference pattern at the object, which is a distance z beyond the focus of the spherical wave, is the time-dependent FZP laser field. The dichroic beamsplitter (transmission at 514 nm/reflection at 595 nm) allows light at 514 nm to transmit and excite the fluorescent object, which fluoresces strongly around the center frequency of 560 nm. This light is reflected by the dichroic beamsplitter and passes through a 595 nm narrow-bandpass emission filter (used to reject the background laser light at 514 nm) and into the photomultiplier tube (PMT). The specimen is scanned through the FZP in a raster pattern using a computer-controlled mechanical x–y scanning platform. The PMT current, which contains 3D information of the object being scanned, is electronically filtered and amplified at 10.7 MHz and demodulated according to the electronic processing unit shown in Figure 4.1, except for the fact that only one channel was used and hence twin-image noise will appear on the reconstruction plane. The demodulated signal is then digitized by a standard PC-based analog-to-digital (A/D) board. This digitized intensity pattern is stored in a 2D array, in accordance with the scanning to produce the hologram.

The fluorescent sample used in the experiment was a solution containing a high concentration of fluorescent latex beads. The beads were 15 μm in diameter with peak excitation at 530 nm and peak emission at 560 nm. These beads made good experimental specimens because they have a quantum efficiency near 100%, have broad excitation and emission bands, and have fluorescent lifetime shorter than 10 ns. To demonstrate the depth-discriminating capability of the system, the fluorescent object consisted of two wires placed next to each other parallel to the optical axis but with their ends at slightly different distances from the focus of lens L_1. A drop of the fluorescent solution was placed on the end of each of these wires. A hologram of this fluorescent sample was recorded and is shown in Figure 4.8. The two drops, easily distinguishable in the hologram, are separated in depth by approximately 2 mm, with the drop on the left at $z_0 \approx 35$ mm and the drop on the right at $z_1 \approx 37$ mm. The area scanned to produce the hologram is about 2×2 mm. The hologram is a 256×256 array with intensity level represented by 256 gray levels.

Once the hologram (Fig. 4.8) has been recorded and stored, the 3D image can be reconstructed either optically or digitally, as discussed in the last section. We emphasize the difference between the holographic technique and standard 3D imaging techniques. In optical sectioning methods, the image is

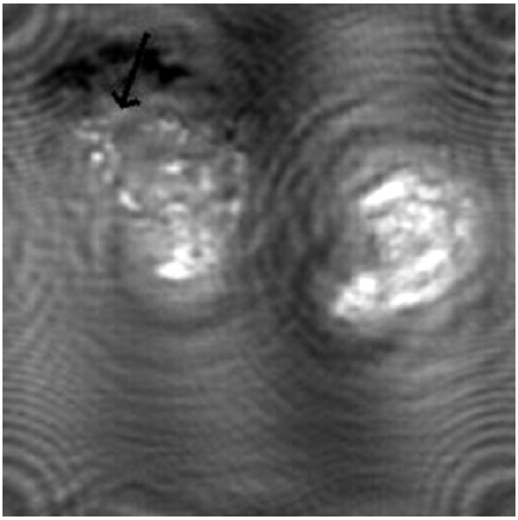
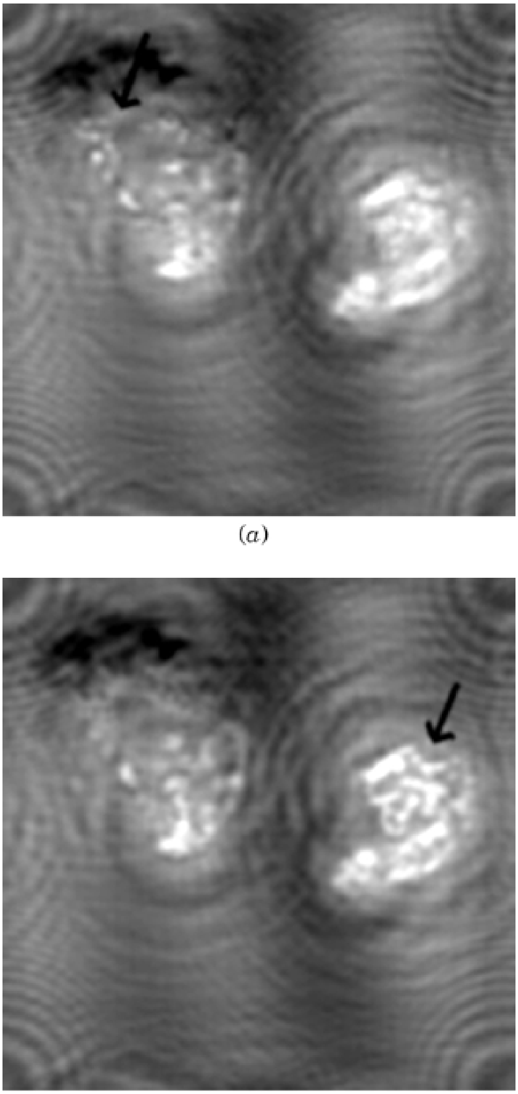


Figure 4.8 Hologram of fluorescent specimen recorded using OSH. Object consists of two drops of solution containing high concentration of fluorescent latex beads separated in depth by about 2 mm. Image is 256-level gray-scale image consisting of 256×256 pixels. Area scanned is about 2×2 mm. (After Ref. 15.)

brought into focus at a chosen depth, say z_0, and the 2D image for that plane is recorded and stored. This process is repeated for as many planes in z as desired, and an image must be recorded and stored for each plane. With OSH, a hologram of the object is recorded and stored with a single 2D scan. Since all the depth information is stored in the hologram, any desired image plane can be brought into focus during image reconstruction. Numerical image reconstruction has been performed on the hologram for two different depths. Figure 4.9*a* is a reconstructed image at $z_0 = 35$ mm, and Figure 4.9*b* is an image reconstruction at $z_1 = 37$ mm. Since the individual attributes in each fluorescent drop are not obvious in Figure 4.9, arrows have been overlaid on the figures to point out certain areas of interest. In Figure 4.9*a*, the fluorescent drop on the left is in better focus than that on the right. The arrow in Figure 4.9*a* points to particular beads that are better imaged when the hologram is reconstructed at depth z_0 than in Figure 4.9*b* at depth z_1. Similarly, the arrow in Figure 4.9*b* points out a string of four beads that can be individually distinguished when the hologram is reconstructed at depth z_1, but are blurred in the image reconstruction plane z_0 in Figure 4.9*a*. No attempt was made to eliminate the twin image in these reconstructions, which explains the residual "fringing" observable in these images. The resolution of the holographic system is limited by the system's numerical aperture (NA), which in turn depends on the focal length of lens L_1, which is 150 mm, and the diameter D of the plane wave focused by lens L_1 ($D = 10$ mm). The diffraction-limited resolution limits



(*a*)

(*b*)

Figure 4.9 (*a*) Reconstruction of hologram shown in Figure 4.8 at depth of $z_0 = 35\,\text{mm}$. Arrow shows individual fluorescent beads which are in focus at this depth. (*b*) Reconstruction of hologram shown in Figure 4.8 at depth $z_0 = 37\,\text{mm}$. Arrow shows four individual fluorescent beads which are in focus at this depth. (After Ref. 15.)

for our system can be calculated from the following equations:

$$\text{Lateral resolution} = \Delta x \sim \frac{\lambda}{2\text{NA}}$$

$$\text{Longitudinal resolution} = \Delta z \sim \frac{\lambda}{(\text{NA})^2}$$

where λ is the wavelength of the laser. The NA of our system is ~ 0.033, which corresponds to diffraction-limited resolution limits of $\Delta x \sim 7.7\,\mu\text{m}$ and $\Delta z \sim 200\,\mu\text{m}$. The 15-$\mu$m bead size is very close to the limit that we can expect to resolve laterally with the current setup and more than an order of magnitude beyond what we can resolve in the depth dimension.

4.5 THREE-DIMENSIONAL IMAGE RECOGNITION

Three-dimensional optical image recognition finds applications in the areas of 3D microscopy, medical imaging and recognition, robotic vision, 3D data acquisition and processing, and optical remote sensing. One of the key operations to achieve 3D optical image recognition is 3D optical correlation, that is, an optical technique capable of performing 3D correlation of a 3D reference object and a 3D target object to be recognized. This is a formidable task.

Research in 2D pattern recognition has its root in the 1960s [16, 17], and it has been rejuvenated in past decades due to the advancement of the development of spatial light modulators [18, 19]. Optical image recognition of 3D objects, on the other hand, is rarely tackled due mainly to the lack of optical systems capable of performing 3D correlation. Some work involving trans-dimensional mappings has been done in optics, but it exclusively dealt with 1D-to-2D or 2D-to-1D transformation [20–22]. A relevant 3D-to-2D mapping has been proposed that involves sampling along the 3D object's depth and hence represents the object by 2D sectional depth images [23]. Recently, a 3D joint-transform correlator has been demonstrated in that 2D images can be recognized even when the images are translated along the z direction. The technique involves capturing the 3D information by transverse displacement (i.e., x and y scanning) of a charge-coupled device (CCD) camera [24]. However, these two proposed techniques suffer a major drawback of dimensional reduction in that the number of these many 2D sampling images should be high for good resolution performance. To alleviate the need for many 2D sampling images, range images have been used, but coding schemes must be applied to the input scene and the reference image that makes the proposed system lack real-time capability [25]. Mostly recently, a planar encoding theory of 3D images has been proposed that does not require the many 2D images [26].

In this section, we investigate the use of OSH for 3D image recognition; that is, the goal is to recognize a 3D object, an object of three spatial coordinates. In essence, our proposed optical system performs essentially the correlation of two pieces of holographic information pertaining to the two 3D objects, a 3D reference object and a 3D target object, to be matched, and hence a 3D recognition is possible. In the technique, there is no need to record 2D images for 3D object representation as the holographic technique is employed. We let the planar distribution of a 3D reference object to be $R(x, y; z)$, that is, $I_o(x, y; z) = R(x, y; z)$, and the reference complex hologram, according to Eq. 4.11, is

$$H_R(x, y) = \int \exp\left[-j \frac{\pi}{\lambda z} (x^2 + y^2) \right] \odot R(x, y; z)\, dz \tag{4.15}$$

After the complex hologram of the reference object is generated, the object is removed. We shall call the recorded hologram a reference hologram as the object being recorded is the reference object. The reference hologram will be used in the system to recognize a 3D target. To be precise, in the recognition stage, one of the masks carries the holographic information of the reference object; that is, we let $m_1(x, y) = \ H_R(x, y)$ and keep $m_2(x, y) = 1$ as in the holographic recording stage. A 3D target object with its planar distance given as $O(x, y; z)$ is now placed z away from the masks of the scanning system. According to Eq. 4.8, with $I_o(x, y; z) = O(x, y; z)$ the output is

$$\begin{aligned} i(x, y) &= \int [m_{1z}(x, y) m_{2z}^*(x, y)] \odot O(x, y; z)\, dz \\ &= \int [H_R(x, y) \otimes h(x, y; z)] \odot O(x, y; z)\, dz \\ &\sim H_R(x, y) \odot \left[\int \exp\left(-j \frac{k_0}{2z} (x^2 + y^2) \right) \odot O(x, y; z) \right] dz \\ &= H_R(x, y) \odot H_O(x, y) \\ &= c(x, y) \end{aligned} \tag{4.16}$$

where $H_O(x, y)$ is the complex hologram of $O(x, y; z)$. Equation 4.16 demonstrates that the system can perform correlation of two holographic information and hence 3D recognition is possible. If the reference object and the target object are the same, that is, $O = R$, we will have a strong correlation peak, as shown in Figure 4.10, where Re$[c(x, y)]$ is plotted. Note that we have used O and R, as shown in Figure 4.2*a*, for simulations. Figure 4.11*a* shows a different 3D target object O, and the resulting correlation is shown in Figure 4.11*b*. We note the lack of a strong correlation peak due to the mismatch of the two 3D objects. The selection of $m_1(x, y) = H_R(x, y)$ to make the system become a 3D recognition system can be understood physically. Referring to Figure 4.12,

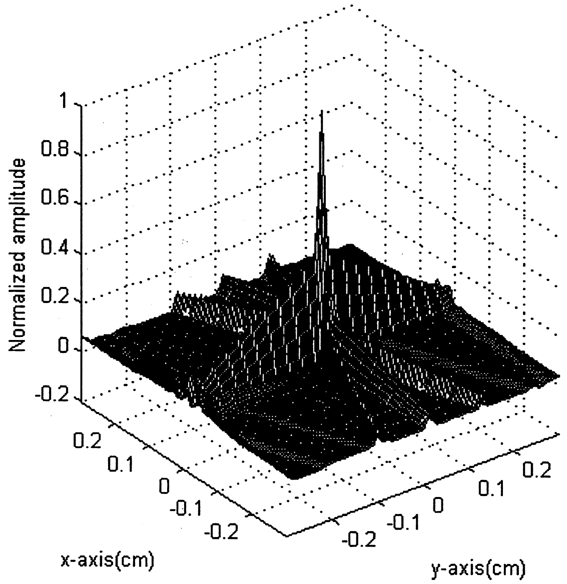


Figure 4.10 Normalized correlation output (0.4 × 0.4 cm) when target object is matched with reference object. (After Ref. 5.)

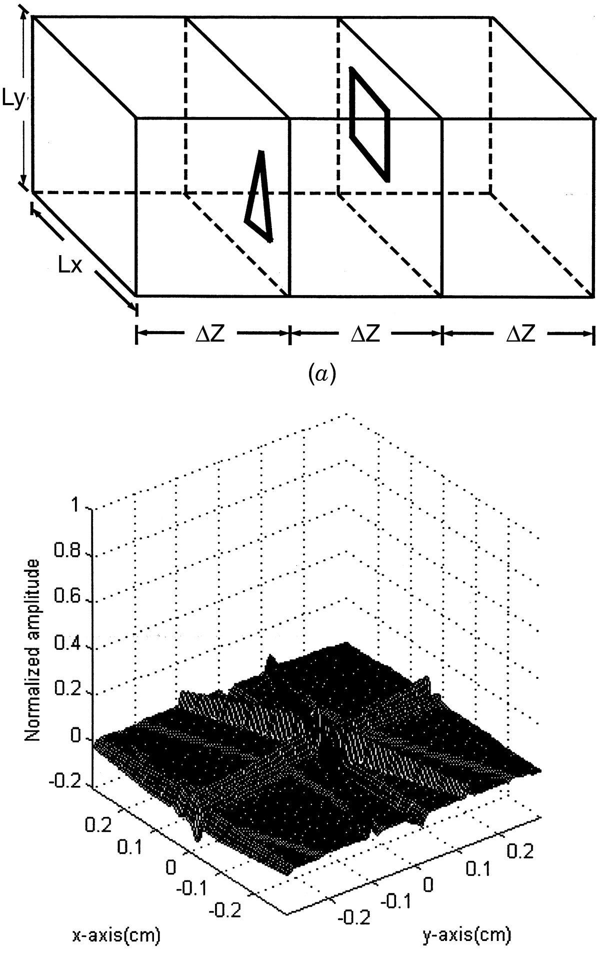
$m_1(x, y)$ generates a holographic 3D image (dotted lines) over the target object (solid lines). As the holographic image is 2D scanned over the target, a strong correlation peak would result if the target is matched to the holographic image.

4.6 PREPROCESSING OF HOLOGRAPHIC INFORMATION

In this section, we discuss that proper manipulation of the interfering waveforms used to scan the objects can result in holograms having unique properties. In fact, when the object is scanned by a structured beam, this is considered a type of preprocessing. Some important applications of preprocessing of holographic information can be found in the works of Molesini et al. [27], Vikram and Billet [28], and Ozkul et al. [29]. In the context of preprocessing in OSH, it is instructive to develop a transfer function approach to describe the processing. With reference to Eq. 4.8 and considering a planar input distribution $I_o(x, y; z)$ located z away from the masks, we have again the output $i(x, y; z)$ given by

$$i(x, y; z) = [m_{1z}(x, y)m_{2z}^*(x, y)] \odot I_o(x, y; z) \tag{4.17}$$

We define the 2D Fourier transform [4] as $\mathscr{F}\{u(x, y)\}_{k_x, k_y} = \iint u(x, y)\exp(jk_x x + jk_y y)\,dx\,dy = U(k_x, k_y)$ with k_x and k_y denoting the spatial



(*a*)

(*b*)

Figure 4.11 (*a*) Three-dimensional target object. (*b*) Correlation output when target object in (*a*) is scanned; reference object shown in Figure 4.2(*a*) (after Ref. 5).

frequencies and with the uppercase function U indicating the transform of the lowercase function u. We have, by taking the Fourier transform of Eq. 4.17,

$$\mathscr{F}[i(x, y; z)] = \mathscr{F}^*[m_{1z}(x, y)m_{2z}^*(x, y)]\mathscr{F}[I_o(x, y; z)] \tag{4.18}$$

We now define the optical transfer function (OTF) of the system as

$$\begin{aligned} \mathrm{OTF}(k_x, k_y; z) &= \mathscr{F}[i(x, y; z)]/\mathscr{F}[I_o(x, y; z)] \\ &= \mathscr{F}^*[m_{1z}(x, y)m_{2z}^*(x, y)] \end{aligned} \tag{4.19}$$

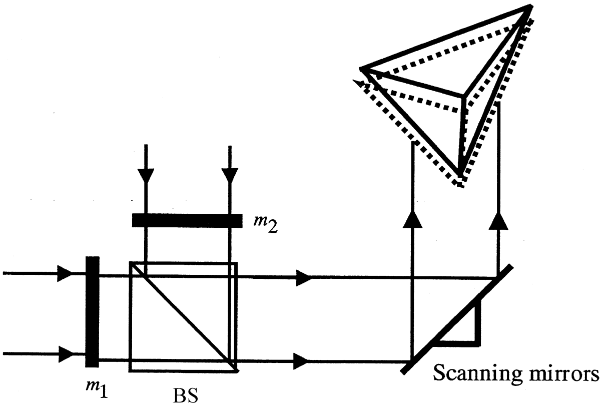


Figure 4.12 Physical interpretation of 3D recognition.

when we substitute Eq. 4.18 into Eq. 4.19. The OTF can be expressed directly in terms of masks $m_i(x, y)$ by substituting Eq. 4.2 into Eq. 4.19. We then have

$$\mathrm{OTF}(k_x, k_y; z) = \exp\left[-j\frac{z}{2k_0}(k_x^2 + k_y^2)\right] \times \iint m_1^*(x', y')$$
$$m_2\left(x' + \frac{z}{k_0}k_x, y' + \frac{z}{k_0}k_y\right)\exp[-j(x'k_x + y'k_y)]\,dx'\,dy' \qquad (4.20)$$

In the context of OSH, we again let $m_1(x, y) = \delta(x, y)$ and $m_2(x, y) = 1$. The OTF of optical scanning holography, $\mathrm{OTF}_{\mathrm{osh}}$, is then given by

$$\mathrm{OTF}(k_x, k_y; z) = \exp\left[-j\frac{z}{2k_0}(k_x^2 + k_y^2)\right] = \mathrm{OTF}_{\mathrm{osh}}(k_x, k_y; z) \qquad (4.21)$$

According to Eq. 4.19 and 4.21, the output spectrum in scanning holography is

$$\mathscr{F}[i(x, y; z)] = \exp\left[-j\frac{z}{2k_0}(k_x^2 + k_y^2)\right]\mathscr{F}[I_o(x, y; z)] \qquad (4.22)$$

It is important to point out that, for scanning holographic recording, the object's spectrum, $\mathscr{F}\{I_o(x, y; z)\}$, is filtered by a complex Fresnel zone pattern-type function $\exp[-j(z/2k_0)(k_x^2 + k_y^2)]$. If we now let $m_1(x, y) = \delta(x, y)$ and leave $m_2(x, y)$ as is, the OTF becomes, from Eq. 4.20,

$$\mathrm{OTF}(k_x, k_y; z) = \exp\left[-j\frac{z}{2k_0}(k_x^2 + k_y^2)\right] m_2\left(\frac{z}{k_0}k_x, \frac{z}{k_0}k_y\right)$$
$$= \mathrm{OTF}_{\mathrm{osh}}(k_x, k_y; z)m_2\left(\frac{z}{k_0}k_x, \frac{z}{k_0}k_y\right) \qquad (4.23)$$

and the object's spectrum now is

$$\mathscr{F}[i(x, y; z)] = m_2\left(\frac{z}{k_0}k_x, \frac{z}{k_0}k_y\right)\mathrm{OTF}_{\mathrm{osh}}(k_x, k_y; z)\mathscr{F}[I_o(x, y; z)] \quad (4.24)$$

The result of Eq. 4.24 can be interpreted as follows. The second and the third products of the right side of the equation are considered as scanning holographic recording of the object. The multiplication of the resulting holographically recorded object's spectrum by function m_2 is considered as preprocessing of the holographic information of the object. In fact, this novel real-time preprocessing of holographic information has been explored by Schilling and Poon [30]. Real-time preprocessing of holographic information can be performed by simply choosing mask m_2 accordingly. In their paper, they have chosen a filtering operation specified by an off-centered Gaussian function, which has led to an edge extraction of the original image upon holographic reconstruction. We shall further demonstrate edge extraction of 3D images upon reconstruction in this section. Instead of using the planar distributions in Figure 4.2*a* as our 3D object, we choose Figures 4.13*a* and *b* as our planar distributions for the simulations. Hence $I_o(x, y; z)$ in Eq. 4.24 again consists of two planar distributions: the "circle" pattern and the "rectangular" pattern separated by different depths. A mask of the form of a difference-of-Gaussian function is used in the simulations for filtering [31]:

$$m_2(x, y) = \exp[-a(x^2 + y^2)] - \exp[-b(x^2 + y^2)] \quad (4.25)$$

To employ Eq. 4.24, we make the substitutions $x = (z/k_0)k_x$ and $y = (z/k_0)k_y$ in Eq. 4.25 to get the filtering function in the spatial frequency domain:

$$m_2\left(\frac{z}{k_0}k_x, \frac{z}{k_0}k_y\right) = \exp\left[-a\left(\frac{z}{k_0}\right)^2(k_x^2 + k_y^2)\right] - \exp\left[-b\left(\frac{z}{k_0}\right)^2(k_x^2 + k_y^2)\right] \quad (4.26)$$

For some values of *a* and *b*, Figure 4.14 shows a 1D plot of Eq. 4.26 across the origin of the frequency domain. In the simulations, we use these values of *a* and *b* to process the last two terms of the right side of Eq. 4.24. Figure 4.15 shows preprocessed holographic reconstructions at two depths. As difference-of-Gaussian filtering is approximated by an ω^2-Gaussian shape [32], we expect the filter to extract the edge of the 3D image as it is evident from Figure 4.15.

4.7 CONCLUDING REMARKS

We have presented the principles of optical heterodyne scanning and discussed a real-time holographic recording technique known as optical scanning holo-

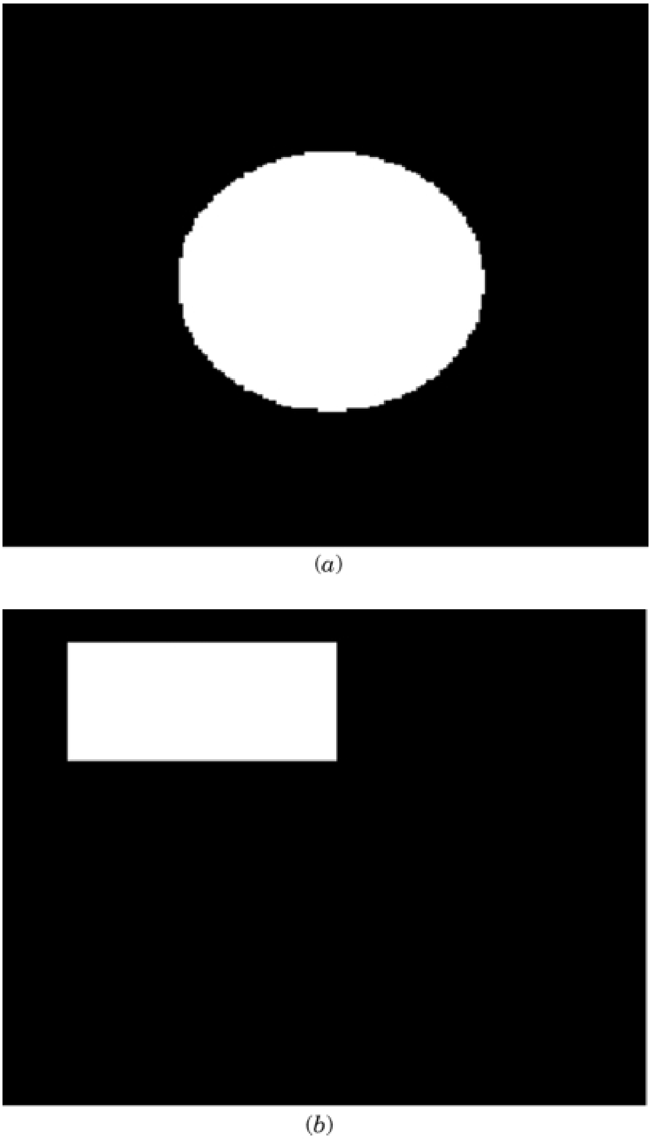


Figure 4.13 (*a*) Circular pattern. (*b*) rectangular pattern.

graphy. By employing OSH, 3D information of an object can be achieved only with a single 2D active optical heterodyne scanning. In the context of OSH, we also have discussed some of its applications, such as twin-image elimination, often encountered in holography, 3D holographic fluorescence microscopy, 3D optical image recognition, and real-time preprocessing of holographic information.

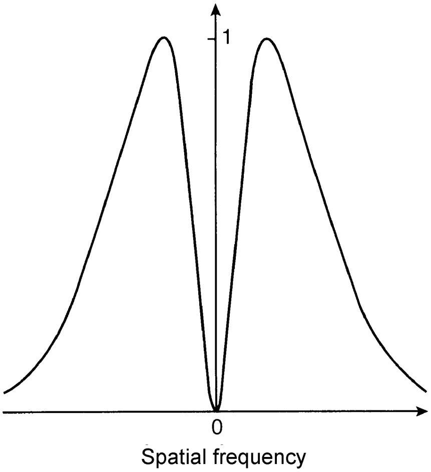
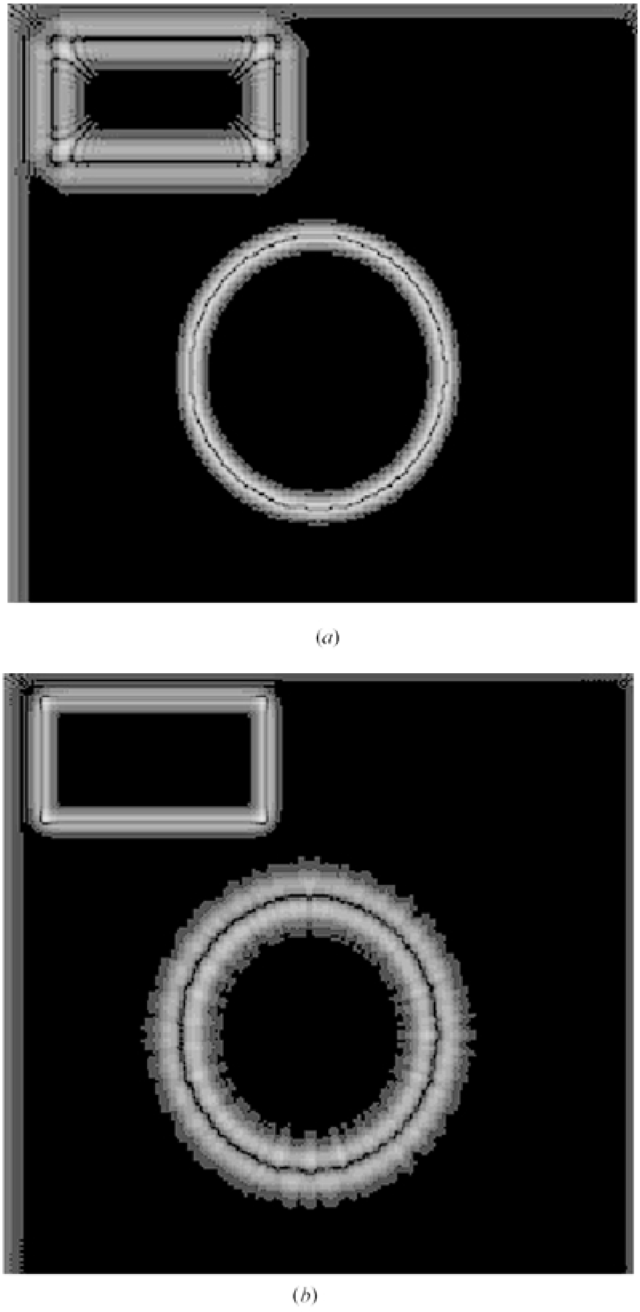


Figure 4.14 Difference-of-Gaussian filtering.

With current advances in spatial light modulators (SLMs) and personal computers, holography is again seriously considered for real-time 3D image display and 3D movies [33–35]. On-axis holography seems to be prevalent nowadays as it presents important advantages such as lower resolution requirement of SLMs and larger viewing angle than that of off-axis alternatives as long as one can find novel ways to eliminate the twin-image noise [36]. Twin-image elimination in OSH requires the use of two real holograms, and computer simulations have shown the effectiveness of the proposed technique. Most recently, experimental verifications of the technique by optically acquiring the two holograms to form a complex hologram and subsequently reconstructing the complex hologram digitally have been reported and 3D image reconstruction without twin-image noise has been demonstrated [37]. It is conceivable that by using two SLMs (one for the display of a sine hologram and the other for that of a cosine hologram), one could achieve all-optical reconstruction without twin-image noise [37, 38]. In passing, we point out that, using polarization optics, a complex hologram of a point source has been generated by adding a cosine and a sine hologram [39].

To our knowledge, for the first time holograms of fluorescent specimens have been recorded by an optical holographic technique as described in this chapter. However, in the "proof-of-principle" stage of the newly developed holographic fluorescence microscopy, the optics has not been optimized for resolution. When imaging takes place in the air, the practical limit for NA is about 0.95. Assuming the setup could be optimized to achieve such a NA, the theoretical limits of resolution for 3D holographic fluorescence microscopy by OSH are $\Delta x \sim 0.3\,\mu\text{m}$ and $\Delta z \sim 0.6\,\mu\text{m}$. For biological applications, such resolutions are needed and should be demonstrated in the future. It is appropriate to point out that the holographic method studied can be applied



(*a*)

(*b*)

Figure 4.15 Preprocessed holographic reconstruction: (*a*) image focused on circle; (*b*) image focused on square.

to 3D biomedical applications as fluorescent imaging of objects in turbid media has been demonstrated recently by Indebetouw et al. [40].

In 3D optical image recognition, 3D image matching is achieved by 2D correlation of holographic data, and hence we call the technique 3D holographic correlation. Computer simulations have verified the proposed idea. However, the correlation output does not give a strong correlation peak when the location of the 3D target is shifted along the depth direction because holographic correlation is basically a 2D correlation process and hence the technique is not z invariant. In fact, the correlation peak is smeared out or broadened by a convolution process, reminiscent of defocused imaging of coherent optical systems [5]. A technique that extracts the 3D location of holographic correlation using the so-called power-fringe-adjusted filter [41, 42] and the Wigner distribution [43, 44] has been proposed recently to achieve z invariance [45]. Experimental verifications of holographic correlation and its newly proposed 3D shift-invariant scheme are still needed.

Regarding preprocessing holographic information, it is important to point out that preprocessing not only can improve the reconstructed image quality [27] but also can lead to the reduction of information content to be recorded in the hologram [46]. In fact, by using one of the masks as the form of the difference-of-Gaussian function as demonstrated in Section 4.6, we can record narrow-band holographic information on the hologram. Indeed, this concept provides a powerful coding scheme for holographic recording as in general one could use any structured beams for scanning by simply modifying either of the masks in the scanning system. Techniques for information reduction or coding are important considerations if, for instance, holographic information is to be transmitted through some channel to a remote site. Of course, another way to deliver the holographic information to a remote site is wireless transmission. This could be easily done with the holographic information right at the output of the photodetector, already riding on a temporal carrier Ω (see Fig. 4.1, right of the bandpass filter). This information could be directly amplified and radiated by an antenna, a viable scheme for TV (or wireless) transmission of holographic information.

In summary, some technological innovations in OSH have been demonstrated, and various potential applications are being studied. Optical scanning holography is a real-time holographic recording technique, and with the use of SLMs, real-time 3D imaging and display might become a reality in the future. We submit that OSH is simple and yet powerful for 3D imaging in general and hope that this chapter will stimulate further research in real-time holography and its various applications.

ACKNOWLEDGMENTS

I would like to thank my student Taegeun Kim for his help with most of the simulations reported in this chapter. The study of optical scanning holography has been supported by the National Science Foundation.

REFERENCES

1. T.-C.Poon and A. Korpel, "Optical transfer function of an acousto-optic heterodyning image processor," *Opt. Lett.* **4**, 317–319 (1979).
2. C. J. Kuo (Ed.), *Opt. Eng.* **35**, special issue on "Electronic Holography" (1996).
3. A. Korpel, "Acousto-optics," in R. Wolfe (Ed.), *Applied Solid State Science*, Vol. 3, Academic, New York, 1972.
4. P. P. Banerjee and T.-C. Poon, *Principles of Applied Optics*, Richard D. Irwin, 1991.
5. T.-C.Poon and T. Kim, "Optical image recognition of three-dimensional objects," *Appl. Opt.* **38**, 370–381 (1999).
6. G. Indebetouw and T.-C. Poon, "Novel approaches of incoherent image processing with emphasis on scanning methods," *Opt. Eng.* **31**, 2159–2167 (1992).
7. K. Doh, T.-C. Poon, M. H. Wu, K. Shinoda, and Y. Suzuki, "Twin-image elimination in optical scanning holography," *Opt. Laser Technol.* **28**, 135–141 (1996).
8. T.-C. Poon, "Scanning holography and two-dimensional image processing by acousto-optic two-pupil synthesis," *J. Opt. Soc. Am. A* **2**, 621–627 (1985).
9. K. Shinoda, Y. Suzuki, M. Wu and T.-C. Poon, "Optical heterodyne scanning type holography device," U.S. Patent 5,064,257 (1991).
10. B. D. Duncan and T.-C. Poon, "Gaussian beam analysis of optical scanning holography," *J. Opt. Soc. Am. A* **9**, 229–239 (1992).
11. T.-C. Poon, M. Wu, K. Shinoda, and Y. Suzuki, "Optical scanning holography," *Proc. IEEE* **84**, 753–764 (1996).
12. P. Davidovits and M. D. Egger, "Scanning optical microscope," U.S. Patent 3,643,015 (1972).
13. T. Wilson and C. Sheppard, *Theory and Practice of Scanning Optical Microscopy*, Academic, New York, 1984.
14. J. Pawley, "Fundamental limits in confocal microscopy," in *Handbook of Biological Confocal Microscopy*, 2nd ed., Plenum, New York, 1995, Chapter 2.
15. B. W. Schilling, T.-C. Poon, G. Indebetouw, B. Storrie, K. Shinoda, Y. Suzuki, and M. H. Wu, "Three-dimensional holographic fluorescence microscopy," *Opt. Lett.* **22**, 1506–1508 (1997).
16. A. B. Vander Lugt, "Signal detection by complex spatial filtering," *IEEE Trans. Inf. Theory* **IT-10**, 139–145 (1964).
17. C. S. Weaver and J. W. Goodman, "A technique for optically convolving two Functions," *Appl. Opt.* **5**, 1248, 1249 (1966).
18. M. A. Karim and M. S. Alam (Eds.), *Opt. Eng.* **37**, special Issue on "Advances in recognition technique" (1998).
19. T.-C. Poon, T. Hara, and R. Juday (Eds.), *Appl. Opt.* **37**, special issue on "Spatial Light Modulators: Research, Development, and Application" (1998).
20. W. W. Stoner, W. J. Miceli, and F. A. Horrigan, "One-dimensional to two-dimensional transformations in signal correlation," in W. T. Rhodes, J. R. Fineup, and B. E. A. Saleh (Eds.), *Transformations in Optical Signal Processing*, *Proceedings of SPIE*, Vol. 373 (1981).

21. W. T. Rhodes, "The falling raster in optical signal processing," in W. T. Rhodes, J. R. Fineup, and B. E. A. Saleh (Eds.), *Transformations in Optical Signal Processing*, *Proceedings of SPIE*, Vol. 373 (1981).

22. A. E. Siegman, "Two-dimensional calculations using one-dimensional arrays, or 'Life on the Skew,'" *Comput. Phys.*, Nov./Dec., 74–75 (1988).

23. J. Hofer-Alfeis and R. Bamler, "Three- and four-dimensional convolution by coherent optical filtering," in W. T. Rhodes, J. R. Fienup, and B. E. A. Saleh (Eds.), *Transformations in Optical Signal Processing*, *Proceedings of SPIE*, Vol. 373 (1981).

24. J. Rosen, "Three-dimensional optical Fourier transform and correlation," *Opt. Lett.* **22**, 964–966 (1997).

25. E. Paquet, P. Garcia-Martinez, and J. Garcia, "Tridimensional invariant correlation based on phase-coded and sine-coded range images," *J. Opt.* **29**, 35–39 (1998).

26. Y. B. Karasik, "Evaluation of three-dimensional convolution by use of two-dimensional filtering," *Appl. Opt.* **36**, 7397–7401 (1997).

27. G. Molesini, D. Bertani, and M. Cetica, "In-line holography with interference filters as Fourier processors," *Opt. Acta* **29**, 479–485 (1982).

28. C. S. Vikram and M. L. Billet, "Gaussian beam effects in far-field in-line holography," *Appl. Opt.* **22**, 2830–2835 (1983).

29. C. Ozkul, D. Allano, and M. Trinite, "Filtering effects in far-field in-line holography," *Opt. Eng.* **25**, 1142–1148 (1986).

30. B. W. Schilling and T.-C. Poon, "Real-time preprocessing of holographic information," *Opt. Eng.* **34**, 3174–3180 (1995).

31. T.-C. Poon and K. C. Ho, "Real-time optical image processing using difference-of-Gaussians wavelet," *Opt. Eng.* **33**, 2296–2302 (1994).

32. T.-C. Poon, J. W. Park, and G. Indebetouw, "Optical realization of textural edge extraction," *Opt. Commun.* **65**, 1–6 (1988).

33. P. St. Hilaire, "Holographic video: The ultimate visual interface?" *Opt. Photon. News*, August, p. 35 (1997).

34. T.-C. Poon, K. Doh, B. W. Schilling, K. Shinoda, Y. Suzuki, and M. H. Wu, "Holographic three-dimensional display using an electron-beam addressed spatial light modulator," *Opt. Rev.* **4**, 567–517 (1997).

35. B. P. Ketchel, C. A. Heid, G. L. Wood, M. J. Miller, A. G. Mott, R. J. Anderson, and G. J. Salamo, "Three-dimensional color holographic display," *Appl. Opt.* **38**, 6159–6166 (1999).

36. R. Piestun, J. Shamir, B. Wesskamp, and O. Bryngdahl, "On-axis computer-generated holograms for three-dimensional display," *Opt. Lett.* **22**, 922–924 (1997).

37. T.-C. Poon, T. Kim, G. Indebetouw, B. W. Schilling, M. H. Wu, K. Shinoda, and Y. Suzuki, "Twin-image elimination experiments for three-dimensional images in optical scanning holography," *Opt. Lett.* **25**, 215–217 (2000).

38. K. Shinoda, T.-C. Poon, M. Wu, and Y. Suzuki, "Twin-image elimination apparatus and method," U.S. Patent 5,805,316 (1998).

39. S.-G. Kim, B. Lee, and E.-S. Kim, "Removal of bias and the conjugate image in incoherent on-axis triangular holography and real-time reconstruction of the complex hologram," *Appl. Opt.* **36**, 4784–4791 (1997).

40. G. Indebetouw, T. Kim, T.-C. Poon, and B. W. Schilling, “Three-dimensional location of fluorescent inhomogeneities in turbid media using scanning heterodyne holography,” *Opt. Lett.* **23**, 135–137 (1998).

41. X. W. Chen, M. A. Karim, and M. S. Alam, “Distortion-invariant fractional power fringe adjusted joint transform correlation,” *Opt. Eng.* **37**, 138–143 (1998).

42. M. S. Alam, X.-W. Chen, and M. A. Karim, “Distortion-invariant fringe-adjusted joint transform correlation,” *Appl. Opt.* **36**, 7422–7427 (1997).

43. T. A. C. M. Claasen and W. F. G. Mecklenbrauker, “The Wigner distribution — a tool for time-frequency signal analysis. Part 1: Continuous-time signals,” *Philips J. Res.* **35**, 217–250 (1980).

44. T. A. C. M. Claasen and W. F. G. Mecklenbrauker, “The Wigner distribution — a tool for time-frequency signal analysis. Part 2: Discrete-time signals,” *Philips J. Res.* **35**, 276–300 (1980).

45. T. Kim and T.-C. Poon, “Extraction of 3D location of matched 3D object using power fringe-adjusted filtering and Wigner analysis,” *Opt. Eng.* **38**, 2176–2183 (1999).

46. B. J. Thompson, “Holographic methods for particle size and velocity measurement — Recent advances,” *Proc. SPIE* **1136**, 308–326 (1989).

Three-Dimensional Holographic Imaging. Edited by Chung J. Kuo, Meng Hua Tsai
Copyright © 2002 John Wiley & Sons, Inc.
ISBNs: 0-471-35894-0 (Hardback); 0-471-22454-5 (Electronic)

CHAPTER 5

Tangible, Dynamic Holographic Images

WENDY PLESNIAK, RAVIKANTH PAPPU, and STEPHEN BENTON

Media Laboratory, Massachusetts Institute of Technology
Cambridge, Massachusetts 02139

Good holograms are bewitching. They command our eyes to their images, to search the marvelous realness of surfaces, textures, and minute detail for some aberration, some visual clue that will dissuade us from seeing as real what we know is not. Yet while they seem to assemble the very molecules of physical matter for us to ponder, they also present a conundrum: The objects they render appear frozen, lifeless, and confounding to our fingertips.

What if we could render these images animate and touchable, as phantom material that was both dynamic and plastic? Such an ultimate display would be both powerful and magical; it would deliver naturally to our spatial proficiencies, inspire our imaginations, and perhaps even provoke our emotions. Of course, such a display does not yet exist, and many challenges to its development remain. But while current technology still leaves us in the shadow of such a goal, we are beginning to see it as more real than chimerical.

To this end, we will describe our first experiments with tangible, dynamic holographic images. Our prototype system, called the holo–haptic system, comprises a sizable arsenal of computers and both commercial and custom hardware. The visual images it produces for these experiments are monochromatic and postcard sized and depict only simple geometries. The haptic images it produces are felt and shaped with a hand-held device. Thus, the trappings of engineering are anything but transparent to the experience, and the demonstrations themselves are artistically unsophisticated. But by using this agglomeration of technology in simple demonstrations, we can feel and sculpt three-dimensional shapes made only of light — which inches us closer to where we want to go.

5.1 INTRODUCTION

People perceive, think, and act quite naturally in a spatial theater. At a glance, we understand the layout of our environment, the locations and shapes of objects within reach. We apprehend and manipulate these objects without a thought, skillfully, and sometimes even artfully. Yet, even while great attention is turned toward human-centered engineering and interaction design in computer-based applications, the full exploratory and manipulative dexterity of the hand and the sensorimotor advantages of binocular vision are not commonly pressed into service. The reasons for this are twofold: The point-and-click desktop paradigm still dominates our notion about how to interact with computers, and it remains challenging to design and build the new sensing and display technologies that will enable a new way of working.

Obviously, it is not always desirable to take input from mouse-twiddled graphical user interface (GUI) buttons, sliders, and the keyboard or to always deposit output into a flat display window, though these are the archetypal input–output (I/O) channels. Even using virtual reality (VR)–style head-mounted displays combined with body tracking seems now a cumbersome and tired approach. Yet new technologies and methods of using them continue to emerge, and with them comes the promise of new and imaginative tools that cater to our natural, spatial strategies for doing things. Among these technologies are displays and sensors that are minimally or not at all body borne and that allow us to use two eyes and hands within manipulatory, interactive, or reactive workspaces.

The attending interaction design possibilities might better serve human perception and performance: For instance, two-handed input is known to provide some manual and cognitive benefits [1]; and including binocular vision and motion parallax helps a viewer to better understand shape and layout and to plan and execute prehensile movement in the workspace [2, 3]. The coaction of eye and hand in manipulatory space might even be reinforced by spatially colocating the manual work volume with the visual one or by merging the controller and display entirely.

Diminishing the evidence of technology situated *between* our action and a computer-based system's response also shrinks our psychological awareness of the symbolic interpretation, instruction, and actuation that occurs there. In the real world, whether we are hammering a nail or squeezing a block of clay, most actions we exert on objects seem so directly and inextricably linked to the their reactions that we do not normally distinguish between what is *input* and *output*. Mixing this kind of "Newtonian" interaction with the freedom to express system response as anything from physically based to purely poetic offers us rich new territory for engineering and interaction design.

It is within this context that we situate our recent work with electro-holography and force feedback. The specific project we describe here provides a physically based, though stylized, emulation of a task that has a real-life analogue — the lathing of rotating stock. We display a free-standing holographic image combined with a force image that can be directly felt and carved with a hand-held tool. The visual and force images are spatially co-located, engaging eye and hand together and attempting to veil the computation nested between input and display. Many other research efforts, employing a wide range of technologies, have contributed a diverse set of applications related by these elements of interaction and workspace design. Some depart from strictly physically based simulation and exploit the malleability of the computational rules translating action into response; others, like ours, simply try to simulate a task already practiced in the physical world. The following section provides a brief overview of some of these projects, which have either inspired or enabled our holo–haptic work.

5.2 CONTEXT

One project that is thematically related to our work is the "Virtual Lathe" described and presented at SIGGRAPH'92 by Deering [4]. In this demonstration, a head-tracked stereo display showed a computer-graphic stock, spinning about its long axis, which a person could interactively carve using a rod-shaped three-dimensional (3D) mouse. The demonstration underscored popular interest in having a more direct way to interact with virtual prototyping systems. But without force feedback, the Virtual Lathe provided neither the important sense of contact with the simulated stock, nor the feel of carving.

A wide variety of virtual reality (VR) and augmented reality (AR) application areas such as telesurgery, entertainment, and maintenance analysis and repair do employ computational haptics and stereo computer graphics to feel, see, and interact with data. Most existing demonstrations offset the visual and manual workspaces, so that a person manipulates her hand in one place while visually monitoring its action and the system's response on another separate display. Fewer attempts to conjoin eyes and hands in a coincident workspace have been reported. Two compelling examples are Boston Dynamics's virtual reality surgical simulator [5] and the nano Workbench at the University of North Carolina (UNC) at Chapel Hill [6]. Both of these systems use force feedback and head-tracked stereo visual display with liquid crystal display (LCD) shutter goggles, and they let the hand-held device appear to operate directly on the visually displayed data.

Another interesting system that incorporates computational haptics (but no stereo viewing), called the WYSIWYF (What You See Is What You Feel) display [7], has been demonstrated at Carnegie Mellon University. Here the visual display behaves like a movable "magic window" interposed between the viewer's eyes and hand, and through which the hand can be seen interacting with a virtual, tangible scene. The system uses a haptic manipulator and image compositing to present the computer graphically rendered scene overlayed by a video image of the operator's hand/arm and the accompanying force model. Without properly representing occlusion, however, WYSIWYF is unable to always display the correct visual relationship between hand and scene, and it also provides only monocular cues to depth.

Rather than using computational haptic feedback, actual “wired” physical controllers can be employed. These interface objects act as physical handles for virtual processes, may have on-board sensing, computation and intercommunication, and can be hand manipulated and spatially commingled with visual output. A person using them can enjoy the simplicity of interacting with physical objects while observing the outcome displayed on or near the controller, in a separate location, or in the ambient environment. Several of these efforts have been presented recently: for instance, EuroPARC’s Digital Desk [8], MIT Media Lab’s metaDESK [9], and IlluminatingLight [10]. Yet, while providing whole-hand interaction and richly programmable visual feedback, using physical controllers restricts the bulk and tactual feel as well as the physical *behaviors* of the input devices to be bound by physical mechanics.

Conversely, a system that has programmable haptic display but restricted visual output is Dimensional Media’s High Definition Volumetric Display. This system incorporates force feedback and a reimaging display, which employs optical components to relay and composite images of already existing 3D objects and/or 2D display screens. As a result, using a force feedback device to interact with the optical output (which can appear strikingly realistic) to modify the geometry of the displayed object is not possible.

Holography is another optical display technique that can project spatial images into a viewer’s manual workspace. The combination of haptics and holography was first investigated by researchers at De Montfort University in an object inspection task [11]. In this work, visual display was provided by a reflection transfer hologram that presented an aerial image of a control valve while a computer-controlled tactile glove provided coincident haptic display of the same data. Subsequent informal experiments in combining reflection transfer holograms with force feedback were also performed at the MIT Media Laboratory’s Spatial Imaging Group. Since reflection holograms require front overhead illumination for image reconstruction, the interacting hand could literally block the holographic image in both of these holo–haptic efforts.

This problem was addressed by employing full-parallax edge-illuminated holograms in combination with a force- feedback device for the inspection of 3D models [12]. The edge-illuminated hologram format allowed hand movements in the visual workspace in front of the hologram plane without blocking illumination (Fig 5.1). Thus, a viewer could haptically explore the spatially registered force model while visually inspecting the holographic image details over a wide field of view. The DeMontfort and MIT holo–haptic displays were static, however; no dynamic modification could be made to the displayed image.

5.3 HAPTICS AND HOLOGRAPHIC VIDEO

Haptics is a general term referring to elements of manual interaction with an environment. This interacting may be done by either human hands or sensing machines in an environment that may be physical or simulated. In our case,

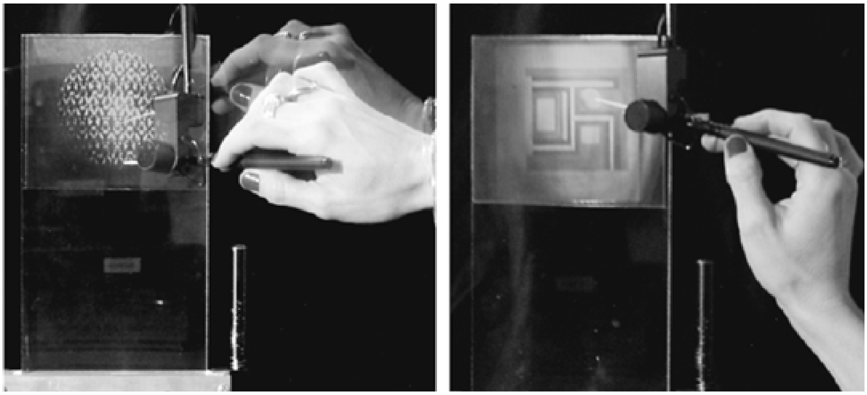


Figure 5.1 Edge-illuminated haptic holograms.

the interactions are accomplished by human hands that sense force information as they explore and manipulate.

By bringing together computational haptics and electroholography, we hoped to render simple scenes that could be seen, felt, and *modified* in the manual workspace. Of course, there is little value in demonstrating a multimodal simulation with a stable haptic simulation, but a visual frame rate of only one frame every *few* seconds; electroholography currently suffers from limited computation and communication bandwidth, leading to just this problem. To increase our visual update rate, we chose to take two simplifying measures: First, rather than being entirely recomputed, the hologram itself would be updated only in regions of change; second, the underlying object geometry would only be modified by interaction in a constrained way.

For our simulated object and demonstration, we chose a cylindrical stock spinning about its vertical axis that can be carved along its length into an arbitrary surface of revolution. The holographic and force images of this model are spatially and metrically registered and provide a free-standing 3D representation of the model to be lathed in the workspace (Fig. 5.2). For positional tracking and force display, we use the Phantom Haptic Interface from Sensable Technologies and for visual display the MIT second-generation holographic video (holovideo) system. This combination of display devices gives us a visuomanual workspace of about $150 \times 75 \times 75$ mm^3.

The haptic simulation is produced by modeling the object as a cubic B-spline surface with compliance, surface texture, static and dynamic friction, mass, and rotational speed. The endpoint of the Phantom is monitored for contact with the model, in which case the appropriate restoring force is computed and displayed. The holograms are produced by first populating the object model with a collection of spherical emitters and then computing their interference with a collimated reference wave; after normalizing, this pattern is sent to the holovideo display. In the combined holo–haptic workspace, a

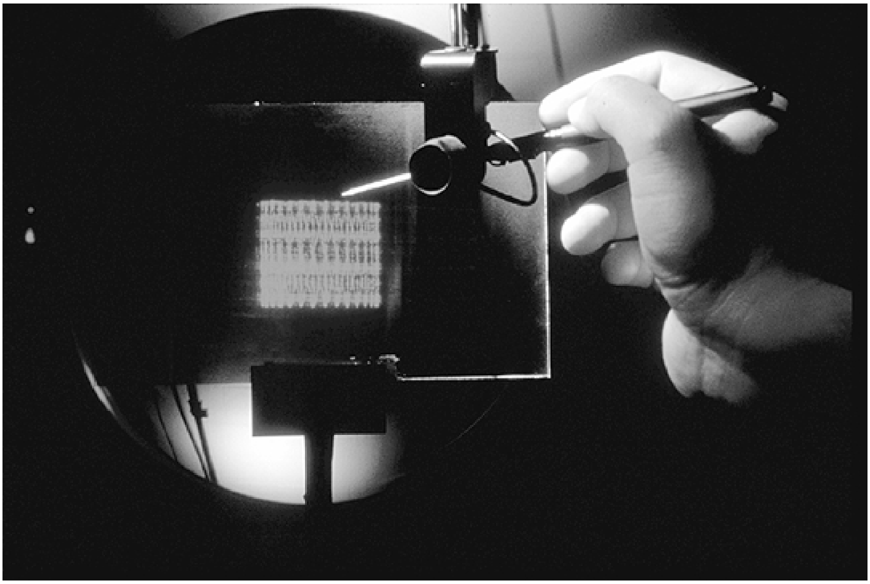


Figure 5.2 Dynamic holo–haptic lathe.

person can inspect and carve the holographic image while both seeing the image change and feeling accompanying forces. Rather than computing a brand new hologram each time the model is modified, the final hologram is assembled from a set of five *precomputed* ones. A more detailed description of holographic and haptic modeling and display and of the final system follows.

5.4 HOLOGRAPHIC VIDEO SYSTEM ARCHITECTURE

As previously mentioned, we employ the second generation of holovideo in this work. This system is capable of displaying monochromatic, horizontal-parallax-only (HPO) images in a volume of $150 \times 75 \times 75$ mm^3, and with a viewing angle of 30°. The 3D image produced by holovideo supports the most important depth cues: stereopsis, motion parallax, occlusion, and many other pictorial and physiological cues as well.

Generally speaking, the second-generation system accepts two inputs: a computer-generated hologram (CGH) and light. Its output is a free-standing, 3D holographic image whose visual and geometric characteristics depend on how the CGH was computed. Each CGH contains 36 megasamples, at 1 byte per sample, apportioned into 144 lines of 256 kilosamples each. The methods of computing these holograms, delivering them to the display, and producing an image are described in the following sections.

5.4.1 Optical Pipeline

The design strategy [13] for the second-generation holovideo display was to exploit parallelism wherever possible, both optically and electronically, such that the approach would be extensible to arbitrarily large image–sized displays. To produce an image volume of $150 \times 75 \times 75$ mm^3, two 18-channel acousto-optic modulators (AOMs) were used, with AOM channels modulating beams of helium–neon laser light in parallel. Six tiled horizontal mirrors scan across the output, matched to the speed of the signal in the AOM, such that the relayed image of the diffraction pattern in the AOM is stationary. As the mirrors scan from left to right, one AOM provides 18 lines of rastered image. When the mirrors return from right to left, the second crossfired AOM provides the next 18 lines of rastered image. A vertical scanner images each 18-line pass below the previous one, with eight horizontal scans in all, providing $18 \times 8 = 144$ vertical scan lines in the final image.

This resulting image is HPO, with video resolution in the vertical direction and holographic resolution in the horizontal direction. To match the shear-mode active bandwidth of the tellurium dioxide AOM crystal, we need to produce a signal with a bandwidth of approximately 50 MHz. So that the output sampling satisfies the Nyquist criterion, we use a pixel clock of 110 MHz. As mentioned earlier, each horizontal line of the display is 256 Kbytes of holographic fringe pattern; 144 of these *hololines* make up the final hologram, yielding 36 Mbytes of information per frame. Since the display has no persistence, it must be refreshed at 30 Hz, thus requiring an average data rate of 1 G pixel per second from the frame buffers.

An 18-channel reconfigurable frame buffer was required to drive the display. While no known commercially available frame buffer is capable of the required data rate, we were able to adapt the Cheops Imaging System [14] to the task. Cheops is a data flow architecture digital video platform developed at the MIT Media Laboratory. The Holovideo Cheops system provides six synchronized frame buffers to drive our 256 Kbyte × 144 display as well as a high-speed interface to host processors and a local data flow processing card for decoding of encoded or compressed image formats. The entire optical pipeline of the second-generation system is depicted in Figure 5.3.

5.4.2 Computational Pipeline

The production of a traditional optical hologram or holographic stereogram requires an object-modulated wave and an interfering reference wave and results in a real fringe pattern recorded in a photosensitive medium. In computational holography, we start with a three-dimensional, mathematical or computer-graphic model of a scene and compute a holographic fringe pattern from it. In this case, the fringe pattern is simply an array of numbers, but when written to a spatial light modulator, it can diffract input light in a directed fashion to produce a spatial image of the original model. We will describe two

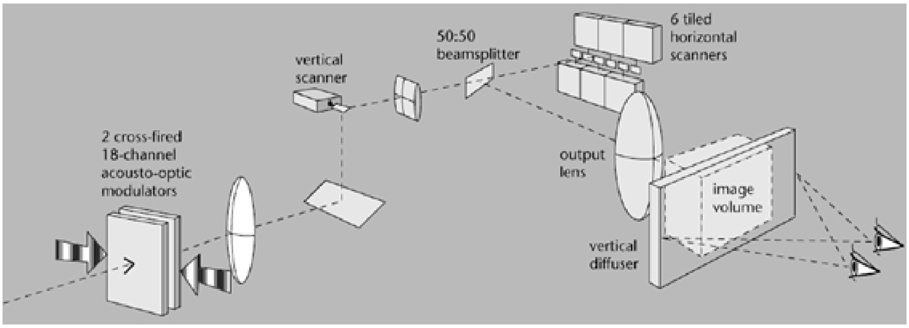

6 tiled horizontal scanners
50:50 beamsplitter
vertical scanner
2 cross-fired 18-channel acousto-optic modulators
output lens
image volume
vertical diffuser

Figure 5.3 Optical pipeline.

distinct ways of generating a CGH from the same object model, the interference modeling and the stereogram modeling approach. Both methods were tested to generate holographic images for the holo–haptic lathe. In both cases, the initial object model and the final display pipeline are identical but the intervening algorithmic subsystems are distinct.

5.4.2.1 Interference Modeling Approach A "fully computed hologram" is a fringe pattern resulting from the computational modeling of the interference process. In the ideal case, the wavefront reconstructed by the display would be perceptually indistinguishable from that generated by the original object. In practice, however, there are significant departures from this ideal. These departures are a consequence of shortcuts taken to make the computation and display of fringe patterns tractable and also due to our technological limitations.

As such, computation of the fringe pattern proceeds through three stages: scene modeling, occlusion processing, and interference modeling. The scene modeling subsystem uses standard computer-graphics techniques to generate an intermediate wireframe or shaded polygonal description of the scene. To represent the object as a collection of point sources, we populate each polygon with a series of self-luminous points and assign a location, amplitude, and initial phase to each of them. This particular representation for the object field was chosen because point sources have a particularly simple mathematical form, and interference patterns arising from them are fairly simple to compute. Artifacts of spatially and temporally coherent illumination are diminished by randomly varying the interpoint spacing, which is on the order of 10 points/mm, or by assigning uniformly distributed random initial phases. Each point is then assigned to one of the 144 hololines according to its vertical projection onto the hologram plane. Since we are computing HPO holograms, each object point will contribute only to its assigned hololine.

The sorted points are then passed to an occlusion processing system, described in more detail elsewhere [15], which computes for each point radiator the set of regions on the hololine to which it contributes. With this information, the fringe pattern can be rendered. Since there may be hundreds or hundreds of thousands of points in a scene, computation time in all stages of hologram generation (from modeling through fringe rendering) is highly dependent on object complexity.

Fringe rendering is accomplished by approximating the classical interference equation for each sample on the hololine. The contribution from each subscribing point radiator is totaled, and the intensity at the current sample is determined. The final result is normalized and quantized to the range 0–255 to represent each sample as a 1-byte quantity. This hologram computation is currently implemented on an SGI Onyx workstation. The final hologram, which takes on the order of 10 min to compute (depending on object complexity), is then dispatched to Cheops via a SCSI link at about 10 Mbits per second; the final image is then viewable on the display. The entire computational pipeline, from modeling to display, is shown in Figure 5.4.

5.4.2.2 Stereogram Modeling Approach A holographic stereogram uses wavefront reconstruction to present a finite number of perspective views of a scene to an observer, whereas fully computed holograms offer continuous parallax through the view zone. A stereogram's discretization of parallax views results in a discrete approximation to the ideal reconstructed wavefront [16], and this approximation permits a considerable increase in computational speed.

The underlying principle in our approach to stereogram computation is to divide the process into two independently determinable parts. The diffractive

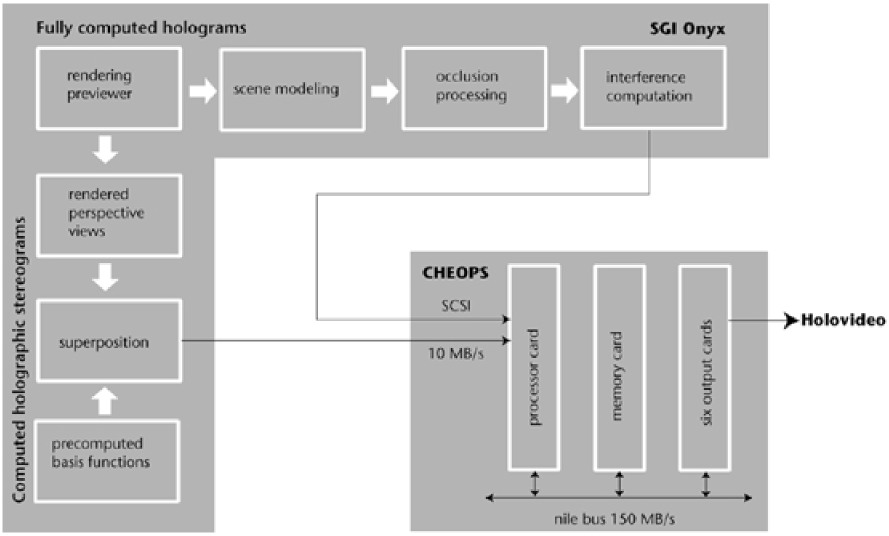


Figure 5.4 Generalized computational pipeline.

part is unvarying for a given display geometry and can be precomputed, stored, and retrieved from a table when required. This computation is achieved by positing a set of basis functions, each of which should diffract light in a certain direction. Basis functions are currently determined using an iterative optimization technique with initial conditions that describe spatial and spectral characteristics of our particular display and view zone [17]. Each of these basis functions is scaled in amplitude by a coefficient determined from the perspective views. The superposition of these scaled basis functions results in a holographic element, or *hogel*—a small segment of the hololine. If the perspectives change, only the new scaling coefficients must be determined before the superposition can be carried out again.

Computer-generated holographic stereograms admit input data from a variety of sources: from computer graphic renderers or shearing and recentering optical capture systems, for instance. Here, occlusion processing is bundled in for free. But despite the versatility and computational appeal of computed stereograms, they suffer from certain shortcomings; the abrupt change in phase from one perspective to another results in perceptible artifacts including enhanced speckle when coherent illumination is used. Additionally, increasing the number of perspectives while using the same 1 byte/sample framebuffer leads to a decrease in the dynamic range available to each basis function and a consequent decrease in diffraction efficiency.

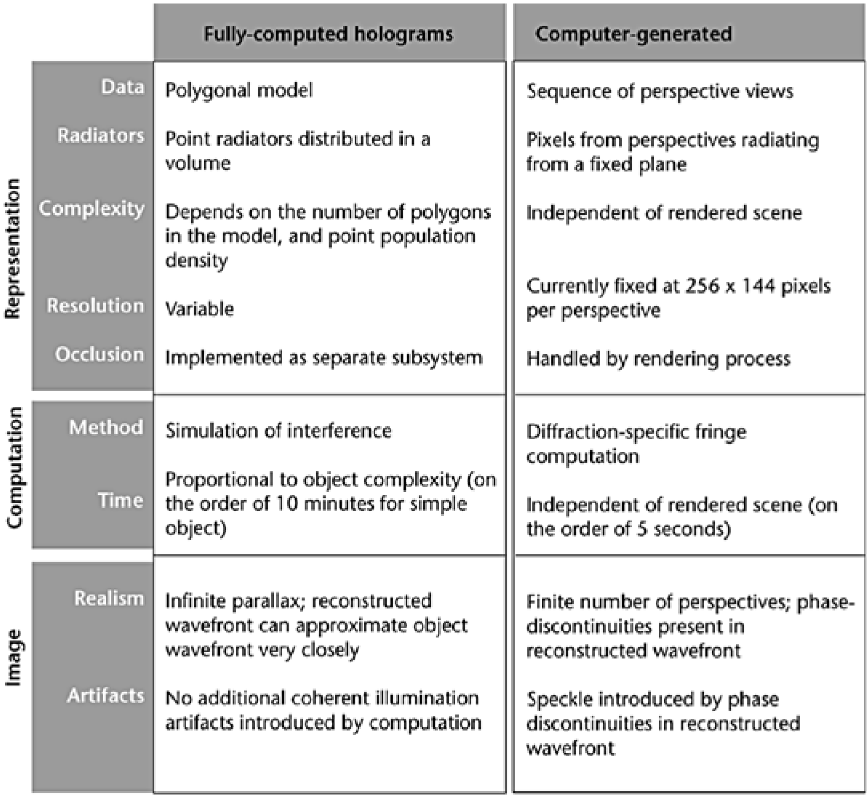
Computer-generated holographic stereograms are currently computed from 32 prerendered perspective views using either an SGI Onyx or Origin and sent to Cheops via SCSI for display. The stereogram computing pipeline is also shown in Figure 5.4. The time required to compute a stereogram is on the order of 1–6 s, depending on computational platform, and the computation is roughly independent of scene complexity.

A detailed comparison between computed stereograms and fully-computed fringe patterns as well as the computation time associated with each is given in Figure 5.5. The figure makes evident the trade-off between image realism and the speed of hologram generation. No matter which method is used, the fundamental issues of computation and communication bandwidth must still be addressed. Developing more efficient representations for the fringe pattern or techniques for decoupling fringe computation from the complexity of the object remain worthwhile areas of investigation.

5.5 HOLO–HAPTIC LATHE IMPLEMENTATION

5.5.1 System Overview

Three separate processes support our holo–haptic display: a *haptics module*, which performs force modeling; the *holovideo module*, which precomputes



		Fully-computed holograms	Computer-generated
Representation	Data	Polygonal model	Sequence of perspective views
	Radiators	Point radiators distributed in a volume	Pixels from perspectives radiating from a fixed plane
	Complexity	Depends on the number of polygons in the model, and point population density	Independent of rendered scene
	Resolution	Variable	Currently fixed at 256 x 144 pixels per perspective
	Occlusion	Implemented as separate subsystem	Handled by rendering process
Computation	Method	Simulation of interference	Diffraction-specific fringe computation
	Time	Proportional to object complexity (on the order of 10 minutes for simple object)	Independent of rendered scene (on the order of 5 seconds)
Image	Realism	Infinite parallax; reconstructed wavefront can approximate object wavefront very closely	Finite number of perspectives; phase-discontinuities present in reconstructed wavefront
	Artifacts	No additional coherent illumination artifacts introduced by computation	Speckle introduced by phase discontinuities in reconstructed wavefront

Figure 5.5 Comparing fully-computed holograms and computed holographic stereograms.

holograms and drives rapid local holographic display updates based on changes to the model; and the *workspace resource manager* (WRM), which links the two. More specifically, the WRM is notified by the haptics module of geometry changes imparted to the model by an interacting user. It determines the regions of the hologram affected by new model changes and the closest visual approximation to the haptic change, and then makes requests to the holovideo module for the corresponding local hologram updates. The holovideo module assembles the updated chunk of hologram from a set of precomputed holograms and swaps them into the one currently displayed. From the point of view of a user, who is holding the stylus and pressing it into the holographic image, a single multimodal representation of the simulation can be seen and felt changing in response to the applied force. The system architecture is shown in Figure 5.6.

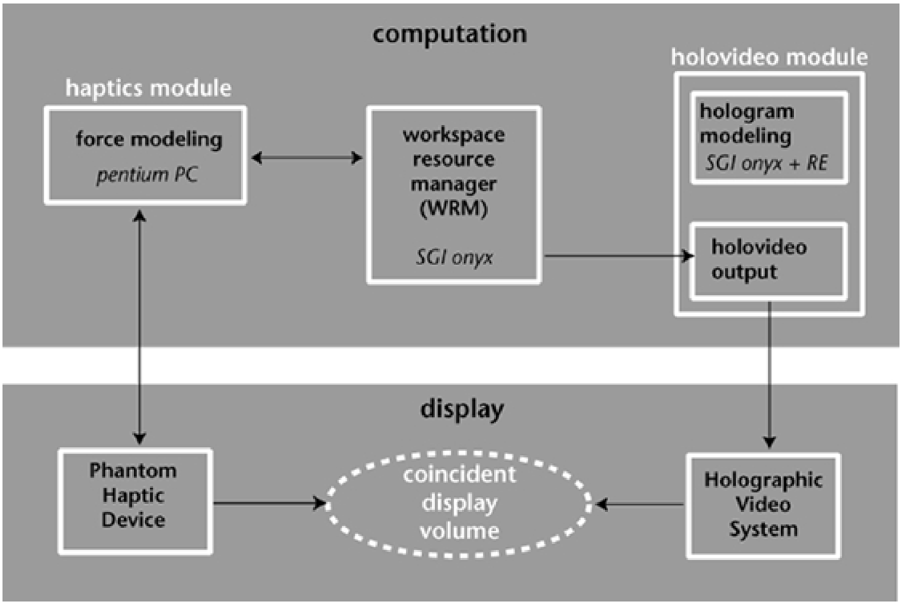


Figure 5.6 Dynamic holo–haptic system architecture.

5.5.2 Haptic Modeling and Display

As mentioned previously, we use the Phantom haptic device, which interfaces to the body via a hand-held stylus. The stylus can be used to probe a simulated or mixed-reality scene and displays force, when appropriate, back to the user. Six encoders on the device are polled to compute the stylus tip's position, and this information is checked against the geometry of our haptic stock. If contact is detected between the stylus tip and the stock model, appropriate torque commands are delivered to the device's three servomotors; thus a restoring force is felt by the hand holding the stylus. The device has an addressable workspace of about $290 \times 400 \times 560$ mm^3.

The haptic stock, initially and in subsequent stages of carving, is represented as a surface of revolution with two caps. It has a mass of 1 g, an algorithmically defined vertical grating (with 1 mm pitch and 0.5 mm height) as a surface texture, static and dynamic frictional properties, and stiff spring bulk resistance. The haptic stock rotates about its vertical axis at 1 rev/s and straddles a static haptic plane (which spatially corresponds with the output plane of the holovideo optical system). The haptic plane is modeled with the same bulk and frictional properties as the stock.

The haptic stock maintains rotational symmetry about its vertical axis initially and in all subsequent stages of carving. Its radius profile is represented by a cubic B-spline curve; initially, all control points, **P**, are set to the same radius value (25 mm) to let us begin lathing a cylinder. Control points on the

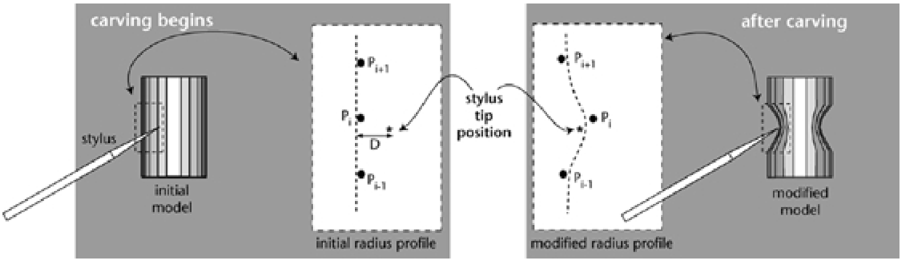


Figure 5.7 Lathing the haptic model.

radius profile curve are modified as force is exerted on the stock at height h, between control points $\mathbf{P}_i$ and $\mathbf{P}_{i+1}$. A new radius for the entire surface of revolution at this height can be computed by evaluating the nonuniform rational B-spline formulation, and this change is immediately reflected in the model geometry. The stock can be felt to spin beneath the user's touch, and when pressed with enough force (when the surface has been penetrated by some threshold distance D), its surface deforms (Fig. 5.7).

The haptic model can be carved away from its original radius (25 mm) down to a minimum radius (15 mm); the minimum radius is enforced so that once the stock has deformed this much, the control points will update no further. The control point density was derived though trial and error to enable fairly intricate carving without permitting deep notches, which introduce instabilities into the haptic model.

5.5.3 Precomputed Holograms and Limited Interaction

Ideally, haptic interaction could arbitrarily modify the object model, and realistic visual feedback would be displayed in concert with carving. However, as mentioned earlier, we must take several simplifying measures to achieve near-real-time simulation. First, we limit the way the underlying object geometry can be modified by working with an object model that always maintains rotational symmetry, as described above. Second, we precompute a set of holograms and reassemble the final hologram from them; the resulting final hologram displays an *approximation* to the model's shape.

The haptic model of the stock is continuous and varies smoothly along its carved profile. In our implementation, translating changes in this model to hologram updates requires going through an intermediate representation. This intermediate representation (dubbed the "stack") treats the stock as a pile of 120 disks, each of some quantized radius nearest to the haptic stock radius at a corresponding height. We select from a set of five radii, ranging from the initial radius of the haptic stock, down to the minimum radius permitted by carving. The number of disks in the stack represents the number of display lines occupied by the final holographic image, and also corresponds to the physical

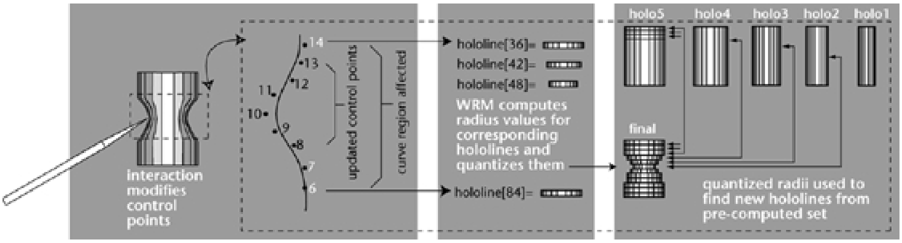


Figure 5.8 Method of propagating haptic model changes to holovideo display.

height of the haptic model. It is an image of the *stack* that is reconstructed holographically, yielding a visual image that is an approximation to the accompanying force image.

To efficiently *assemble* a hologram of the stack, we first pre-compute a set of five holograms of cylinders, each having a different radius determined from the set mentioned above. A hologram of the stack is assembled by using appropriate lines from each of the pre-computed holograms at appropriate locations in the final one. For instance, if a region in the middle of the haptic stock has abruptly been shaved down to its minimum radius, only the middle lines on the holovideo display are changed by swapping in corresponding lines from the minimum-radius hologram. The process is depicted in Figure 5.8.

Since we are precomputing holograms for this work and thereby relaxing our need for rapid computation, we chose to use the more realistic images afforded by the fully computed method. While holographic stereograms are faster to produce, the algorithm described earlier produces image artifacts, and the final holographic images lack the sharpness and dynamic range that enhance the appearance of solidity.

5.6 RESULTS

When an operator carves the holographic stock with the Phantom, the hologram image changes due to the force apparently applied by the tip of the stylus. The resulting shape can be explored by moving the stylus tip around the surface without exerting too much force (Fig. 5.9). Physical objects in the workspace may also be explored, so that both physical and simulated forces can be displayed to the operator alternatively in the same workspace. When the operator maintains the correct viewing position for holovideo, the perception of a single multimodal stimulus is convincing, and the experience of carving a hologram is quite inspiring. Additionally, once the model has been carved into “finished” form, it can be dispatched to a 3D printer that constructs a physical hard copy of the digital design (Fig. 5.10).

Of course, this demonstration is still a prototype; it exhibits low frame rate (10 frames/s), lag (0.5 s), and many of the intermodality conflicts described in

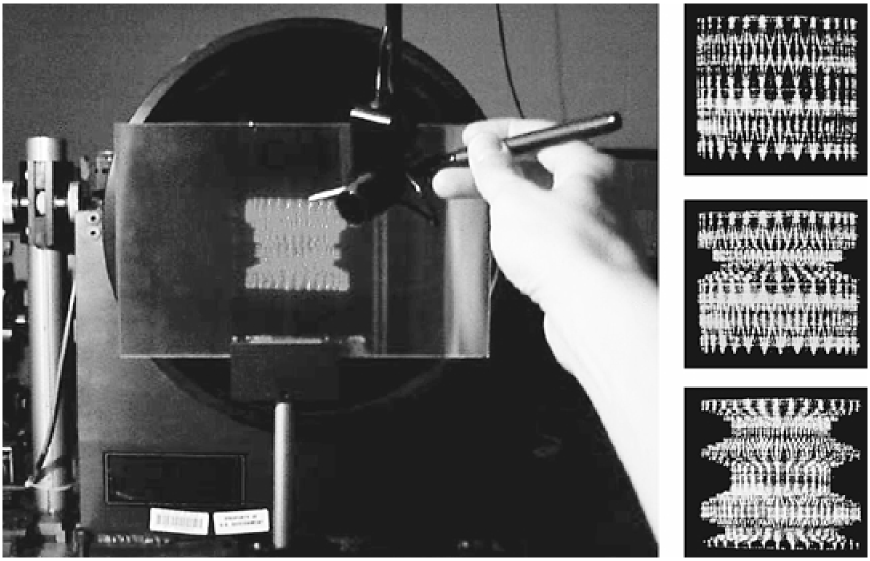


Figure 5.9 Using the holo–haptic lathe.

the following section. We also present a tremendous modal mismatch since our haptic simulation models a spinning stock, but the visual representation does not spin. To represent a spinning holographic image, we must update all the hololines spanned by the image at a reasonable rate; when visual update can be more rapid, of course, the visual and haptic dynamics should match.

Differences between the haptic feedback in our simulation and the feeling of carving on an actual lathe are also important to note. Among them are that

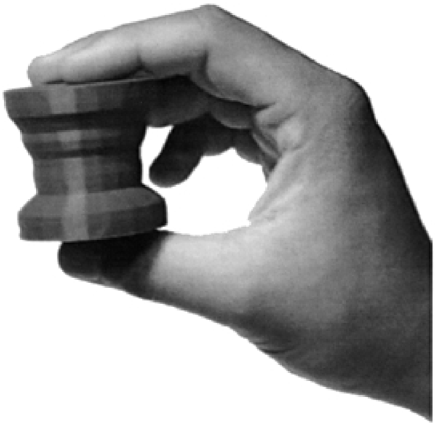


Figure 5.10 Physical prototype of carved stock.

the simple material properties we currently simulate are quite different from those of wood or metal moving against a cutting tool. Additionally, since a "cut" applied at an instantaneous position on the surface of revolution results in a modification that extends around the *entire* circumference of the shape, a person does not experience the feeling of continuously removing material as the stock spins under the stylus. Of course, another obvious departure from the real-world task is the change in orientation of the lathe axis.

5.7 MODALITY DISCREPANCIES AND CUE CONFLICTS

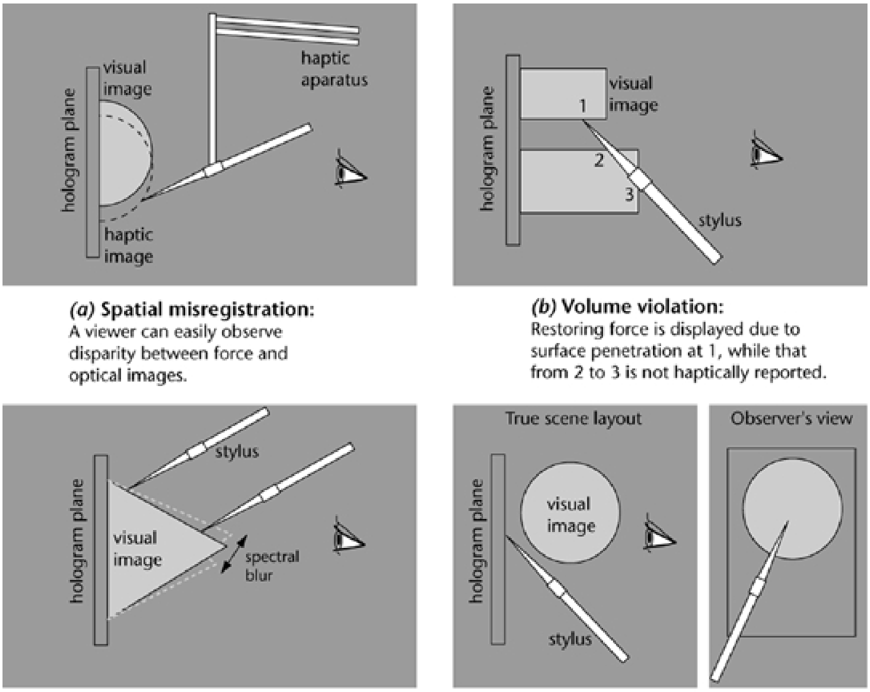
As we readily observe in our everyday interactions, harmonious multisensory stimulation usually gives rise to correct perception of objects and events. The broad body of work on multisensory interaction indicates that some disparity between visual and haptic information can distort the overall percept while still being tolerated. The ability of sensorimotor systems to adapt to discordant sensory input permits us to perform well even in the presence of distortion, so long as sensory feedback is available. This fact is extremely useful in offset visual-haptic workspace configurations, wherein the tracked hand or device position is represented as a graphical element on the visual display and the user never actually visually observes her hand. In such workspace configurations, slight spatial misregistrations or changes in scale between the visual and haptic display can be virtually unnoticeable. Yet too much intermodality disparity can cause the visual and haptic cues to be perceived as arising from entirely separate events and may be quite confusing or annoying.

Tolerances are lower still when visual and haptic workspaces are superimposed. In our coincident holo–haptic workspace, we observed several conflicts between what is seen and what is felt; these intra- and intersensory conflicts are described in turn below.

5.7.1 Spatial Misregistration

When exploring a surface with the Phantom and visually monitoring the device, simultaneous visual and haptic cues to the surface location are available. When we *feel* contact, the visible location of the stylus tip is perceived to be colocated with the haptic surface. During contact, if the holographic surface and the haptic surface are not precisely aligned, the misregistration is strikingly obvious to vision. These conflicting visual cues erode the impression of sensing a single object. Instead, the impression of two separate representations is evident. This condition is shown in Figure 5.11*a*.

If visual and haptic models are perfectly registered, a viewer's eyes are in the correct viewing location, and the stylus tip is touched to a detail on the holographic image, touch, stereopsis, and horizontal motion parallax reinforce the perception that the stylus and the holographic surface detail are spatially



***(a)* Spatial misregistration:** A viewer can easily observe disparity between force and optical images.

***(b)* Volume violation:** Restoring force is displayed due to surface penetration at 1, while that from 2 to 3 is not haptically reported.

***(c)* Material property mismatch:** In this example, the image feels stiff everywhere, but exhibits visible blur that changes with depth.

***(d)* Occlusion violation:** In the *true* spatial layout, the stylus is positioned behind the holographic image of the sphere. In the *apparent* depth ordering, the stylus is seen occluding (in front of) the image.

Figure 5.11 Cue conflicts to depth and layout in holo–haptic systems.

colocated. However, as is the case for all HPO holograms, the lack of vertical parallax causes a slight vertical shift that accompanies head motion and increases with image depth. Thus, spatial misregistration is always potentially present in haptic HPO holograms, but with full-parallax holograms, precisely matched and colocated visual and force representations of a scene can be displayed.

5.7.2 Occlusion Violations

Occlusion is perhaps the most powerful cue to layout in a scene. When we see the image of an object being blocked by the image of another, we understand the occluded object to be farther from our eye than the occluding one. In our holo–haptic system, it is possible to position the haptic apparatus between hologram and image and actually block its reconstruction; in an observer's view of the scene, occlusion relationships contradict other depth cues reporting true scene layout, as shown in Figure 5.11*d*. Even in the presence of correct depth accounting from stereopsis and motion parallax, perception appears to favor the depth ordering reported by occlusion relationships.

5.7.3 Volume Violations

Obviously, holograms present spatial images that cannot by themselves exhibit a restoring force when pushed upon by an object. With no haptic simulation to detect collisions with model surfaces and to display contact forces, the haptic apparatus is free to pass through the holographic image undeterred. Our haptic simulation can prevent a single point on the stylus from penetrating the model, but current device limitations preclude emulation of the kind of multipoint contact that occurs in the physical world.

During each haptic control loop cycle, the simulation checks for a surface collision all along the stylus probe; even if it finds many, it can only compute and display forces for one. If a model surface has been penetrated by the stylus tip, it is assumed the viewer's primary attention is focused there, and forces due to this collision are computed and displayed. However, if not the tip but other points along the probe have penetrated the model, then the collision closest to the tip is used for computation and display.

The situation permits another kind of occlusion violation, which we call a volume violation, to occur, as shown in Figure 5.11*b*. While the stylus tip is seen and felt in contact with some geometry, the stylus may be rotated around its tip and swept through proximal holographic image volume. Parts of the user's hand may also penetrate the visual image while the stylus tip is in contact with the force image. Seeing both physical objects and holographic image coexist in the same physical volume presents a confusing impression of depth and object solidity in the scene.

5.7.4 Visual–Haptic Surface Property Mismatch

Upon observing a visual scene, we form certain expectations about the material properties and surface characteristics of objects we see. Thus, when something appears blurry and soft but its surfaces feel hard and smooth, the effect can be quite startling. Designing a correspondence between visual and haptic material modeling is good policy in multimodal display unless the disparity between these is an element of interest.

An instance of this problem arises with the chromatic blur accompanying broad-spectrum illumination of holograms — not an issue in the current instantiation of holovideo but still worth mentioning. Depth-related blurring throughout the image volume already challenges the impression of image solidity (Fig. 5.11*c*), but adding coincident haptic display causes further difficulty. In this case, an image's visual properties change substantially with depth though its force properties remain the same. Thus in parts of the image volume (typically close to the hologram plane), the multimodal simulation can be very convincing, while the modal outputs seem to break into two distinct and unrelated simulations elsewhere.

5.8 IMPLICATIONS FOR MIXED-REALITY DESIGN

The work described in this chapter offers haptic interaction with holographic images on the tabletop; this marks a long-held goal in the field of holography. Holographic images in the manipulatory space are accompanied by real objects as well (at very least the hand and haptic apparatus). In the resulting mixed-reality setting, visual, haptic, and physical behavior differences between the holographic image and juxtaposed physical objects can be quite striking.

Even if we have done our best to render the holographic images with a solid, 3D appearance, intermodal cue conflicts and many types of discrepancy between spatial images and real objects call attention to the boundary between simulation and reality. Noticeable distinction between real and synthetic objects may not necessarily impact performance in this space, but to the extent that we want to render a *physically believable* scene, we need to consider the underlying issues more carefully.

Based on observations in our laboratory and discussions with users of our systems, we have compiled a preliminary set of guidelines for generating physically believable visual–haptic displays in mixed-reality settings. We suggest that physical believability depends on how well the stimuli representing a simulated object would correspond to stimuli generated by an actual physical instantiation of that object. Rendering methods and display characteristics are obviously important factors. Additionally, all sensory modalities employed in a spatial display should act in concert to model some basic rules that, based on our experience, physical objects usually obey. We group these guidelines into *display*, *rendering*, and *modeling* factors for presenting physically believable multimodal simulations in coincident workspaces:

Display Factors

Simulated and real objects should appear with the same luminance, contrast, spatial resolution, color balance, and clarity.

Visual and force images of *objects* should have “stable” spatial and temporal properties (no perceptible temporal intermittance, spatial drift, or wavering).

No time lag should be detectable between a user’s action and the multimodal response or effect of that action in the workspace.

A viewer’s awareness of display technology should be minimized.

Rendering Factors

Computer-graphic rendering or optical capture geometry should match the system viewing geometry.

Illumination used in simulated scenes should match the position, intensity, spread, and spectral properties of that in the real scene, and simulated shadows and specular reflections should not behave differently.

Optical and haptic material properties, as represented, should be compatible (a surface that *looks* rough shouldn't *feel* soft and spongy).

Modeling Factors

The volumes of simulated objects should not interpenetrate those of real or other simulated objects.

Occlusion, stereopsis, and motion parallax cues should report the same depth relationships.

Convergence and accommodation should provide compatible reports of absolute depth.

Accommodation should be permitted to operate freely throughout the volume of a simulated scene.

The range of fusion and diplopia should be the same for simulated and real scenes.

All multisensory stimuli should appear to arise from a single source and should be in precise spatial register.

Undoubtedly, more issues remain to be added to this list; the factors noted above already prescribe high technological hurdles for visual and haptic display designers.

5.9 CONCLUSION

We set out to demonstrate an experimental combination of display technologies that engage both binocular visual *and* manual sensing. The stylized holo–haptic lathe we chose to implement for this demonstration can be easily manipulated by inexperienced users but elicits the greatest enthusiasm from those familiar with the inherent pleasure in skillfully working materials with their hands. This work has illuminated some of the intra- and intersensory conflicts resident in a coincident visual–haptic workspace and has helped us begin to qualify the requirements for rendering a physically believable simulation in a mixed-reality setting.

Within the field of holography, this work is a simple demonstration of a long-held goal. Not long ago, building a holographic video system that could display interactive moving images itself seemed an intractable problem. However, not only are we currently able to play back prerecorded digital holographic "movies," but we can also propagate primitive changes in underlying scene geometry to the image in near-real time. These changes are achieved by updating the hologram locally, only in regions of change, and not by recomputing the entire fringe pattern. Combining a force model with the spatial–visual image finally allows fingertips to apply a “reality test” to these compelling images and provides the most intimate way of interacting with them.

Our broader agenda is to suggest new ways of developing and working with spatial computational systems as innovative sensing and display technologies become available. In particular, the combination of holographic and haptic technologies with sophisticated computational modeling can form a unique alloy — a kind of digital plastic — whose material properties have programmable look, feel, and behavior. We look forward to the evolution of such systems and the exciting possibilities for their employ in the fields of medicine, entertainment, education, prototyping, and the arts.

REFERENCES

1. A. Leganchuk, S. Zhai, and W. Buxton, “Manual and cognitive benefits of two-handed input: An experimental study,” *Trans. Computer-Human Interaction* **5**(4), 326–359 (1998).
2. P. Servos, M. A. Goodale, and L. S. Jakobson, “The role of binocular vision in prehension: A kinematic analysis,” *Vision Res*., **32(**8), 1513–1521 (1992).
3. J. J. Marotta, A. Kruyer, and M. A. Goodale, “The role of head movements in the control of manual prehension,” *Exp. Brain Res.* **120**, 134–138 (1998).
4. M. Deering, “High resolution virtual reality,” Proceedings SIGGRAPH’92, *Comput. Graphics* **26** 195–202 (1992).
5. R. Playter, “A novel virtual reality surgical trainer with force feedback: surgeon vs medical student performance,” *Proceedings of the Second PHANToM Users Group Workshop*, October 19–22, Dedham, MA, 1997.
6. R. M. Taylor, W. Robinett, V. L. Chi, F. P. Brooks, Jr., W. V. Wright, R. S. Williams, and E. J. Snyder, “The nanomanipulator: A virtual-reality interface for a scanning tunneling microscope,” *Computer Graphics: Proceedings of SIGGRAPH ’93*, 1993.
7. Y. Yokokohji, R. L. Hollis, and T. Kanade, “Vision-based visual/haptic registration for WYSIWYF display,” *International Conference on Intelligent Robots and Systems*, 1996, pp. 1386–1393.
8. P. Wellner, W. Mackay, and R. Gold “Computer augmented environments: Back to the real world,” *CACM* **36**(7) (1993).
9. H. Ishii and B. Ullmer “Tangible bits: Towards seamless interfaces between people, bits and atoms,” in *Proceedings CHI ’97, ACM*, Atlanta, March 1997, pp. 234–241.
10. J. Underkoffler and H. Ishii, “Illuminating light: An optical design tool with a luminous-tangible interface,” in *Proceedings of CHI ’98*, 1998, pp. 542–549.
11. M. R. E. Jones “The haptic hologram,” in *Proceedings of SPIE*, *Fifth International Symposium on Display Holography*, Vol. 2333, 1994, pp. 444–447.
12. W. Plesniak and M. Klug “Tangible holography: Adding synthetic touch to 3D display,” in S. A. Benton (Ed.), *Proceedings of the IS&T/SPIE’s Symposium on Electronic Imaging*, *Practical Holography XI*, 1997.

13. P. St.-Hillaire, M. Lucente, J. D. Sutter, R. Pappu, C. J. Sparrell, and S. Benton, "Scaling up the MIT holographic video system," in *Proceedings of the Fifth International Symposium on Display Holography*, Lake Forest College, July 18–22, SPIE, Bellingham, WA, 1994.
14. J. A. Watlington, M. Lucente, C. J. Sparrell, V. M. Bove, Jr., and I. Tamitani, "A hardware architecture for rapid generation of electro-holographic fringe patterns," *Proceedings of SPIE Practical Holography IX*, SPIE, Bellingham, WA, 1995.
15. J. Underkoffler, "Toward accurate computation of optically reconstructed holograms," S.M. Thesis, Media Arts and Sciences Section, Massachusetts Institute of Technology, 1991.
16. M. W. Halle, "Holographic stereograms as discrete imaging systems," in S. A. Benton (Ed.), *SPIE Proc. Practical Holography VIII*, SPIE, Bellingham, WA, 1994, pp. 73–84.
17. M. Lucente, "Diffraction-specific fringe computation for electro-holography," Ph.D. Thesis, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, 1994.

Three-Dimensional Holographic Imaging. Edited by Chung J. Kuo, Meng Hua Tsai
Copyright © 2002 John Wiley & Sons, Inc.
ISBNs: 0-471-35894-0 (Hardback); 0-471-22454-5 (Electronic)

CHAPTER 6

Preliminary Studies on Compressing Interference Patterns in Electronic Holography

HSUAN T. CHANG

Department of Electrical Engineering
National Yunlin University of Science and Technology
Touliu Yunlin, 64002, Taiwan

6.1 INTRODUCTION

Holography has attracted the attention of many researchers for more than three decades since its invention by Gabor [1]. In conventional holography, the three-dimensional (3D) object information on a holographic film is recorded, and then the developed film is used for the 3D object reconstruction. Due to the exhaustive film exposure and development processes, the conventional hologram is only available for non-real-time applications. To solve this problem, several real-time holographic techniques have been developed recently. The best-known technique is the optical heterodyne scanning technique [2, 3] proposed by Poon et al. The other technique is a micrographic camera for instrumentation purpose [4]. Recently, an electronic holographic apparatus [5] was invented whose electrical output represents the magnitude and phase of coherent light reflected from a 3D object and distributed over the aperture of the apparatus. This apparatus provides a coherent beam that illuminates the object to create a speckle pattern in an aperture bounding an optical sensing arrangement. A reference beam derived from the same source as the illuminating beam illuminates the sensing aperture directly and creates fringes in the speckle pattern. The optical sensing arrangement consists of an 128 × 128 charge injection device (CID) [6] camera with plural optical detectors arranged in a standard rectangular array to sense the magnitude and spatial phase of each speckle on the spatial and time average. The sampled outputs of the CID detectors are processed to isolate the magnitude and phase information representing the complex optical wave front of the hologram from irrelevant terms created by the interference process.

Apparently, other intensity-recording devices such as the charge-coupled device (CCD) [7, 8] can be used in the electronic holographic apparatus. In addition, since the holographic information is converted into electrical signals, the information can then be displayed at a spatial light modulator (SLM) such as the electron-beam-addressed SLM (EBSLM) [9] or the liquid crystal light valve (LCLV) [10] (through a high-resolution cathode ray tube (CRT) [11]) for 3D object reconstruction [12, 13]. The technique above is thus named *electronic holography* [14–20]. If the response time of the intensity-recording device and SLM is fast enough, then the technique can also be used for real-time applications. The holographic film now can be replaced by a high-resolution CCD, which can record the optical information and then convert it to digital information.

Computers can assist us to handle the whole process of electronic holography. However, the original analog holographic information should be converted into digital format. Therefore, the interference pattern generated in CCDs will be digitized and thus can be processed by the computer. Conventional holograms own a large scale of size, for example, 5×10 cm. Although a hologram can be replaced by a smaller size CCD, the digitized holographic information becomes very huge because of its high resolution. For example, a CCD with 1024×1024 pixels and 256 gray levels will generate a 32-Mbyte data file. This file size is much larger than common digital still images. Therefore, it should be compressed for ease of storage and transmission. A high compression ratio is expected to reduce the huge data amount of the holographic information. As a result, however, compression distortion must be introduced, and hence we must examine the performance based on different compression methods. In this chapter, we first investigate the characteristics of the interference pattern. Then we propose a novel method to enhance the light efficiency of the holography. The sampling and quantization effects on the digitized holographic information are briefly introduced. A nonlinear quantization model is used to reduce the quantization noise. Before proposing interference pattern compression techniques, the digital hologram is downsized and subsampled to examine the reconstruction results. Finally, a Joint Picture Expert Group (JPEG)–based technique [21] is used to compress the interference pattern that has been transferred to a gray-scale image.

6.2 CHARACTERISTIC OF INTERFERENCE PATTERN

The architecture of conventional holography is shown in Figure 6.1. Conventional holography utilizes an object wave and a reference wave to collect and record the interference pattern on the hologram. Let us denote the wave distribution of the reference wave and the object wave at the hologram plane

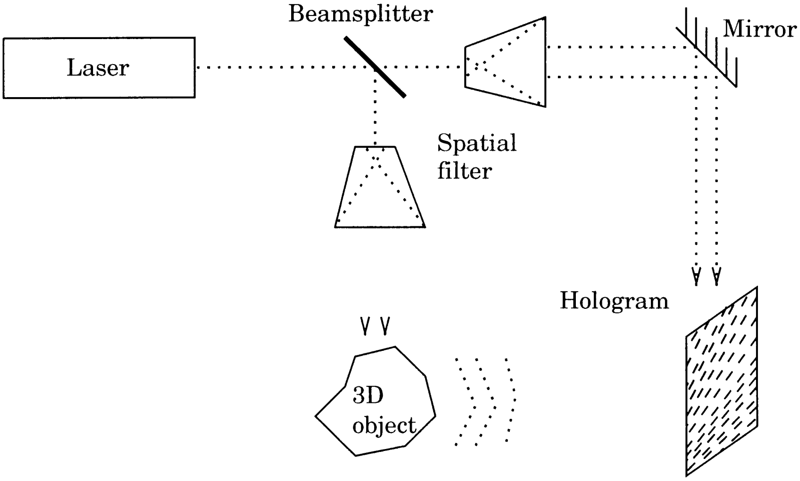


Figure 6.1 Architecture of conventional holography.

by $Ae^{-j2\pi\alpha y}$ and $a(x, y)e^{j\phi(x,y)}$, respectively. If the recording plane is placed at the position (x_0, y_0, z_0) from the object, the amplitude distribution $a(x, y)$ and the phase distribution $\phi(x, y)$ on the recording plane become

$$a(x, y) = \frac{a_0}{r} \tag{6.1}$$

$$\phi(x, y) = kr \tag{6.2}$$

where a_0 is the amplitude of the point source,

$$r = \sqrt{(x - x_0)^2 + (y - y_0)^2 + (z - z_0)^2},$$

and $k = 2\pi/\lambda$ is the wave number. Here, (x, y) denotes the spatial position at the plane of interest, and the spatial frequency of the reference plane wave is given by $\alpha = \sin\theta/\lambda$, where λ is the wavelength of light and α is the incident angle in the $-y$ direction normal to the recording plane. We thus have the input light intensity on the hologram as

$$\begin{aligned} I(x, y) &= |Ae^{-j2\pi\alpha y} + a(x, y)e^{j\phi(x,y)}|^2 \\ &= A^2 + a^2(x, y) + 2Aa(x, y)\cos[\phi(x, y) + 2\pi\alpha y] \end{aligned} \tag{6.3}$$

This input light intensity is called the *interference pattern* of the object and reference waves. It is always positive since it is an intensity signal. As shown in Eq. 6.3, the phase information $\phi(x, y)$ of the object wave is preserved in the cosine-modulated term $2Aa(x, y)\cos[\phi(x, y) + 2\pi\alpha y]$.

Now we analyze the intensity signal $I(x, y)$ on the recording plane. It is found that A^2 is a bias term and the original light distribution of the object is squared. Thus the variance of the object wave is enlarged. On the other hand, in the third term of Eq. 6.3, the original light distribution of the object is multiplied by a cosine term in which the phase is determined by the phase information of both the original object and the reference wave.

We can Fourier transform Eq. 6.3 and calculate the spectrum of the signal $I(x, y)$. With the aid of Fourier spectrum analysis, the spectrum of the signal on the recording plane is wider than the spectrum of the signal in which only the object wave is considered. Therefore, in the spatial domain, the interference pattern of the signal $I(x, y)$ incident on the recording plane is much more complicated than the original one of the object wave. This is also why we usually cannot obtain any information from the hologram since the original signal has been modulated to high frequencies.

We can expect that the interference pattern has two main characteristics: First, the contrast of the signal is high due to the generated squared term $a^2(x, y)$. Second, the original distribution of the object signal is cosine modulated with the spatial frequency $\phi(x, y) + 2\pi\alpha y$. Usually the spatial frequency of the reference plane wave is high since the wavelength λ is on the order of 10^{-10} m such that $\sin\theta/\lambda$ leads to a large value. Since A is a constant, the original distribution $a(x, y)$ is cosine modulated by a high angular frequency, and thus the interference pattern is shown to be with more high-frequency contents.

In our experiments, we use a point source as the object and the width of the point source is infinitely small. The wavelength of the light source is assumed to be 632 nm. The Fresnel diffraction theory [1] is used to predict the diffraction pattern in the holography. The point source is placed 50 cm in front of the CCD, and the plane reference wave is tilted at the angle $\frac{1}{4}\pi$ in the $-y$ direction and incident on the CCD. Given the amplitude of the point source and the reference wave, that is, $a(x, y)$ and A, the light intensity on the CCD is calculated using Eq. 6.3. The pixels in the CCDs are in the standard rectangular array. The spacing between two neighboring pixels is assumed to be zero.

In the recording stage, each pixel in the CCD acts linearly, and thus no nonlinear effects [22] occur while converting the light intensity to a digital signal. The CCD sensor has pixels of 10 μm pitch. To digitally process and transmit the holographic signals, the analog-to-digital (A/D) converter is used to quantize the received intensity signals. The resolution (the number of available bits) of the A/D converter affects the performance of the quantizer. The higher the resolution of the converter, the smaller the quantization noise. Here we choose 8-bit resolution (256 gray levels) in the A/D converter.

6.3 ELECTRONIC HOLOGRAPHY

The reconstruction of the 3D image is obtained by illuminating the same reference wave on the hologram and can be seen by the human eye. As shown in Eq. 6.3, only the cosine-modulated term contributes to the original 3D information. The term $a^2(x, y)$ is useless in the reconstruction step. Therefore, if we can eliminate this term, the light (diffraction) efficiency will be improved.

6.3.1 A Novel Architecture

To eliminate the $a^2(x, y)$ term, we here propose a novel scheme with the help of a digital computer. The architecture of the proposed digital holography is shown in Figure 6.2. In this figure, we employ a 50/50 beamsplitter in front of the object. Two CCDs are placed in the same length of the optical path after the beamsplitter. Three black shields are placed to isolate the light from other undesirable light sources. The reference wave is incident only to one of the CCDs. Let CCD 1 receive the summed light intensity from both the object and reference wave. In contrast, CCD 2 receives the light intensity reflected from the object only. The light intensity on CCD 1 now becomes

$$I_1(x, y) = A^2 + \tfrac{1}{4}a^2(x, y) + Aa(x, y)\cos[\phi(x, y) + 2\pi\alpha y] \tag{6.4}$$

and the light intensity on CCD 2 becomes $I_2(x, y) = \frac{1}{4}a^2(x, y)$. Both intensity signals are transmitted to the computer and proceed with the subtraction operation. Assume the amplitude A of the plane reference wave can be known in advance. We can eliminate the tern A^2 in the computer and then obtain the light intensity signal

$$I_3(x, y) = Aa(x, y)\cos[\phi(x, y) + 2\pi\alpha y] \tag{6.5}$$

in which only the cosine-modulated term is left. In the reconstruction stage, the useless part corresponding to $I_2(x, y)$ is eliminated, and thus a higher holographic diffraction efficiency can be achieved.

The light intensity captured by the CCD is then converted into digital format. That is, the intensity signals are sampled and quantized. Before the digitizing process, we make the first assumption that the CCD recording is linear while converting light intensity to the electrical signal. That is, the nonlinear effects on the recording device are neglected. However, the low spatial resolution of the CCD is the major drawback of digital holography. The maximum angle $\theta_{\max}$ between the reference and the object wave is dependent on the maximum spatial frequency that can be resolved by the recording medium [18, 23, 24]. Conventional holographic recording media have resolution as high as 5000 lines/mm. But CCD cameras have resolutions of 100 lines/mm. Therefore, the maximum angle between two interference waves is

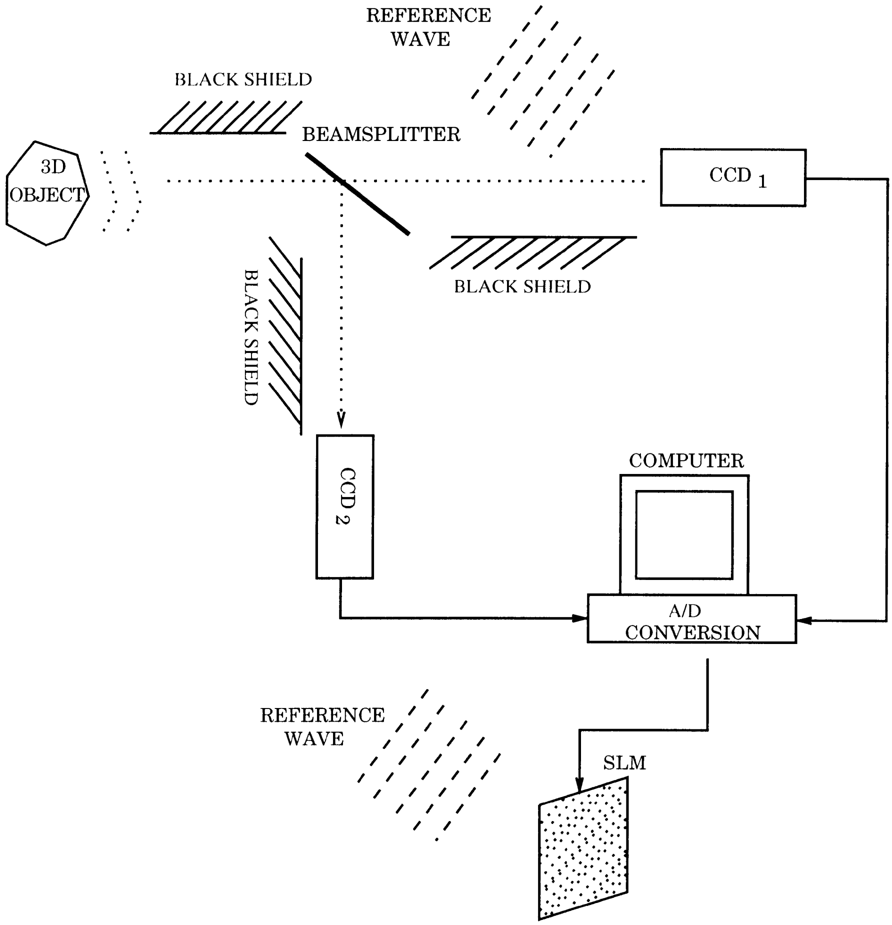


Figure 6.2 Architecture of proposed high-efficiency holography that can eliminate $a^2(x, y)$ term in conventional holography.

limited to only a few degrees. In our second assumption, we suppose that the resolutions of the CCDs will be high enough such that the maximum degree $\theta_{\max}$ is not limited to some degrees.

Figure 6.3*a* shows the light distribution at the CCD for the point source only. The intensity distribution of light, that is, interference pattern, which is obtained from Eq. 6.3, corresponding to this point source and the reference wave is shown in Figure 6.3*b*. As shown in this figure, more high-frequency contents are generated. If the light intensity corresponding to Eq. 6.5 in which the object is eliminated, with the proposed scheme the interference pattern becomes that shown in Figure 6.3*c*. It looks very similar

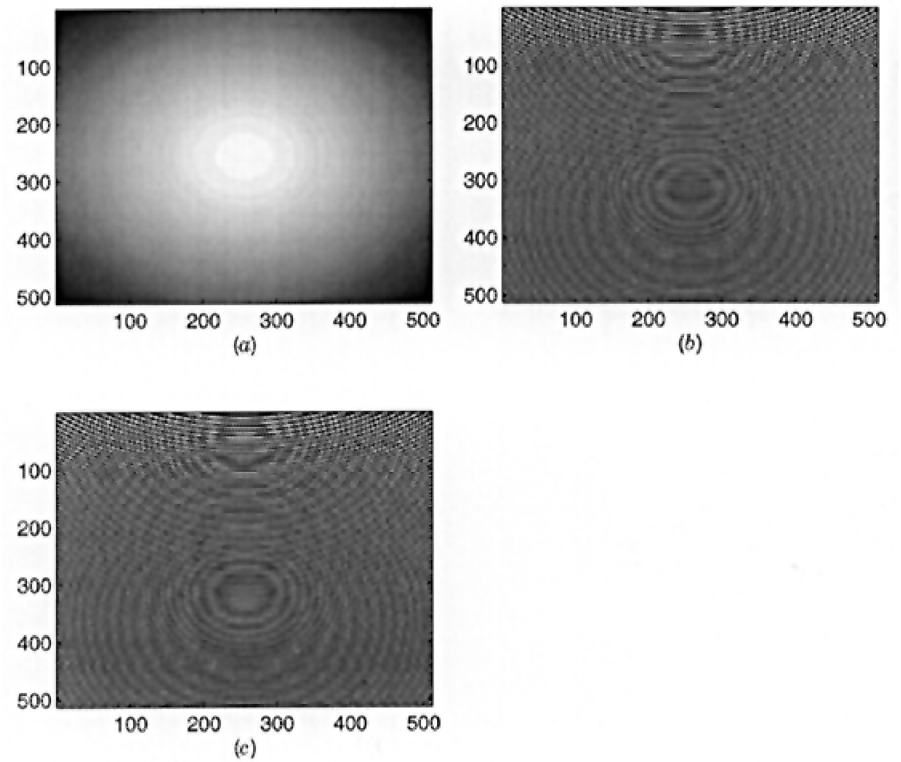


Figure 6.3 (*a*) Object intensity distribution and two holograms obtained from (*b*) conventional and (*c*) proposed digital holography.

to the pattern shown in Figure 6.3*b* since the $a^2(x, y)$ term is very small while compared with the other terms in Eq. 6.4. Note that the interference patterns and the power spectrum are represented by the monotone images with 256 gray levels.

The Fourier spectrum corresponding to Figure 6.3 is obtained by taking the absolute value of the *fast Fourier transform* (FFT) on the interference patterns and is shown in Figures 6.4*a–c*. As shown in Figure 6.4*a*, the recorded light intensity of the point source is almost band limited in low frequency. When the reference wave interferes, the cosine modulation shifts the baseband signal to higher band. Figure 6.4*b* proves this result. If we discard the $a^2(x, y)$ part, the Fourier spectrum in Figure 6.4*c* shows that the corresponding frequency components also disappear. The Fourier spectrum left only corresponds to the intensity signal shown in Eq. 6.5. Apparently, this spectrum is almost the same as that shown in Figure 6.4*b*.

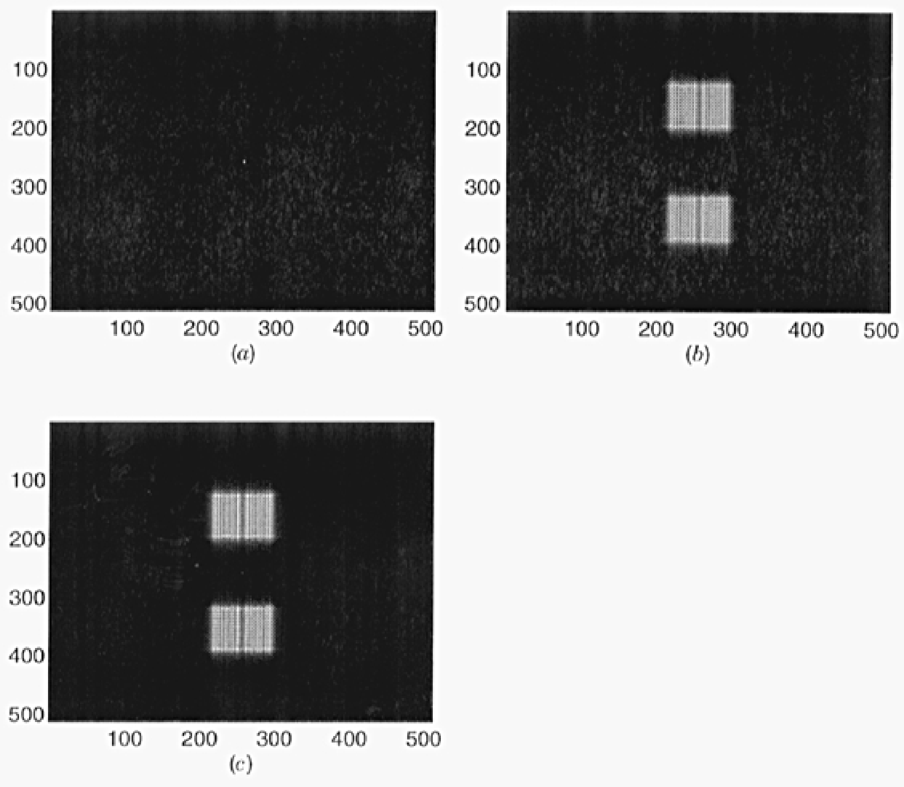


Figure 6.4 Fourier spectrum of interference patterns for (*a*) object wave only, (*b*) conventional holography, and (*c*) proposed holographic architecture.

6.4 SAMPLING AND QUANTIZATION

Sampling and quantization problems in digital holography have been discussed elsewhere in [25, 26]. It is noted that the sample spacing used in holographic data processing is usually selected to be a half wavelength since it can achieve the Nyquist rate. However, to avoid the alias error [27] in digital holography, it is of course not practical to take an infinite number of samples at the Nyquist rate or higher. Therefore an optimal spacing can be determined by minimizing the correlation among the image components. When we use only a given fixed number of data samples for the wave field, there is a trade-off between the sampling frequency and the aperture size for imaging.

From the mathematical derivation shown in the previous section, we can obtain the final light intensity signal $I_f(x, y) = Aa(x, y)\cos[\phi(x, y) + 2\pi\alpha y]$ in the computer. This representation is an analog form. In numerical evaluations, the recording plane is divided into discrete elements. Suppose that the CCD targets have $N \times M$ pixels with pixel dimensions $\Delta x \times \Delta y$. Then the discrete representation of the signal becomes

$$\begin{aligned} I_d[n,m] &= I_f(n\,\Delta x,\ m\,\Delta y) \\ &= Aa(n\,\Delta x,\ m\,\Delta y)\cos[\phi(n\,\Delta x,\ m\,\Delta y) + 2\pi\alpha m\,\Delta y] \end{aligned} \tag{6.6}$$

where n and m are integers and $1 \leqslant n \leqslant N$, $1 \leqslant m \leqslant M$.

6.4.1 Uniform Quantization

In the quantization process, we simply measure the maximum and minimum discrete signal $I_d[n,m]$ to obtain its dynamic range. The uniform quantization step δ for L gray levels can be determined by

$$\delta = \frac{\max(I_d[n,m]) - \min(I_d[n,m])}{L-1} \tag{6.7}$$

for all n and m. Consequently, this quantization process unavoidably introduces the distortion. It has been shown that the quantization level L should not be less than 256 (i.e., 8-bit resolution) such that a high signal-to-noise ratio (SNR) can be obtained [14].

6.4.2 Nonuniform Quantization

Since the contrast of the images shown in Figure 6.3 is high, we also investigate their corresponding histograms. Figure 6.5 shows the histogram of the inter-

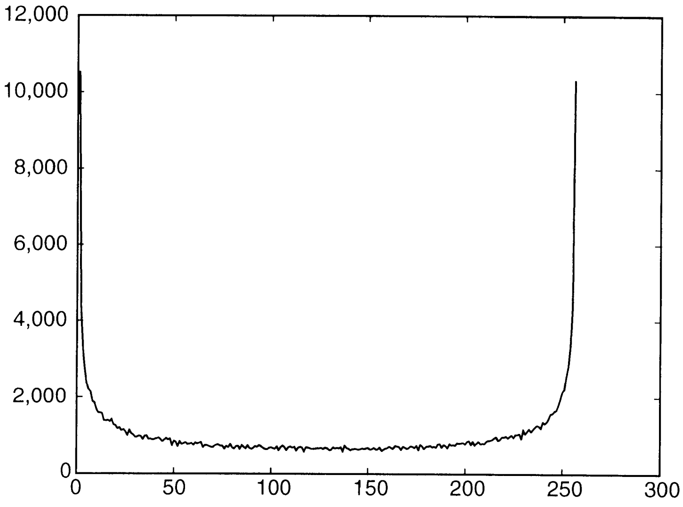


Figure 6.5 Histogram of interference pattern for proposed holographic architecture.

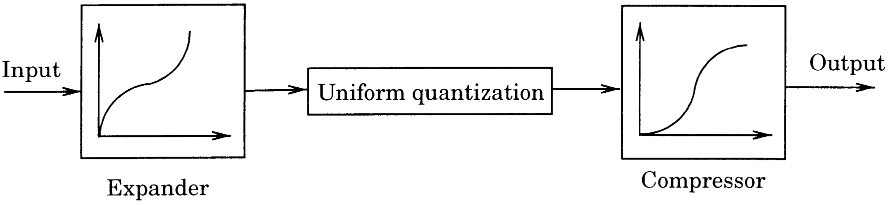


Figure 6.6 Expressor model of nonuniform quantization.

ference pattern in Figure 6.3*c*. Most pixels are concentrated in the very low and very high gray levels. This is very different from natural images. We might be able to equalize the histogram distribution and thus the conventional compression techniques are easier to be applied on the interference patterns. As shown in Figure 6.5, most pixel values are distributed in the gray levels regions 0–10 and 250–255. The quantization error can be further reduced if we apply a nonuniform quantization on the original analog signals. Performing the nonuniform quantization on the original signal can obtain a more uniform distribution of the quantized signals. This can be achieved by using a method similar to the compander model [28]. We call this the *expressor* model, and it is composed of the expander first and then the compressor. Figure 6.6 shows the expressor model for performing uniform quantization on a nonuniform distribution. Since the most important compander is the logarithmic compander, the characteristic of the proposed expressor is also designed by approximating the logarithmic curve for positive values. However, this is somewhat different from the original design.

As shown in Figure 6.6, the positions of the expander and compressor are reversed compared with the compander architecture. The curve has to be shifted such that it can cover and operate for pixels in both the high and low gray levels. Here we employ the μ-law characteristic on the expressor model. The μ-law used in the compressor is defined by

$$|v_2| = \frac{\log(1 + \mu|v_1|)}{\log(1 + \mu)} \tag{6.8}$$

where v_1 and v_2 are normalized input and output values. Therefore the inversed μ-law used in the *expander* becomes

$$|v_2| = \frac{(1 + \mu)^{|v_1|} - 1}{\mu} \tag{6.9}$$

Figure 6.7 shows the input/output relationships based on different μ values (1, 10, 50, 100). The received signals are normalized in the range between -1 and $+1$ such that Eqs. 6.8 and 6.9 can be applied on the normalized signals directly. Once this nonuniform quantization process is complete, the quantized signals are then mapped into the 256 gray levels 0–255 for display purpose.

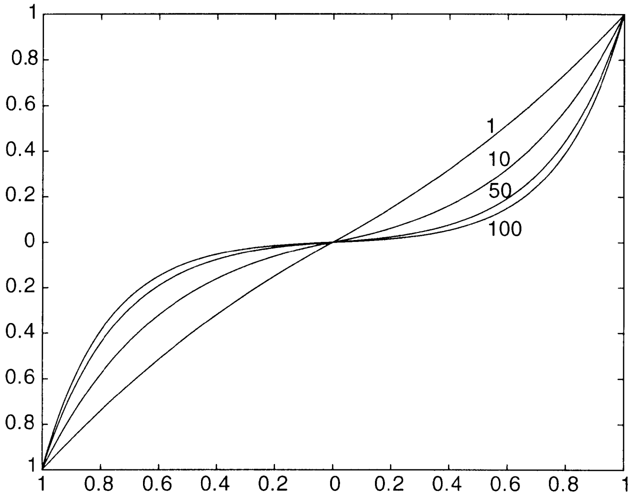


Figure 6.7 Expander law: μ-law.

Figure 6.8 shows the histograms obtained from the nonuniform quantization based on the expressor model with different μ values. The pixel gray levels concentrate in the median range when the parameter μ increases. Compared with the histogram shown in Figure 6.5, the parameter μ should be set around 1 so that a more uniform distribution can be obtained. Therefore better reconstruction of the 3D image can be expected. In our simulation, μ

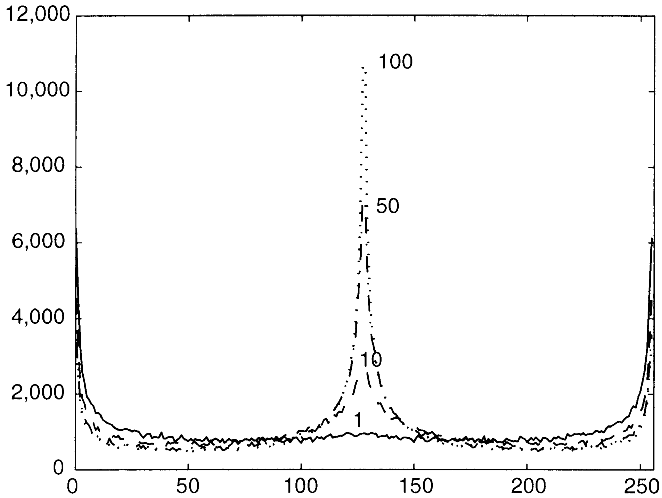


Figure 6.8 Histograms obtained from nonuniform quantization based on expressor model with different μ values.

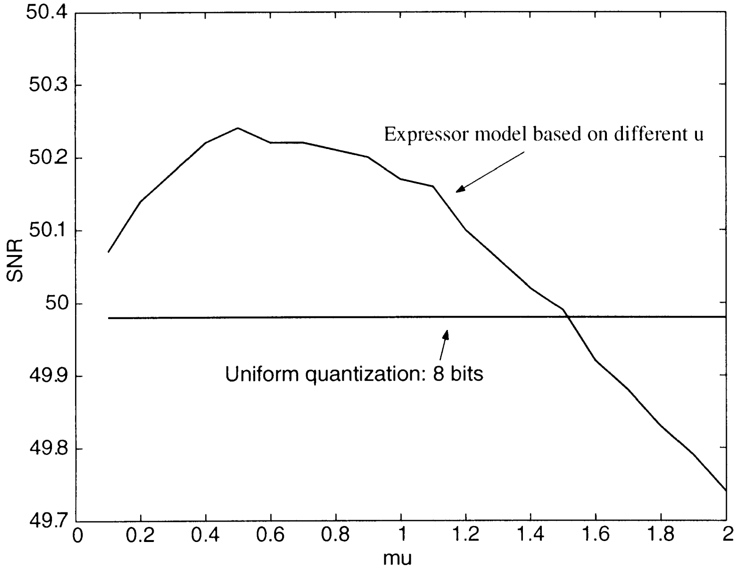


Figure 6.9 The SNR comparison between uniform (8 bits) and nonuniform quantization at different μ values.

ranges from 0.1 to 2 and increases by 0.1. As shown in Figure 6.9, the SNR is improved when μ is less than 1.5. However, the improvement is less than 1 dB and it is not significant. The reason is that the numbers of pixels in very high and very low gray levels are only a small part of all the pixels. Therefore, we can utilize the uniform quantization for original interference patterns to avoid complicated calculation in the expressor model.

6.5 COMPRESSION OF INTERFERENCE PATTERN

Now, we have obtained the digitized interference patterns of the electronic holography. It can be displayed in a visual form like a gray-scale image. However, the data amount is huge when the resolution required for the CCD is as high as, for example, 2048 × 2048 [16, 24]. For efficient storage and transmission through the Internet, a compression scheme should be employed to reduce the data amount so that the reconstructed 3D information can be well preserved.

In an image coding framework, many compression techniques have been proposed already. For example, the JPEG-based technique, vector quantization, wavelet transform, and fractal-based schemes are well known and have shown good performance. However, the interference pattern in digital holography is very different from the natural images or artifact objects and shapes.

Moreover, the final target is to reconstruct a 3D image. The coded interference pattern must be transmitted to a display such as an SLM and illuminated by the original reference wave. Therefore, we should calculate the SNR for the reconstruction results, not for the reconstructed interference pattern.

In this section, we will investigate some possible ways to compress the interference pattern. For example, we can downsize or subsample the original interference pattern to achieve the compression purpose. It has been shown that only a small part of the hologram stores some sort of the whole information. We will make computer simulations based on different hologram sizes. We will subsample the digitized interference pattern based on different sampling rates. Finally, the JPEG-based technique is used to compress the visualized interference pattern. The SNRs of three cases will also be determined to evaluate their performance.

6.5.1 Downsizing

A conventional hologram can reconstruct the original wave front even when we only have a part of entire hologram. Therefore, an intuitive method to reduce the amount of data contained in the interference pattern is to select part of the information for reconstructing the 3D image. The hologram size used in our original simulation is 512×512 with pitch 10^{-5} m. Here four different hologram sizes are used, 256×256, 128×128, 64×64, and 32×32, with the same pitch as in our reconstructing stage. The centers of the four smaller holograms are placed in the same position as the original one. The corresponding light intensity distributions of the reconstructed point source are shown in Figures 6.10*a–d*. As shown in the figure, the reconstruction peak is wider when the size of the hologram is smaller. That is, a larger hologram owns a higher resolution in the recontructed 3D image and vice versa.

We also investigate the SNRs for the above cases. The SNR is defined as

$$\mathrm{SNR} = 10\log_{10}\frac{\text{signal power}}{\text{noise power}} \quad (\mathrm{dB}) \tag{6.10}$$

Figure 6.11 shows the SNRs based on the different sizes of the holograms in Figure 6.10. These SNRs are all less than 0 dB, which means that the SNR decreases significantly when the compression ratio (CR) is more than 4. However, the point object is an extreme case since the power of the original signal is very small.

6.5.2 Subsampling

Another method to reduce the amount of holographic information is to subsample the original interference pattern. In our experiments, we subsampled the original 512×512 hologram based on different sampling periods (2, 4, 8,

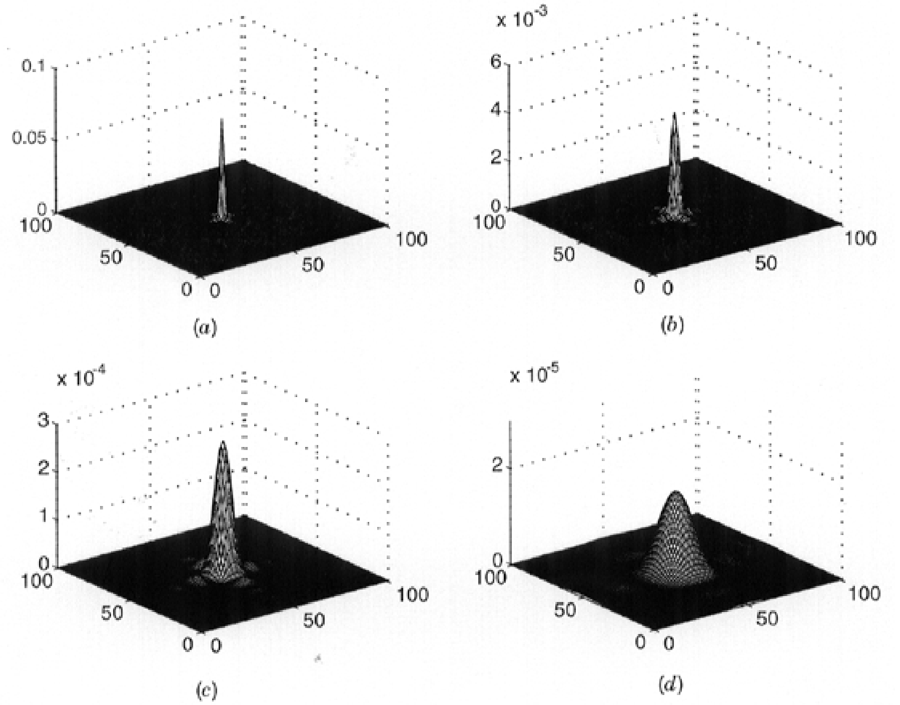


Figure 6.10 Constructed point source at different sizes of digital hologram: (*a*) 256 × 256, (*b*) 128 × 128, (*c*) 64 × 64, and (*d*) 32 × 32.

and 16 pixels), to obtain different subsampled holograms (256 × 256, 128 × 128, 64 × 64, and 32 × 32). Figures 6.12*a–d* show the reconstructed point source based on different subsampled periods. As shown in the figure, the intensity of the reconstruction peak decreases significantly as the sampling period increases. This is resonable since fewer sample points contribute to the signal reconstruction. In contrast, the reconstructed peaks are as wide as the original peak. However, the aliasing effects appear when the sampling period is 16 pixels.

The SNRs corresponding to different sampling periods are given in Figure 6.13. These SNR results are unacceptable and very similar to Figure 6.11. Therefore, both downsizing and subsampling are not efficient methods for compressing the digitized interference pattern in a hologram.

6.5.3 JPEG-Based Coding Technique

The JPEG-based method can also be applied to the digitized interference pattern. The interference pattern is first quantized into 256 gray levels such that it can be displayed as a gray-scale image. The maximum and minimum original

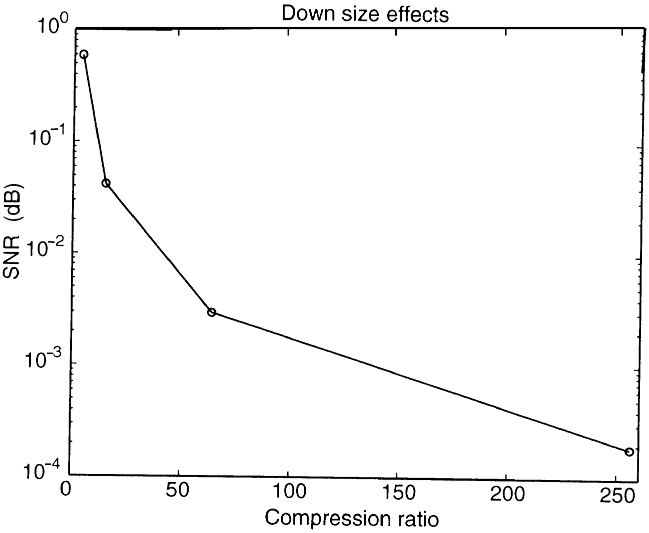


Figure 6.11 Corresponding SNR (dB) at different sizes of digital hologram: (*a*) 256 × 256, CR = 4; (*b*) 128 × 128, CR = 16; (*c*) 64 × 64, CR = 64; and (*d*) 32 × 32, CR = 256.

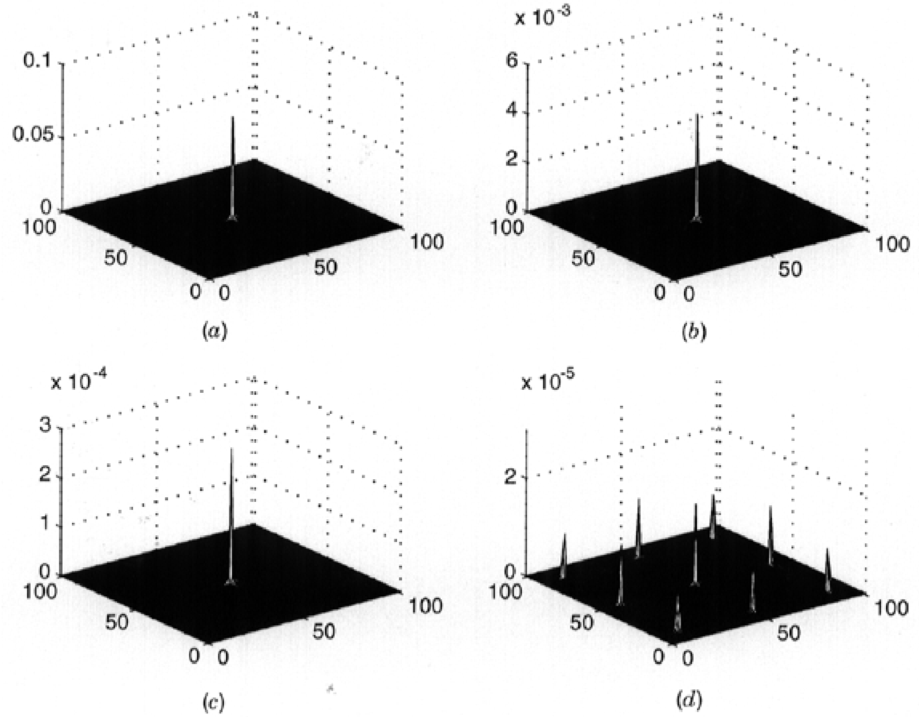


Figure 6.12 Reconstructed point source at different subsampled sizes on original digital hologram of size 512 × 512: (*a*) 256 × 256; (*b*) 128 × 128; (*c*) 64 × 64, and (*d*) 32 × 32.

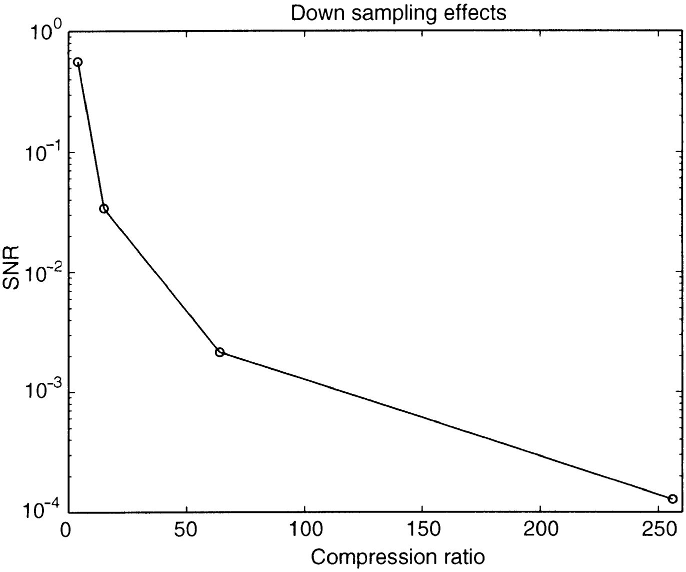


Figure 6.13 Corresponding SNR (dB) at different sampling rates on original digital hologram of size 512 × 512: (*a*) 256 × 256, CR = 4; (*b*) 128 × 128, CR = 16; (*c*) 64 × 64, CR = 64; and (*d*) 32 × 32, CR = 256.

quantities are recorded since both are required in the reconstruction stage. Then we compress this image using a JPEG-based technique. During the reconstruction stage, the compressed image is decompressed and then is combined with the stored maximum and minimum to restore the interference pattern. The restored interference pattern is sent to an SLM or other display device and then is illuminated by the reference wave so that a distorted wave front of the original 3D object is obtained. Figure 6.14 shows the SNR results under different compression ratios using the JPEG technique. As shown in the figure, the SNR values vary considerably under various compression ratios. Since some important parts of the interference pattern are hidden in the high-frequency components, the high-frequency components attenuated by the JPEG-based technique may result in severe degradation of the original 3D information. It is obvious that the JPEG-based coding technique is not well suited for compressing the digitized interference pattern.

6.6 SUMMARY

In this chapter, we propose a novel architecture for electronic holography so that the diffraction efficiency of the reconstructed 3D image can be increased. We study the characteristics of the digitized interference pattern at both the spatial and frequency domains. An expressor model is proposed to quantize

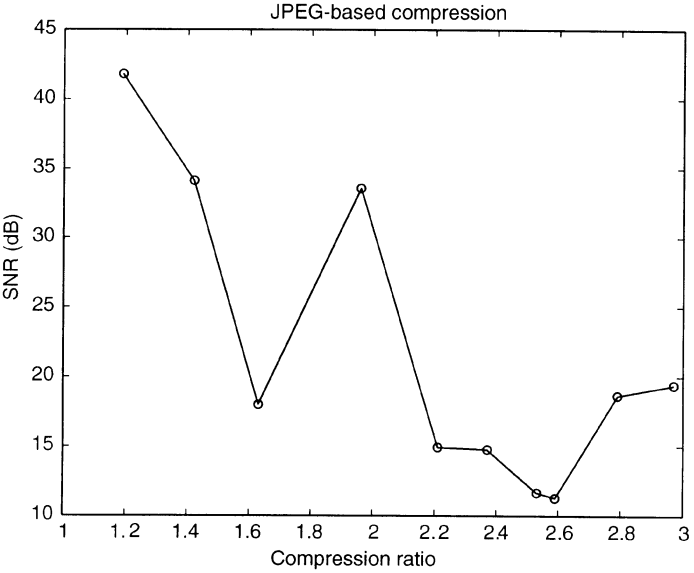


Figure 6.14 SNR (dB) versus compression ratio based on JPEG methods for digital hologram.

the original interference pattern such that the distortion introduced by quantization is reduced. Three methods — downsizing, subsampling, and JPEG-based techniques — are used to compress the information of the digitized interference pattern. However, from simulation results, we found that the SNR performance in the reconstruction stage is not acceptable. Therefore in future work we will propose new techniques that are specialized for compressing the interference pattern such that both the compression ratio and the fidelity of the reconstructed 3D image are high enough for transmission and storage purposes. Finally, some real objects will be used to generate the real digital hologram and to test the performance of the proposed techniques for compressing the digitized interference pattern in electronic holography.

ACKNOWLEDGMENT

Support of the National Science Council, Taiwan, under grant NSC 89-2213-E-324-028 is gratefully acknowledged.

REFERENCES

1. J. W. Goodman, *Introduction to Fourier Optics*, 2nd ed., McGraw-Hill, New York, 1993.

2. B. D. Duncan, T. C. Poon, M. H. Wu, K. Shinoda, and Y. Suzuki, “Real-time reconstruction of scanned optical holograms using an electron beam addressed spatial light modulator,” *J. Modern Opt.* **39**, 63–80 (1992).

3. T. C. Poon, B. W. Schilling, M. H. Wu, K. Shinoda, and Y. Suzuki, “Real-time two-dimensional holographic imaging by using an electron-beam-addressed spatial light modulator,” *Opt. Lett.* **18**, 63–65 (1993).
4. S. B. Gurevich, N. V. Dunaev, V. B. Konstantinov, S. A. Pisarevskaya, V. F. Relin, and D. F. Chernykh, “Compact holographic device for testing of physico-chemical processes under microgravity conditions,” *Proc. AIAA/IKI Microgravity Sci. Symp.* 351–355 (1991).
5. J. Chovan, W. A. Penn, J. J. Tiemann, and W. E. Engeler, “Electronic holographic apparatus,” U.S. Patent 4,974,920 (1990).
6. H. Burke and G. J. Michon, “Charge-injection imaging: Operation techniques and performances characteristics,” *IEEE Trans. Electron Dev.*, **23**, 189–195 (1976).
7. I. Furukawa, K. Kashiwabuchi, and S. Ono, “Super high definition image digitizing system,” *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.* **3**, 529–532 (1992).
8. CCD Image Sensors and Cameras, Dalsa Inc., Canada.
9. EBSLM X3636, Hamamatsu Photonics K. K. Central Research Laboratory, Japan (1993).
10. FLC-SLM X4601, Hamamatsu Photonics K. K. Central Research Laboratory, Japan.
11. Miniature CRT, Miyota Co., Japan.
12. L. Onural, G. Bogdagi, and A. Atalar, “New high-resolution display device for holographic three-dimensional video: Principles and simulations,” *Opt. Eng.* **33**, 835–844 (1994).
13. P. S. Hilaire, S. A. Benton, and M. Lucente, “Synthetic aperture holography: A novel approach to three-dimensional displays,” *J. Opt. Soc. Am., Part A* **9**(11), 1969–1977 (1992).
14. C. J. Kuo and H. T. Chang, “Resolution studies for electronic holography,” *Opt. Eng.* **34**(5), 1352–1357 (1995).
15. C. J. Kuo, “Electronic holography,” *Opt. Eng.* **35**(6), 1528 (1996).
16. E. Marquardt and J. Richter, “Digital image holography,” *Opt. Eng.* **37**(5), 1514–1519 (1998).
17. H. Rabal, J. Pomarico, and R. Arizaga, “Light-in-flight digital holography display,” *Appl. Opt.* **33**(20), 4358–4360 (1994).
18. U. Schnars and W. Juptner, “Digital recording and reconstruction of holograms in hologram interferometry and shearography,” *Appl. Opt.* **33**(20), 4373–4377 (1994).
19. H. Chen, M. Shih, E. Leith, J. Lopez, D. Dilworth, and P. C. Sun, “Electronic holographic imaging through living human tissue,” *Appl. Opt.* **33**(17), 3630–3632 (1994).
20. M. Lucente, “Holographic bandwidth compression using spatial subsampling,” *Opt. Eng.* **35**(6), 1529–1537 (1996).
21. W. B. Pennebaker and J. L. Mitchell, “JPEG: Still image data compression standard,” Van Nostrand Reinhold, New York, 1993.
22. C. J. Kuo and H. T. Chang, “Performance evaluation of noisy hologram,” *Proc. SPIE* **2000**, 270–278 (1993).
23. U. Schnars and W. Juptner, “Direct recording of holograms by a CCD target and numerical reconstruction,” *Appl. Opt.* **33**(2), 179–181 (1994).

24. U. Schnars, T. M. Kreis, and W. P. O. Juptner, “Digital recording and numerical reconstruction of holograms: Reduction of the spatial frequency spectrum,” *Opt. Eng.* **35**(4), 977–982 (1996).
25. H. Lee and G. Wade, “Sampling in digital holographic reconstruction,” *J. Acoust. Soc. Am.* **75**(4), 1291–1293 (1984).
26. N. C. Gallagher, Jr., “Optimum quantization in digital holography,” *Appl. Opt.* **17**(1), 109–115 (1978).
27. J. P. Allebach, N. C. Gallagher, and B. Liu, “Aliasing error in digital holography,” *Appl. Opt.* **15**(9), 2183–2188 (1976).
28. S. Haykin, *An Introduction to Analog and Digital Communications*, John Wiley & Sons, New York, 1989, pp. 410–412.

Three-Dimensional Holographic Imaging. Edited by Chung J. Kuo, Meng Hua Tsai
Copyright © 2002 John Wiley & Sons, Inc.
ISBNs: 0-471-35894-0 (Hardback); 0-471-22454-5 (Electronic)

CHAPTER 7

Holographic Laser Radar

BRADLEY W. SCHILLING
U.S. Army CECOM RDEC
Night Vision and Electronic Sensors Directorate
Fort Belvoir, VA 22060

7.1 INTRODUCTION

An innovative holographic recording technique called optical scanning holography (OSH) has been addressed in detail in another chapter of this book. A derivative of the OSH system is described here and applied to the laser radar problem. The technique is similar in operation to a standard laser radar system in which the image is built up pixel by pixel as the laser pattern is scanned over the object. The main difference between a standard laser radar and the holographic laser radar is in the scanning field. A typical laser radar system employs a spot scan while a holographic laser radar is based on scanning with an optically heterodyned Fresnel zone pattern (FZP). Since the system scans with a structured laser beam (the FZP), three-dimensional (3D) information is encoded in the image as an interference pattern, not by measuring time of flight, as in a standard laser radar. The resulting interference pattern is a hologram of the object or scene.

This technique is noteworthy for several reasons, stemming primarily from the scanning and electronic aspects of the system. For instance, due to the electronic nature of the technique, the photographic processes associated with traditional holographic recording can be avoided. In fact, real-time holography has been demonstrated with a similar FZP scanning technique [1] Since a hologram recorded using electronic holography is a digital image, it is suitable for storage on a computer, convenient transmission using standard communication devices, and other digital manipulations, such as digital compression or processing. Another advantage, inherent in holography in general, is that in a hologram 3D information is stored in a 2D array, in this case a 2D digital image. The electronic nature of the system offers another advantage over traditional holographic recording systems in the area of spectral flexibility.

Theoretically, it is possible to record holograms by this technique in any spectral band where a coherent (laser) source and detector combination exists. Therefore, the technique is particularly well suited to multicolor holography and offers the possibility of holographic recording at infrared wavelengths.

The scanning aspect of the technique offers another set of advantages over traditional holography by relaxing the size constraints of the object or scene to be recorded. With this system, it is feasible to record holograms of large-scale objects, or scenes, since the hologram is built up pixel by pixel as the structured optical field is scanned over the object. In our research, special attention has been placed on demonstrating the ability to record holograms of 3D reflective objects as a basis for recording holograms of large-scale objects and/or a wide-field-of-view scene.

7.2 BACKGROUND AND THEORY

7.2.1 Holographic Recording

This technique, a derivative of OSH, is based on scanning the object or scene with a FZP and collecting the reflected light via a photodetector. First introduced by Poon and Korpel two decades ago, OSH has been demonstrated experimentally at the proof-of-principle level primarily on transmissive point and slit-type objects [2, 3]. Optical scanning holography has also been applied to 3D fluorescence microscopy with some success [4]. Here, the FZP scanning technique is combined with standard laser radar principles in an effort to record holograms of realistic, large-scale, reflective objects.

A number of system-level modifications have been incorporated into the OSH setup to adapt the technique to holography of 3D reflective objects. For instance, a diverging FZP field is generated, consisting of the superposition of two spherical waves of different radii of curvature, to scan the object. The diverging FZP scanning field is a key enabler for holographic recording of large objects over a large field of view. The FZP interference pattern is scanned over the object, with intensity reflectance $|\Gamma(x, y, z)|^2$, and the reflected light is collected by a photodetector. The scanning action results in the spatial convolution of the scanning field intensity $I_s(x, y; z)$ and the object, thus encoding each individual point in the object by a FZP. In this way, the photodetector current contains 3D information pertaining to the scanned object. The current is digitized and synchronized with the scanning action of the mirrors to form a digital image that is a hologram of the scanned object. The hologram can be described mathematically in terms of the convolution of the scanning field and the object as follows:

$$t(x, y) = I_s(x, y; z) * |\Gamma(x, y, z)|^2 \tag{7.1}$$

where the asterisk denotes convolution and $t(x, y)$ is the hologram of the object. Note that the hologram is a function of the lateral dimensions, x and y, but not the longitudinal dimension, z. In digital holography, the depth information is stored in a 2D digital image in the fringes of the interference pattern. The scanning field intensity has the form

$$I_s(x,\ y;\ z) \propto \cos\left[\frac{\pi}{\lambda z}(x^2 + y^2)\right] \tag{7.2}$$

where λ is the laser wavelength [5]. For the purposes of modeling and simulation, the FZP function of Eq. 7.2 is multiplied by a Gaussian multiplier, simulating that the pattern exists in a laser beam, giving

$$I_s(x,\ y;\ z) = \exp\left[-\frac{(x^2 + y^2)}{\alpha}\right]\cos\left[\frac{\pi}{\lambda z}(x^2 + y^2)\right] \tag{7.3}$$

where α represents the Gaussian roll-off.

7.2.2 Point Spread Function

It is useful to consider the scanning holography system in terms of its impulse response, or point spread function (PSF). Consider the object $|\Gamma(x, y, z)|^2$ as the input to a linear and shift-invariant (in x and y) system. If the resulting hologram $t(x, y)$ is taken as the output, the PSF is defined as the hologram of an ideal point object [6]. By definition, the input and output of a linear shift-invariant system are related according to

$$t(x,\ y) = h_\delta(x,\ y;\ z) * |\Gamma(x,\ y,\ z)|^2 \tag{7.4}$$

where $h_\delta(x,\ y,\ z)$ is the PSF. We see from examination of Eqs. 7.1 and 7.4 that the PSF and scanning field intensity are equivalent for this system, provided the system's receiver has an angular field of view larger than the beam divergence of the scanning pattern. This equivalence also follows from Eq. 7.1 since for a point object input $|\Gamma(x, y, z)|^2 = \delta(x = 0,\ y = 0,\ z = z_0)$ the output, or hologram, is equal to the scanning field evaluated at $z = z_0$. Explicitly, Eq. 7.1 is used to define the PSF as

$$\begin{aligned} h_\delta(x,\ y;\ z_0) &= I_s(x,\ y,\ z) * \delta(x = 0,\ y = 0,\ z = z_0) \\ &= I_s(x,\ y,\ z = z_0) \end{aligned} \tag{7.5}$$

In summary, the system PSF is equivalent to the scanning field and is dependent on the distance between the scanning mirrors and the object.

7.2.3 Image Reconstruction

Optical image reconstruction is the process that re-creates the image, or information pertaining to the object, from the recorded interference pattern, which is the hologram. In traditional holography, image reconstruction is accomplished by illuminating the hologram with the same reference wave used for recording. For digital holographic systems, such as computer-generated holography or OSH, optical image reconstruction is accomplished by displaying the hologram on an electro-optic device, such as a spatial light modulator (SLM). The purpose of the SLM in such a setup is to modulate coherent light in accordance with some input, in this case the hologram. In this manner, the hologram can be illuminated by a reference wave resulting in the image reconstruction [7]. Since the holographic recording technique described here is an electronic technique, the hologram exists as a digital image, or more precisely as a 2D array of numbers. In this case, an alternate option exists for image reconstruction based entirely on numerical methods.

For numerical image reconstruction, we make use of our a priori knowledge of how the hologram was recorded and, by Fourier techniques, extract the object information from the hologram. The result is, again, a digital image, typically displayed by computer. Digital holograms have been reconstructed numerically via a complete simulation of the traditional (optical) image reconstruction [5]. A similar but more straightforward technique for the numerical image reconstruction of digital holograms is described here. From Eq. 7.4 we see that the image reconstruction $|\Gamma(x, y, z)|^2$ can be extracted from the hologram $t(x, \mathrm{y})$ through a deconvolution operation with the system PSF. The system PSF can be modeled mathematically or obtained experimentally by using a point source as the input and recording the output. Each technique has advantages.

In the mathematically modeled case, the PSF is calculated from our knowledge of the scanning field, as given in Eq. 7.3. This technique is simple for a known scanning pattern and offers flexibility since the simulated pattern can be generated for various x, y, z and Gaussian roll-off values. In addition, the mathematical model of the system PSF will be "perfect," for example, noiseless, exactly circular, precise fringe spacing. Reconstruction via this perfect mathematical version of the PSF does have some drawbacks. Since the object data are encoded into the hologram experimentally, via an imperfect scanning field, the most accurate reconstruction will result from deconvolution with that same imperfect field. The actual scanning field is a product of nonideal optical, electro-optical, and electronic components as well as imperfect optical alignment. In fact, the more the actual scanning field differs from Eq. 7.3, the less likely it is that the mathematically generated field will be successful in reading out the information stored in the hologram. Obtaining an experimentally recorded PSF for image reconstruction will reduce potential distortion in the image reconstruction stemming from differences between the actual and modeled PSF.

Having the capability to obtain, experimentally, the system PSF and subsequently deconvolve it with the hologram, is important for another reason. Although a FZP scanning field is used in this experiment, other scanning patterns are certainly possible and may offer advantages over the standard FZP. In the past, for instance, an annular FZP scanning field has been proposed as a method of processing the holographic data during recording [5]. Scanning with an annular FZP was shown (via computer simulation) to result in edge extraction upon image reconstruction. If this or other very complex scanning fields are used to record holographic images, it may be difficult or impossible to accurately model the scanning field. If so, deconvolution with an actual system PSF is the only way to extract the image data from the hologram.

For computational efficiency, the deconvolution necessary for image reconstruction is accomplished in the Fourier domain using a 2D fast Fourier transform (FFT) algorithm. Employing the convolution theorem [8], Eq. 7.4 leads to

$$F\{t(x,\ y)\} = F\{h_\delta(x,\ y;\ z)\} \times F\{|\Gamma(x,\ y,\ z)|^2\} \tag{7.6}$$

where $F\{\ast\}$ denotes a Fourier transform. The image reconstruction is therefore calculated by dividing the Fourier transforms of the hologram and the PSF, or

$$F\{|\Gamma(x,\ y,\ z)|^2\} = \frac{F\{t(x,\ y)\}}{F\{h_\delta(x,\ y;\ z)\}} \tag{7.7}$$

An inverse Fourier transform operation completes the computation, resulting in the image reconstruction $|\Gamma(x, y, z)|^2$. For this experimental setup, the form of the PSF is known, so $h_\delta(x, y; z)$ can be generated numerically or recorded experimentally. Both experimentally recorded and numerically simulated PSFs are used for image reconstruction and presented in Section 7.4.

7.3 EXPERIMENTAL BREADBOARD FOR HOLOGRAPHIC LASER RADAR

As mentioned, this approach to holographic laser radar relies on scanning the object or scene with a FZP laser field and collecting the reflected light. For heterodyne detection, the FZP is temporally modulated by using two interfering waves at different temporal frequencies adjusted by acousto-optic devices. The experimental setup necessary to create the heterodyned FZP and scan the object is shown in Figure 7.1. The beam from a laser operating at frequency ω_0 is split into two paths by the first beamsplitter. For our experiment, $\lambda = c/\omega_0 = 632\,\text{nm}$, where c is the speed of light. The first path is simply collimated and then focused by lens L_1, producing a spherical wave at the operating frequency of the laser, ω_0. The second laser path is directed through an acousto-optic modulator (AOM) operating in the Bragg regime and diffracted into many frequency-shifted beams in accordance with the grating

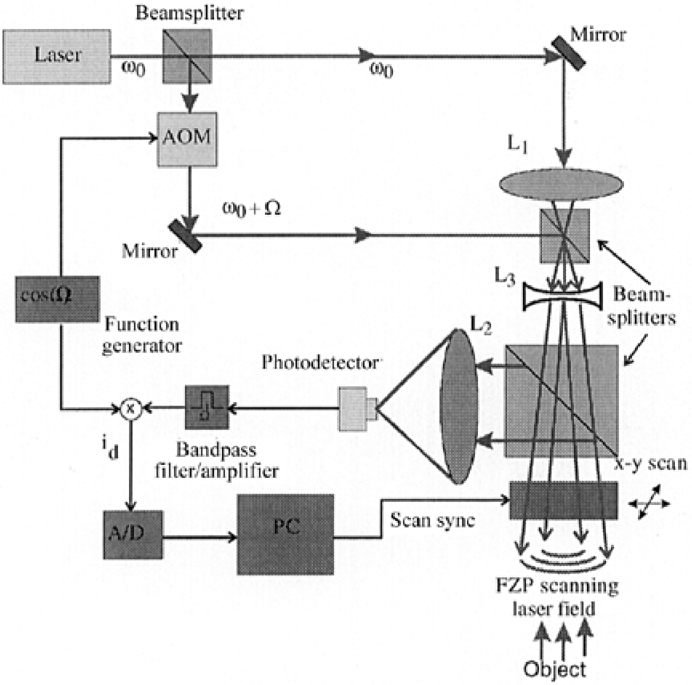


Figure 7.1 Experimental setup for holographic laser radar.

equation. We make use of the first-order diffracted light, which is shifted in frequency by Ω, where Ω is the operating frequency of the AOM. For this experimental setup, $\Omega = 40\,\text{MHz}$. This frequency-shifted beam, at frequency $\omega_0 + \Omega$, is combined collinearly with the first beam at the second beamsplitter. This optically heterodyned laser field is expanded using a diverging lens, L_3, resulting in a diverging interference pattern, which is the desired heterodyned FZP laser field. This interference pattern is scanned over the object, in a raster fashion, with intensity reflectance $|\Gamma(x, y, z)|^2$ and the reflected light is collected by a photodetector. The photodetector and beamsplitter are placed in such a way that the photodetector is scanned over the object or scene along with the scanning field. This experimental configuration helps relax the detector field-of-view requirements, an important consideration when recording holograms of large-scale objects or wide-angle scenes.

As mentioned previously, the scanning action results in the spatial convolution of the scanning field intensity and the object, thus encoding each individual point in the object by a FZP. The photodetector current therefore contains 3D information pertaining to the scanned object. As shown in Figure 7.1, the photodetector current is demodulated, amplified, digitized, and synchronized with the scanning action of the mirrors. The final image is a hologram of the scanned object.

7.4 EXPERIMENTAL RESULTS

To demonstrate the operation of the system described in Section 7.3, holograms have been recorded of simple, 3D reflective objects. One such group of objects, used to record some of the first holograms of this type, consists of three small, circular objects meant to approximate point source objects. Each point object consists of 2 mm aperture in front of a small piece of Scotchlite retroreflective tape and has a unique *x*, *y*, and *z* location. Photographs of this object set are shown in Figures 7.2 and 7.3. Figure 7.2 shows, approximately, the point of view from which the first hologram is recorded. Figure 7.3 is a photograph of the point objects from the side, showing the different depths at which each point object was placed. There are about 15 cm separating each circular object, so $z_1 = 69\,\text{cm}$, $z_2 = 84\,\text{cm}$, and $z_3 = 99\,\text{cm}$, where z represents the distance between the object and the focus of lens L_1. A hologram of this simple, 3D object was recorded using the setup shown in Figure 7.1 and is shown in Figure 7.4.

As described in Section 7.2, image reconstruction is achieved numerically by deconvolving the hologram with the system PSF. Here, we demonstrate that the PSF can be acquired experimentally and subsequently used to read out the 3D information from the hologram. By definition, the system PSF is the response, or output, to an input consisting of an ideal point source. An approximation of the system PSF, $h_\delta(x, y; z = z_n)$, is therefore obtained experimentally by recording a hologram of a point object, $\delta(x, y, z = z_n)$. The first example of an experimentally recorded PSF, $h_\delta(x, y; z = z_1)$, is shown in Figure 7.5. This system PSF was recorded experimentally by obscuring the point objects located at $z = z_2$ and $z = z_3$ (see Figs. 7.2 and 7.3) and recording a hologram in the same fashion as the three-point hologram, shown in Fig. 7.4. Similarly, PSFs for the two other depths of interest, $h_\delta(x, y; z = z_2)$, and

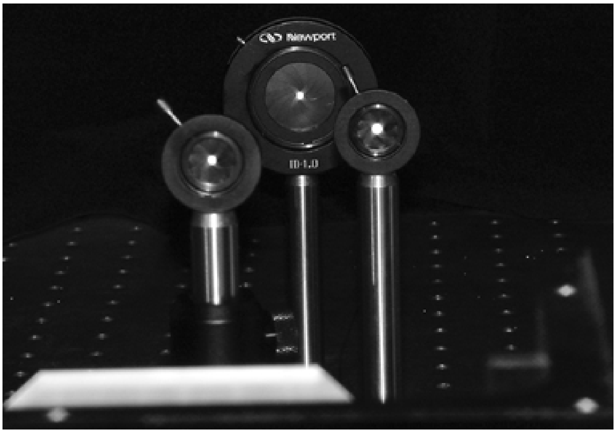


Figure 7.2 Actual 3D reflective object from point of view of hologram recording.

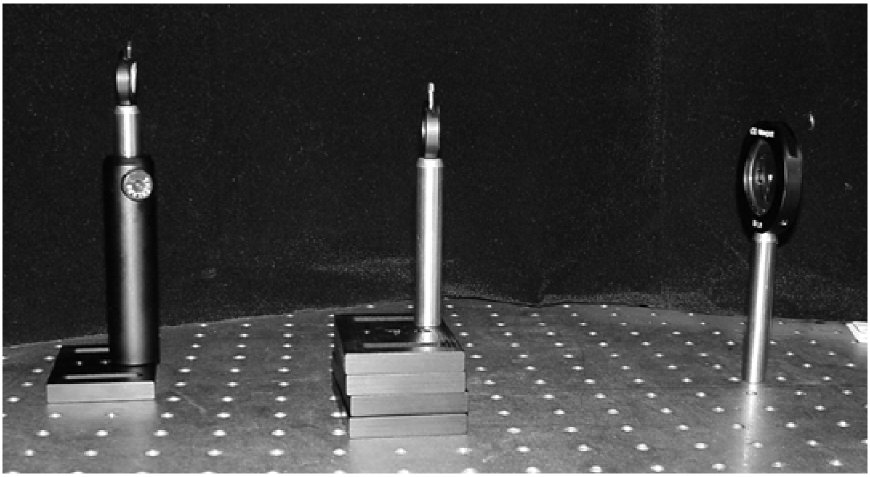


Figure 7.3 Side view of 3D reflective object to show object depth.

$h_\delta(x, y; z = z_3)$, were recorded and are shown in Figures 7.6 and 7.7. Each PSF is used to read out the information stored in the hologram at a single depth. This is possible because the system PSF for any given depth, say $z = z_1$, can be used to reconstruct the image for that particular depth. To demonstrate this, $h_\delta(x, y; z = z_1)$, shown in Figure 7.5, is deconvolved with the hologram $t(x, y)$ (Fig. 7.4), resulting in the image reconstruction shown in Figure 7.8. As

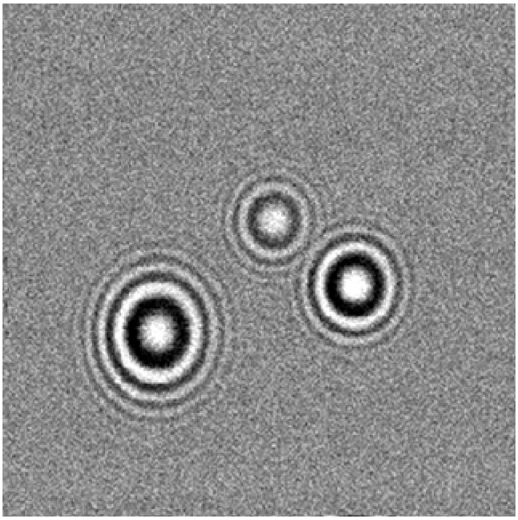


Figure 7.4 Experimentally recorded hologram of three-point objects (see photograph in Figs. 7.2 and 7.3) using holographic laser radar.

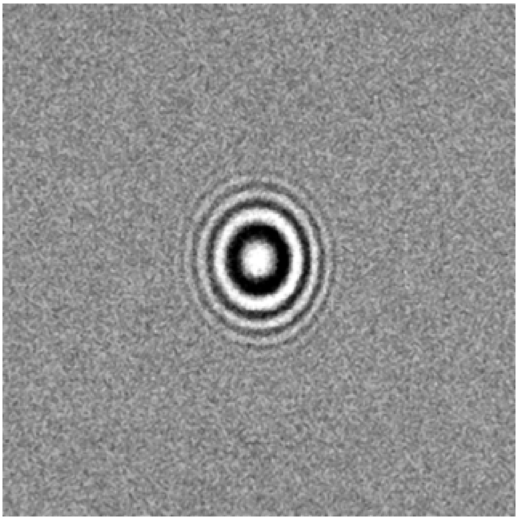


Figure 7.5 Experimentally recorded system PSF for $z = z_1$, denoted $h_\delta(x, y; z = z_1)$.

described above, the deconvolution operation is performed using a 2D FFT algorithm and Eq. 7.5. Note that the point corresponding to the first point object at distance $z = z_1$ is accurately reconstructed in Figure 7.8 as a sharp point, while the other two point objects appear as defocused blur spots in this plane, as expected. The point object located at the z plane corresponding to z_2

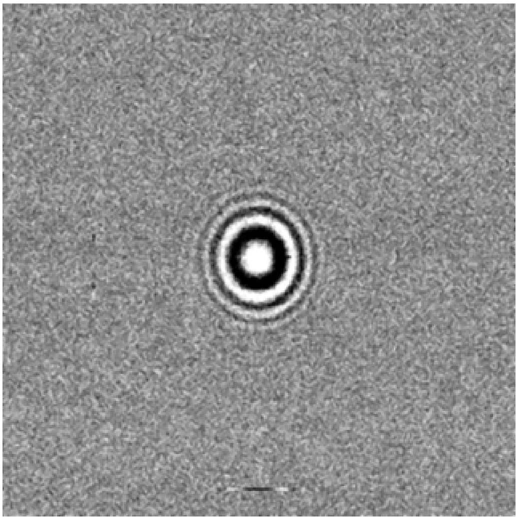


Figure 7.6 Experimentally recorded system PSF for $z = z_2$, denoted $h_\delta(x, y; z = z_2)$.

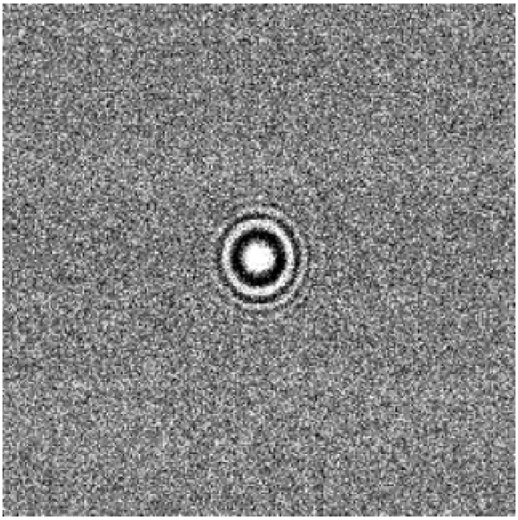


Figure 7.7 Experimentally recorded system PSF for $z = z_3$, denoted $h_\delta(x, y; z = z_3)$.

is accurately reconstructed by deconvolving the hologram with the system PSF for $z = z_2$ [denoted $h_\delta(x, y; z = z_2)$ and shown in Fig. 7.6). The resulting image reconstruction is shown in Figure 7.9. Likewise, using $h_\delta(x, y; z = z_3)$ to reconstruct image plane $z = z_3$ from the hologram results in the digital image shown in Figure 7.10.

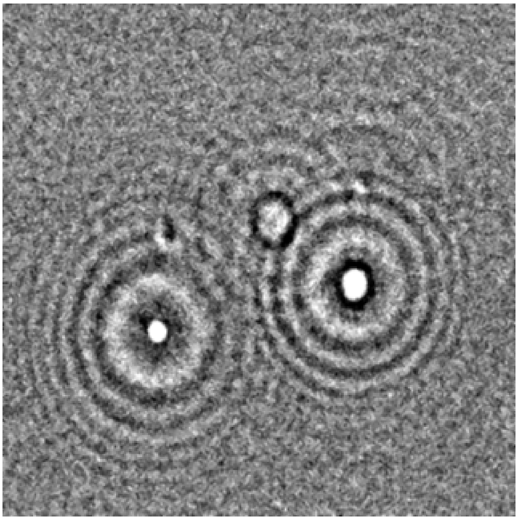


Figure 7.8 Numerical reconstruction of plane $z = z_1$, from hologram in Fig. 7.4.

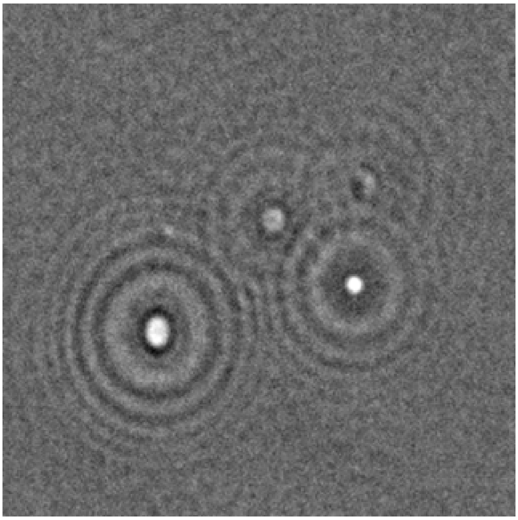


Figure 7.9 Numerical reconstruction of plane $z = z_2$, from hologram in Fig. 7.4.

For comparison, a numerically generated system PSF was generated and used for image reconstruction. The form of the PSF is given by Eq. 7.3, or

$$h'_\delta(x, y; z) = \exp\left[-\frac{(x^2 + y^2)}{\alpha}\right]\cos\left[\frac{\pi}{\lambda z}(x^2 + y^2)\right] \tag{7.8}$$

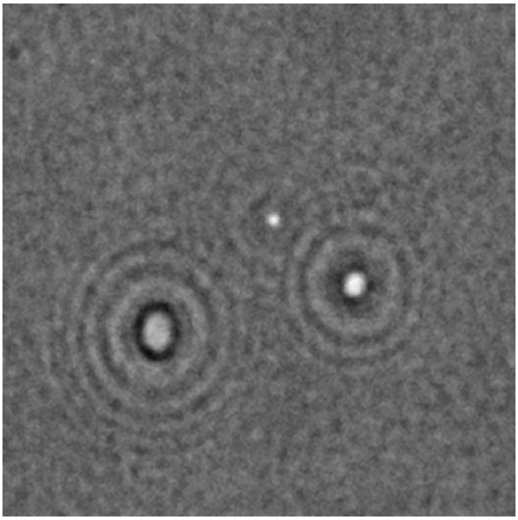


Figure 7.10 Numerical reconstruction of plane $z = z_3$, from hologram in Fig. 7.4.

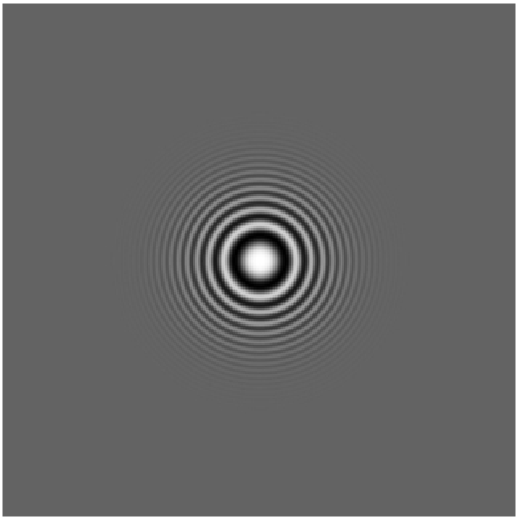


Figure 7.11 Simulated system PSF for $z = z_2$, denoted $h'_\delta(x, y; z = z_2)$.

where the prime indicates the numerically generated PSF. The simulated PSF for $z = z_2$ is shown in Figure 7.11. (The corresponding experimentally recorded PSF is shown in Fig. 7.6.) Note the absence of noise in Figure 7.11 and the presence of multiple fine fringes. Figure 7.12 shows the image reconstruction of plane $z = z_2$ resulting from the numerical deconvolution of Figure 7.11 with

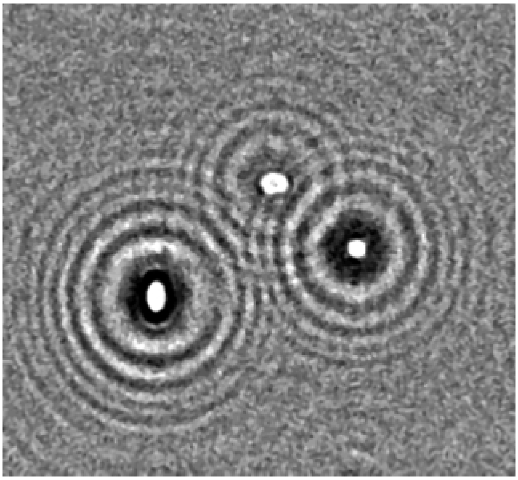


Figure 7.12 Numerical reconstruction of plane $z = z_2$, from hologram in Fig. 7.4 using modeled PSF shown in Fig. 7.11.

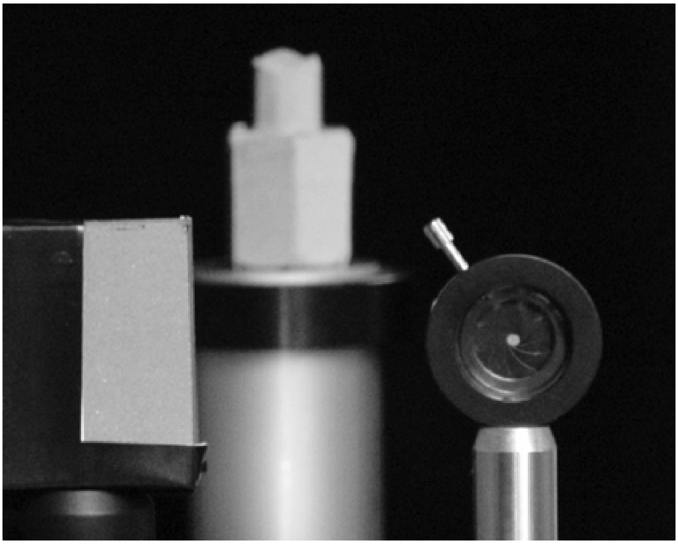


Figure 7.13 Photograph of second data set: point object, rectangular object, and cylindrical object.

Figure 7.4. For comparison, Figures 7.9 and 7.12 are the numerical reconstructions using an experimentally recorded PSF and the mathematically generated PSF, respectively.

Finally, a hologram was recorded of a second set of objects, consisting of a point object, a rectangular object, and a cylindrical object. A photograph of a front view of the second data set is shown in Figure 7.13. The three objects shown in Figure 7.13 are positioned at the same distances from the scanners as the point objects shown in Figures 7.2 and 7.3: $z_1 = 69$ cm, $z_2 = 84$ cm, and $z_3 = 99$ cm. A hologram of the second set of objects was recorded and is shown in Figure 7.14. Image reconstructions are obtained as described above for the first hologram using the three experimentally recorded PSFs, shown in Figures 7.5–7.7. Image reconstructions for the three planes of interest are given in Figures 7.15–7.17. Figure 7.15 shows the image reconstruction for the $z = z_1$ plane, which is the most accurate reconstruction of the point object. Likewise, Figure 7.16 is the reconstruction of the rectangular object, while Figure 7.17 shows the reconstruction for the cylindrical object.

Summarizing our results, the two sets of image reconstructions (Figs. 7.8–7.10 and 7.15–7.17) show that this system is capable of storing three-dimensional information, both location (x, y) and depth (z), pertaining to an object or scene. A particular plane of interest can be reconstructed numerically by deconvolving the hologram with the system PSF corresponding to the desired depth, z. In addition, image reconstruction has been demonstrated via numerical deconvolution with an experimentally acquired system PSF.

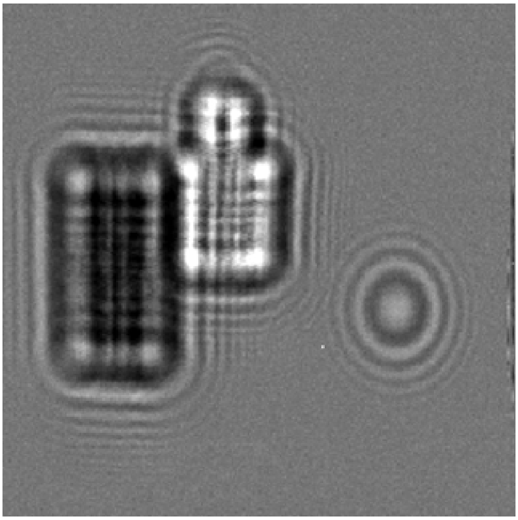


Figure 7.14 Experimentally recorded hologram of three objects (see photograph in Fig. 7.13).

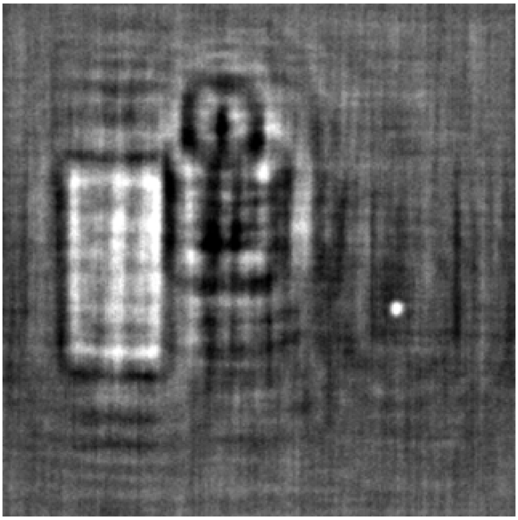


Figure 7.15 Numerical reconstruction of plane $z = z_1$, from hologram in Fig. 7.14.

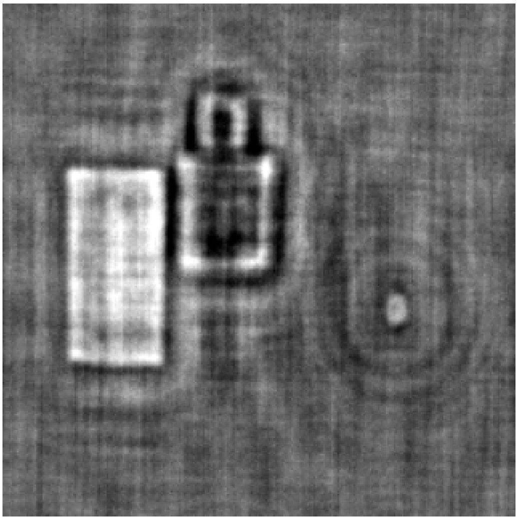


Figure 7.16 Numerical reconstruction of plane $z = z_2$, from hologram in Fig. 7.14.

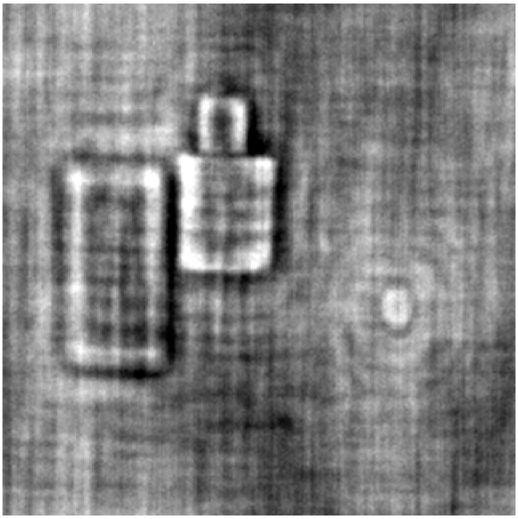


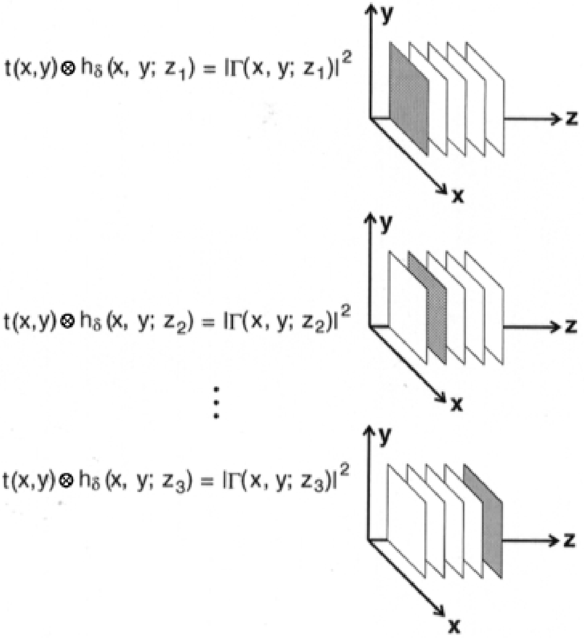
Figure 7.17 Numerical reconstruction of plane $z = z_3$, from hologram in Fig. 7.14.

7.5 ADVANCED NUMERICAL TECHNIQUES FOR HOLOGRAPHIC DATA ANALYSIS

In the situation above, a hologram of three objects placed at three z distances was recorded using the holographic laser radar technique. The system PSF was recorded at each depth of interest and subsequently used to reconstruct object data stored in the hologram. This technique has the drawback that a system PSF must be recorded for each depth of interest to perform image reconstruction for that depth. A more robust reconstruction process, based on numerical Fresnel diffraction, has been developed that can read out all the data stored in the hologram starting with a single, measured PSF at a known depth z_0.

A hologram recorded using the holographic laser radar technique described here contains information pertaining to a continuous 3D volume of space, as shown in Figure 7.18. The volume of data stored in the hologram is represented as a series of images stacked one after another along the z direction. Accessing the various pages of information, each page corresponding to a different depth, is accomplished by deconvolving the hologram with the system PSF corresponding to the depth of interest, according to

$$|\Gamma(x, y, z_n)|^2 = t(x, y) \otimes h_\delta(x, y; z = z_n) \tag{7.9}$$



t(x,y)⊗hδ(x, y; z1) = |Γ(x, y; z1)|2
t(x,y)⊗hδ(x, y; z2) = |Γ(x, y; z2)|2
⋮
t(x,y)⊗hδ(x, y; z3) = |Γ(x, y; z3)|2
y
z
x
y
z
x
y
z
x

Figure 7.18 Hologram contains continuous 3D information from a volume in space. Particular plane is reconstructed via deconvolution operation between hologram and appropriate PSF.

where $\otimes$ denotes deconvolution and $z = z_n$ corresponds to the plane of reconstruction. To (numerically) recover the entirety of the 3D information stored in the hologram, therefore, requires only that the system PSF be available for all z in the volume of interest. This is possible through the application of the Fresnel diffraction formula to an experimentally recorded system PSF. Given a system PSF corresponding to a known plane $h_\delta(x, y; z_0)$, the PSF for another depth, say $h_\delta(x, y; z_n)$, can be determined by propagating the original field intensity pattern to the plane of interest. This is accomplished mathematically through a convolution process with the free-space impulse response in accordance with the following expression [9]:

$$h_\delta(x,\ y;\ z_n) = h_\delta(x,\ y;\ z_0) * h_f(x,\ y; [z_n - z_0]) \tag{7.10}$$

where the asterisk denotes convolution and $h_f(x, y; z)$ is the free-space impulse response given by

$$h_f(x,\ y;\ z) = \frac{1}{j\lambda z} \exp\left[-\frac{j\pi}{\lambda z} (x^2 + y^2) \right] \tag{7.11}$$

From Eq. 7.10, discrete system PSFs are calculated for the volume of interest. Not surprisingly, the convolution of Eq. 7.10 is most efficiently computed by invoking the convolution theorem in conjunction with the FFT algorithm. Specifically, Eq. 7.10 becomes

$$h_\delta(x,\ y;\ z_n) = F^{-1}\{H_\delta(f_x, f_y,\ z_0) \times H_f(f_x, f_y,\ z_0 - z_n)\} \tag{7.12}$$

where $H(f_x, f_y, z)$ is the Fourier transform of $h(x, y; z)$ and f_x and f_y are the spatial frequency-domain variables in the x and y directions, respectively. Note that an expression for the Fourier transform of $h_f(x, y; z)$, commonly refered to as the spatial transfer function for free space, is known [9]:

$$H_f(f_x, f_y;\ z) = \exp\left[j2\pi \frac{z}{\lambda} \sqrt{1 - (\lambda f_x)^2 - (\lambda f_y)^2} \right] \quad \text{for } \sqrt{(\lambda f_x)^2 - (\lambda f_y)^2} < \frac{1}{\lambda} \tag{7.13}$$

For a discretely sampled function, the relationship between x and f_x is $f_x = 1/N\Delta x$, where N is the number of elements in the array. It can be particularly advantageous to use Eq. 7.13 instead of Eq. 7.11 if the choices for x, y, and z lead to significant aliasing in the digital representation. To reconstruct a certain $z = z_n$ plane of the hologram, Eq. 7.12 is substituted into Eq. 7.9, resulting in

$$|\Gamma_r(x,\ y,\ z = z_n)|^2 = F^{-1}\left(\frac{T(f_x,\ f_y)}{H_\delta(f_x,\ f_y;\ z_0) \times H_f(f_x,\ f_y; (z_0 - z_n)} \right) \tag{7.14}$$

To demonstrate the feasibility of this approach and its application to image

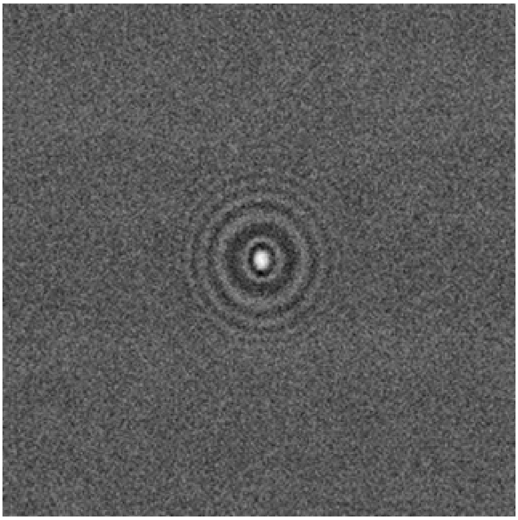


Figure 7.19 Numerically propagated PSF for $z = z_1$ plane.

reconstruction, the algorithm was applied to an experimentally recorded PSF. The PSF, denoted $h_\delta(x, y; z = z_3)$, shown in Figure 7.7, was used as the baseline from which to derive multiple PSFs using Eq. 7.12. The propagated system PSF corresponding to the $z = z_1$ is shown in Figure 7.19 and that for $z = z_2$ is shown in Figure 7.20. These calculated field intensities are subsequently

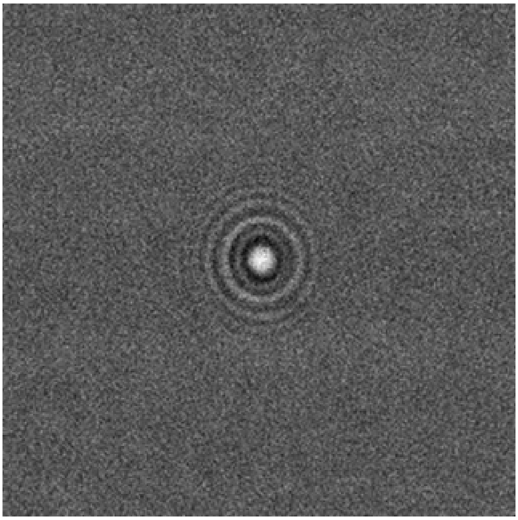


Figure 7.20 Numerically propagated PSF for $z = z_2$ plane.

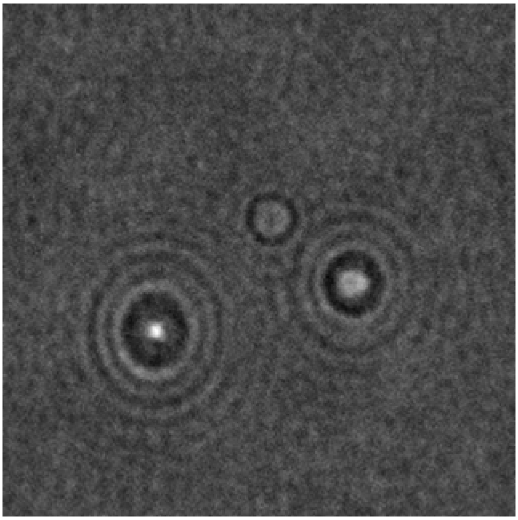


Figure 7.21 Numerical reconstruction of plane $z = z_1$, from hologram in Fig. 7.14 using propagated PSF shown in Fig. 7.19

deconvolved with the original hologram in accordance with Eq. 7.14, resulting in the image reconstructions for the two planes of interest. The reconstructed point images are shown in Figures 7.21 and 7.22 for $z = z_1$ and $z = z_2$, respectively.

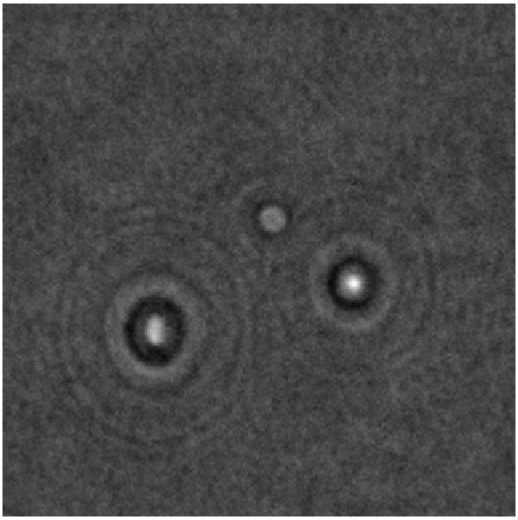


Figure 7.22 Numerical reconstruction of plane $z = z_2$, from hologram in Fig. 14 using propagated PSF in Fig. 7.20.

This approach to data analysis recorded via the holographic laser radar system is important because it represents a way to easily extract all of the 3D data that has been stored in the hologram. Further processing to remove noise from out of focus elements is straightforward and currently under investigation.

7.6 CONCLUSIONS

Holographic laser radar, a laser scanning system for recording 3D information of reflective objects, has been demonstrated. The experimental setup, based on OSH, was described. Holograms are recorded with the system by scanning an optically heterodyned FZP over the object or scene in a raster fashion. Radiation is reflected from the object and detected using a photodetector, resulting in an electronic hologram displayed and stored on a computer. Image reconstruction was addressed in terms of the system PSF and image readout was demonstrated using simulated and experimentally recorded PSFs of the recording system. Numerically propagated PSFs were generated via simulated Fresnel diffraction of an experimentally recorded PSF and used to reconstruct multiple image planes. Holographic recording and image reconstruction was demonstrated for a simple three-point object, and a slightly more complicated three-object data set, using this technique.

REFERENCES

1. T.-C. Poon, B. W. Schilling, M. H. Wu, K. Shinoda, and Y. Suzuki, "Real-time two-dimensional holographic imaging by using an electron-beam-addressed spatial light modulator," *Opt. Lett.* **18**, (1), 63–65 (1993).
2. T.-C. Poon, and A. Korpel, "Optical transfer function of an acousto-optic heterodyne image processor," *Opt. Lett.* **4**, 317–319 (1979).
3. T.-C. Poon, "Scanning holography and two-dimensional image processing by acousto-optic two-pupil synthesis," *J. Opt. Soc. Am. A* **2**, 521–527 (1985).
4. B. W. Schilling, T.-C. Poon, G. Indebetouw, B. Storrie, M. H. Wu, K. Shinoda, and Y. Suzuki, "Three-dimensional holographic fluorescence microscopy," *Opt. Lett.* **22**(19), 1506–1508 (1997).
5. B. W. Schilling, and T.-C. Poon, "Real-time preprocessing of holographic information," *Opt. Eng.*, **34**(11), 3174–3179 (1995).
6. B. E. A. Saleh, and M. C. Teich, *Fundamentals of Photonics*, John Wiley & Sons, New York, 1991.
7. T.-C. Poon, K. B. Doh, B. W. Schilling, K. Shinoda, Y. Suzuki, M. H. Wu, "Holographic three-dimensional display using an electon-beam-addressed spatial light modulator," *Opt. Rev.* **4**(5), 567–571 (1997).
8. J. S. Walker, *Fast Fourier Transforms*, CRC Press, Boston, 1991.
9. J. W. Goodman, *Introduction to Fourier Optics*, McGraw-Hill, New York, 1968.

Three-Dimensional Holographic Imaging. Edited by Chung J. Kuo, Meng Hua Tsai
Copyright © 2002 John Wiley & Sons, Inc.
ISBNs: 0-471-35894-0 (Hardback); 0-471-22454-5 (Electronic)

CHAPTER 8

Photoelectronic Principles, Components, and Applications

OSCAL T.-C. CHEN, WEI-JEAN LIU, ROBIN SHEEN, and JEN-SHANG HWANG

Department of Electrical Engineering
National Chung Cheng University
Chia-Yi, 621 Taiwan

FAR-WEN JIH, PING-KUO WENG, LI-KUO DAI, SHIANG-FENG TANG, and KAUNG-HSIN HUANG

Solid-State Devices Materials Section
Materials and Electro-Optics Research Division
Chung-Shan Institute of Science and Technology
Tao-Yuan, 325 Taiwan

In this chapter, we present the principles, components, and applications of conversions between photons and electrons. To understand photodiodes, light-emitting diodes (LED), and light amplification by the simulated emission of radiation (Laser) diodes, the basic theories behind light-receiving and light-emitting components are briefly reviewed. Based on photodiodes and laser diodes, two photoelectronic applications of the image sensor and optical interconnection systems are introduced. In an image sensor design, the charge-coupled-device (CCD) and complementary metal–oxide–semiconductor (CMOS) technologies are compared to analyze their features, performances, and applications. Passive and active pixel sensors and a read-out circuit for the CMOS image sensor are explored, and fill factors, quantum efficiencies, and fixed pattern noises of these pixel sensors are also illustrated. In optical interconnection applications, we demonstrate several circuit designs of photoreceivers and laser diode drivers operated at 1 GHz by using the Taiwan Semiconductor Manufacturing Company (TSMC) 0.35-μm CMOS technology. To integrate four photodiodes, photoreceivers, laser diode drivers, and laser diode pads, a prototype chip with a die size of 1.8×1.5 mm was implemented for chip-to-chip interconnection. The photo responses of the CMOS photodiodes implemented by N-well P substrate, N-diffusion P-well P substrate, and P-diffusion N-well P substrate are measured. The N-well P-substrate photodiode yields the least dark current. The N-diffusion P-well P-substrate photodiode has the best photo response in visible light.

8.1 LIGHT-RECEIVING COMPONENTS

8.1.1 Principles of Photodiodes

The basic principle behind photodiodes is to convert light into electrical currents. When incident light reaches the diode's sensing region, it stimulates the positive and negative particles in the diode's P–N contact depletion region and areas around the depletion region to create electron–hole pairs (EHPs). This occurrence of EHPs would change the contact region from a semiconducting nature to a conducting one. With this and the potential difference between the P and N layers, the contact region would be able to produce electrical currents. Commonly, there are two ways to illustrate the photoelectronic conversion phenomenon [1]. One is when the light is incident upon an object; the object is able to absorb photoenergy and then release EHPs to produce electrical currents. The other scenario is when an object absorbs enough electrical energy to stimulate electrons on the valence band; these electrons in turn release photons during energy band transfer. The photodiode uses the former method by absorbing photons to produce electrical currents.

To better understand the photodiode, we must first understand the differences between semiconductors, conductors, and insulators. Differences between the semiconductor and the other two are determined by the valence band, band gap, and conduction band. An insulator's band gap is very wide, making it difficult for electrons to move from the valence band to the conduction band. A conductor's band gap has a width of near zero, which means the valence and conduction bands almost overlap making it much easier for electrons to move from one to the other. The width of a semiconductor band gap lies in between those of the insulator and the conductor. Under normal circumstances, semiconductors can be considered insulators. Only when either light or heat is presented to induce electrons to transfer from the valence band to the conduction band, a semiconductor employs a conducting property. The basic conductivity principle of the photodiode is such a process like the semiconductor. When energy from incident light exceeds the band gap, electrons can move from the valence band to the conduction band. This process allows the number of free EHPs to increase, which in turn increases the conductivity of the photodiode. To generate photocurrents, there must be a reverse-biased voltage relationship between P and N layers of semiconductors. Electrons from the P layer would spread in the direction of the N layer when the photodiode has a conducting property. If the distance of the spread is large enough, electrons in the P layer can reach the depletion region, where they are excited by the electrical field due to the reverse-biased voltage of the P–N layers and accelerate into the N layer. However, the number of electrons or holes in the depletion region remains the same. Electrons reaching the N layer would become majority carriers while holes would shift in the opposite direction to the P layer. The EHPs in the depletion region then creates a photocurrent.

The depletion region plays a major role in operating the photodiode under a reverse-biased voltage, which is the most commonly applied method in operating a photodiode. When the P-type semiconductor connects with the N-type semiconductor, it creates a depletion region in the contact area. This depletion region also possesses a semiconducting property; but when there is no incident light, the contact region acts as an insulator. When incident light hits the diode's sensing region, it creates EHPs in and near the depletion region. Those EHPs in and near the depletion region can be directed by the electrical field to move in the right direction so only those EHPs in the area effectively produce carriers to generate photocurrents. Thus, in designing a photodiode, the length and location of the depletion region are crucial considerations because they directly affect the amount of photocurrent produced, which then affects the response speed of the photodiode. A depletion region's capacitance characteristic affects the length of the region. Decreasing the capacitance of the depletion region would mean few carriers are available to form a conducting channel; at the same time it increases the length of the depletion region to allow more time for the carriers to drift. This means the length of the depletion region must be optimized to achieve a good performance of the photodiode. Taking all these into consideration, it is common to operate the photodiode under reverse-biased voltage. There are a few advantages in doing so [2]:

1. Reverse-biased voltage can increase the length of the depletion region, which then decreases the capacitance of the depletion region.
2. Reverse-biased voltage can create a more forceful electrical field within the depletion region to increase drifting of free carriers, which would normally decrease with an increase in the length of the depletion region.
3. Under normal circumstances, it is expected that the photodiode uses the presence of light to control the on/off operation. When there is incident light, the photodiode is turned "on" and generates photocurrents and vice versa when there's no incident light. If forward-biased voltage is applied to the photodiode, it is more likely to be turned "on" when there are no lights and creates currents based on potential differences between the P and N layers of the photodiode. This then fails to meet the expectation of a photodiode using light to control the on/off operation.

8.1.2 Types of Photodiodes

Other than the P–N photodiode, another commonly used photodiode is the P–I–N photodiode. It is different from the P–N photodiode discussed earlier

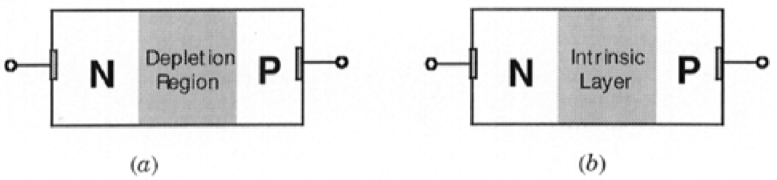


Figure 8.1 (*a*) P–N and (*b*) P–I–N photodiodes.

because a high impedance layer is added between where P and N layers connect. By doing so, it would make the capacitance of the contact region smaller, thus increasing the photodiode's sensitivity to light. Figure 8.1 shows the difference between a P–N photodiode and a P–I–N photodiode. From Figure 8.1*a*, it is clear that the length of the depletion region of a P–N photodiode is controlled by the degree of reverse-biased voltage present. Since the length of a depletion region is normally small using a reverse-biased voltage, a P–I–N photodiode is designed to include a high-impedance intrinsic (I) layer between P and N layers that can function as a depletion region. This would allow designers to easily adjust the length of the depletion region to a level that is optimized for achieving good quantum efficiency and fast response.

Although not commonly used before, the CMOS photodiode is fast gaining presence today. Due to the high integration, low power dissipation, and low cost of the CMOS technology, it is being widely used in designing integrated circuits. Therefore it is not surprising to see it in the photoelectronic industry, particularly in the design of CMOS photodiodes. Some commercial CMOS technologies may not provide the high-impedance layer to design the P–I–N photodiode; thus it is proposed to design the CMOS photodiode shown in Figure 8.2 [3]. Basically it is to make P and N layers by using the P-substrate and N-diffusion materials in the CMOS technology, respectively. There is a

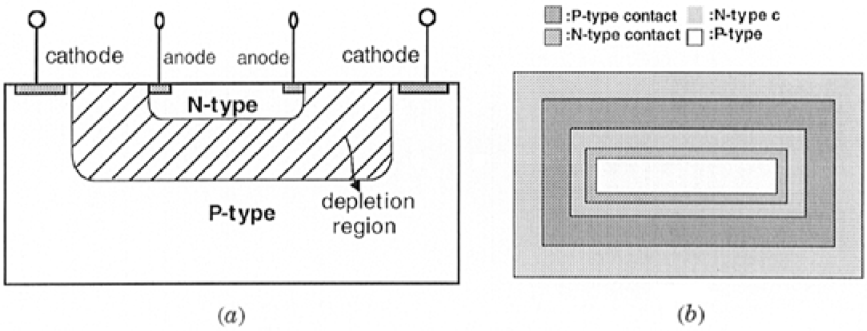


Figure 8.2 CMOS P–N photodiode: (*a*) cross section; (*b*) top view.

P–N contact region between the N-diffusion and P-substrate materials; when light incites on the sensing region, the P–N contact region induces carriers to form a conducting channel. The potential difference between the P and N layers causes the electrons to move from P to N layers while the holes drift from N to P layers; thus photocurrent is produced. Without light, the conducting channel will not form, similar to the switch being turned “off” in a standard photodiode when no light is present.

The cross section of the photodiode using the commercial CMOS technology is quite different from that of the conventional P–N photodiode shown in Figure 8.1*a*. The electrodes of the conventional P–N photodiode are located in two sides, and their electric field is distributed in one direction. Thus the depletion region can distribute in the junction area. Figure 8.2*a* shows the cross section of a CMOS P–N photodiode. We can see that the electrodes of the photodiode are located on the same plane and the depletion region is under this plane. If there is a potential difference between these two electrodes, the direction of the electric field is much different from that of the conventional P–N photodiode. In addition, its depletion region is also very different. Because of the electrodes in the same plane, the current collection capability of the CMOS P–N photodiode is weaker. To overcome this effect, the ring-type electrodes shown in Figure 8.2*b* are designed to enhance the current collection capability.

In addition to the photodiodes discussed above, a sensor unit can also be achieved by using the photogate. Figure 8.3 shows its structure to be similar to that of a sensor unit implemented by CCD technology. Developed by Jet Propulsion Laboratory [4], the photogate is able to store sensed charges under the control switch of the photogate (PG). The control switch of TX will be turned “on” when signals are ready to be sent out. At this time, the charge is transferred to TX’s other output diffusion, which is read by an amplifier circuit. The design of a photogate takes up more space, and therefore fill factor and quantum efficiency are not comparable to those of the other photodiodes. However, it is able to reduce the noise effects to minimal, which a CMOS P–N photodiode cannot do.

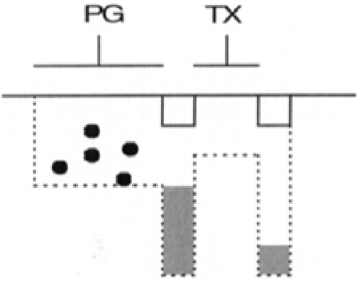


Figure 8.3 Photogate [4].

8.2 LIGHT-EMITTING COMPONENTS

8.2.1 Principles

The emitting of light from a diode results from conversion of electroenergy to photoenergy. To put it simply, when the P–N contact region is under forward-biased voltage, the holes in the P layer would drift to the N layer. This shifting from high- to low-energy levels results in emission of energy in the form of light or heat. In 1916, Albert Einstein confirmed that the interrelationship between light and objects can be described using three basic processes: stimulated absorption, stimulated emission, and spontaneous emission. The process of stimulated absorption was illustrated in the last section in the discussion of an object absorbing light and converting it into electroenergy. Spontaneous emission is when an object releases energy in the form of light or heat when its energy is transformed from high to low energy levels. This is also the theory behind the LED. The more complicated light-emitting process is stimulated emission. Spontaneous emission is different from stimulated emission; spontaneous emission has a shorter life cycle during the excitation stage, which means it does not require external energy to produce radiation. Radiation is instantaneous and natural, forcing ions in the excitation stage back to the lower energy level to emit light. On the other hand, stimulated emission has a longer life cycle during the excitation stage, which means it cannot automatically release energy to return to the lower energy level but instead requires the same amount of incident light to excite its radiation. This phenomenon causes incident light to appear amplified. Radiation light and incident light are very similar in that they have the same energy level, coherent direction, and phase. That said, laser light refers to light amplification by stimulated emission of radiation, which is from where the term *laser* is derived [5].

8.2.2 Types of Components

There are two types of light sources: the LED and laser. The light-emitting process of the LED is of instantaneous radiation, which can be further divided into two types: edge-emitting and surface-emitting components. The hardware design of the edge-emitting component is simpler, as shown in Figure 8.4. From the production perspective it is easier to make the edge-emitting component because the opening where light projects out is located parallel to the P–N contact surface. For the surface-emitting component, Figure 8.5 shows that the opening and the P–N contact surface is perpendicular to each other. This means a well-structured substrate contact must be included in the hardware design to guide the light out. Generally speaking, the surface-emitting component has a higher light-emitting efficiency, which makes it most suitable for light sources of fiber optics.

Laser can also be separated into side-emitting and surface-emitting components from the hardware design perspective. It can also be looked at from its

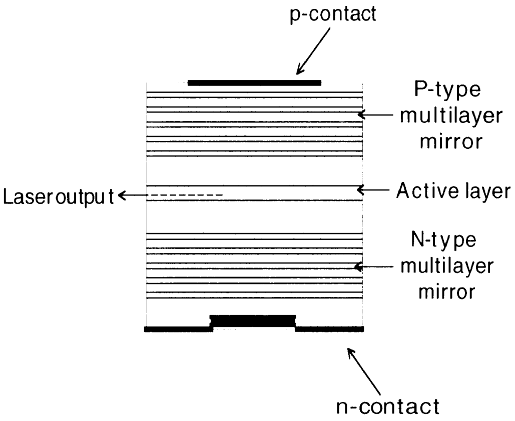


Figure 8.4 Edge emitting component.

different formation processes: solid-state laser, liquid laser, gas laser, and semiconductor laser. Of these, the semiconductor laser is perhaps the most developed and widely used, most notably in the extensive research and development done on the vertical cavity surface-emitting laser (VCSEL). The VCSEL has many advantages, including easy focus and packaging, low threshold currents, single frequency, and single-mode light emission. If used with other semiconductor components such as transistors, passive components, and modulators, the VCSEL would have even more flexibility and potential in what it is able to accomplish, such as in infrared free-space optical communication.

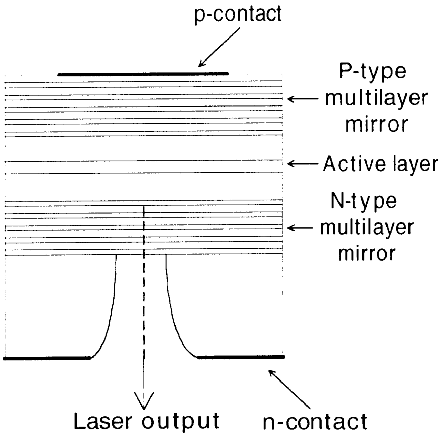


Figure 8.5 Surface-emitting component.

8.3 APPLICATIONS OF PHOTOELECTRONIC COMPONENTS

8.3.1 Image Sensors

The development of transistors opened the door to the microelectronics era. In addition, digital evolution has brought us to a multimedia world. With the digital process, integrated multimedia information can be provided via communication, broadcasting, and Internet networks. In the various media, image is one of the most important sources to access information for mankind because it is easiest to be understood. Normally, humans have reactions to images when the brain senses light, which is reflected from an object and projected into the eyes. Under the same theory, to capture an image, an image sensor unit with the capability to sense light is needed. Among all photoelectronic components, the photodiode is most commonly used as a light sensor component.

Currently CCD and CMOS technologies are being used in the applications of digital image sensors [6]. Both technologies are equipped with photodiodes but differ in how they read signals. The CCD functions by placing many gate electrodes on top of a long substrate. The capacitor between every gate electrode and the substrate is to store charges. Given that each electrode has a different voltage level, these electrodes act as a string of memory cells or shift registers in transmitting voltage from the transistor's source nodes to the drain nodes. These output charges then transmit through the output node to the backend processing circuit. Due to the simplicity and high density of its circuits, the CCD had been widely used in image sensor systems from the start and is still in great demand in the market today.

The CMOS technology has come a long way but is gaining presence in the image sensor system market today [7, 8]. Due to the high integration, low power consumption, and low cost of CMOS technology, it is now widely used in the application of digital and analog circuit designs. To achieve the "camera on a chip" from an image sensor system, using CMOS technology as the integrated process for the entire system is the only option available today. This is why CMOS is beginning to be applied in designing image sensors.

Table 8.1 compares CMOS and CCD image sensors. It is clear from the table that the image resolution for the CCD can reach tens of millions of pixels while the CMOS technology can only achieve a few million pixels. Also, in image quality, the CCD technology is up to 14 bits while the CMOS technology is between 8 and 10 bits. The image resolution and quality values clearly reveal that the CCD is the better application in image sensor system design. The CCD technology had always focused on improving its light sensor capability while the CMOS technology has emphasized electricity characteristics, and that is the main reason CMOS technology is still far behind CCD technology in image resolution and quality.

However, this does not mean that CMOS technology has no advantages over CCD technology. With the improvement in data processing capability of

TABLE 8.1 Comparison of CMOS and CCD Image Sensor Units

Features	CMOS	CCD
Resolution (pixel)	2M	63M
Image quality	8–10 bits	Up to 14 bits
Power consumption	10's of mW	100's of mW
Power supply	Single voltage level	Multiple voltage levels
System integration	Camera on a chip	Needs multiple chips
$X-Y$ addressing	Done on chip	Not possible

the personal computer and the popularization of multimedia, digital image sensor systems are in high demand. Professionals require a higher-resolution image sensor, whereas other users only require an image sensor with enough resolution to be used on multimedia applications. In this regard, CMOS technology has a definite competitive edge over CCD technology. As listed in Table 8.1, CMOS technology demands less power consumption during sensing operations of its image sensor. In power supply, the CMOS image sensor only needs a single supply voltage level while the CCD image sensor requires multiple levels. Multiple levels involve additional voltage suppliers, which means additional hardware is required since it cannot share the same supply voltage with the signal processing circuits. The additional hardware not only increases the production cost but also decreases the CCD's applicability since it is unable to integrate with the other circuit systems. In comparison, the CMOS technology is able to provide the system-on-chip design to allow the sensor unit, read-out circuit, and data processing circuit to integrate on the same chip. This not only decreases cost but also increases speed. In reading signals, the CCD is very much like a two-dimension analog shift register in using the scanning method to read out sensed signals level by level. With the increase in the number of pixels, the system would require much more time in reading signals. Comparatively, a CMOS is much more flexible in that data can be transmitted from any addressed location. This greatly increases the convenience of the backend digital signal processing.

8.3.1.1 Image Sensor Circuits There are two types of image sensor circuits, passive pixel sensor and active pixel sensor, without and with amplifier circuits, respectively. Looking at its amplifier circuit designs, we can further break down the active pixel sensor into two charging modes: linear integration and logarithmic integration. There is also a sensor circuit utilizing photogates that is a little different from the photodiode we have been discussing. A discussion of each image sensor circuit follows.

From Figure 8.6, we see that the passive pixel sensor has a photodiode and a control switch (TX). When the TX is on, the light signals sensed by the photodiode generate charge signals, which in turn become currents passing to

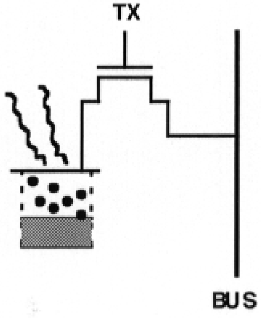


Figure 8.6 Circuit of passive pixel sensor.

the bus. Signals are read with the amplifier circuit, and once that is completed, the photodiode potential is reset to match that on the bus. Because the circuits are simple and the space required for each pixel is limited, the fill factor and quantum efficiency of the passive pixel sensor are very high. However, it is such a simple circuit that it is sensitive to noise. With noise being very large, it is hard for the reading speed to increase, and this limitation is what hinders an image sensor with the passive circuit structure to improve.

To improve the passive circuit structure, an active amplifier is added in the design of the pixel sensor, thus making the circuit an active pixel sensor (APS). Through the amplifier, noises have much smaller impacts on the sensor. With the addition of an active amplifier within each pixel, each unit's circuitry structure becomes more complicated and takes up more space. Consequently, the fill factor and quantum efficiency of an active pixel sensor are much lower. Still, in comparing the overall efficiency of the two systems, the APS is a much better design.

Figure 8.7 shows the sensor circuit in the linear integration mode. In the circuit M1 acts as a source–follower amplifier, with the ability to have the ratio of input/output voltage near 1 : 1, which precisely reflects the voltage value generated from sensing. When the circuit is in action, the reset signal resets the entire circuit. After the reset signal changes from on to off, the process of outputting voltage from the source follower is exhibited by the formula

$$\frac{dV}{dt} = \frac{\eta P_{\mathrm{in}} \lambda}{1.24C} \times G \tag{8.1}$$

where V represents the output voltage from the source–follower, η is the quantum efficiency, P_{in} is the energy of the incident light, λ is the wavelength of the incident light, C is the speed of light in air, and G is the source–follower gain. After a certain integration time, the row select is turned on to output sensed signals onto the data bus to complete the sending of signals. From Eq.

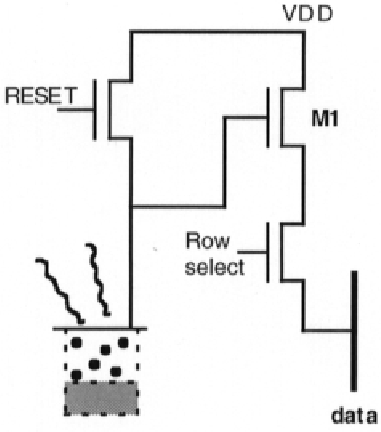


Figure 8.7 Circuit of active pixel sensor: linear integration mode [9].

(8.1), it is clear that the output voltage of this circuit is directly proportional to the integration time and the light intensity and not the sensing area. Therefore this type of circuit has a better fill factor and higher signal-to-noise ratio (SNR). The only disadvantage of this circuit is the integration time, which causes the response speed to improve.

An active pixel sensor in the logarithmic mode is shown in Figure 8.8. The following formula indicates the output voltage of the data bus [9]:

$$V = V_{\mathrm{dd}} - \frac{kT}{q} \ln\left(\frac{I_{\mathrm{ph}}}{I_0}\right) \tag{8.2}$$

where I_{ph} is the sensed photo current, I_0 is a constant, q is the electron energy,

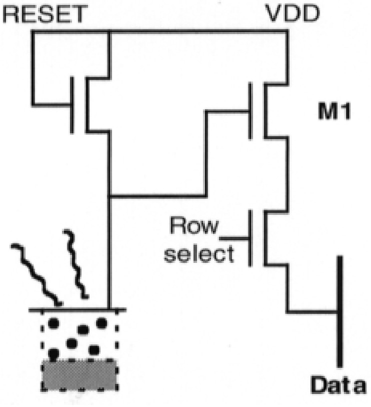


Figure 8.8 Circuit of active pixel sensor: logarithmic mode [9].

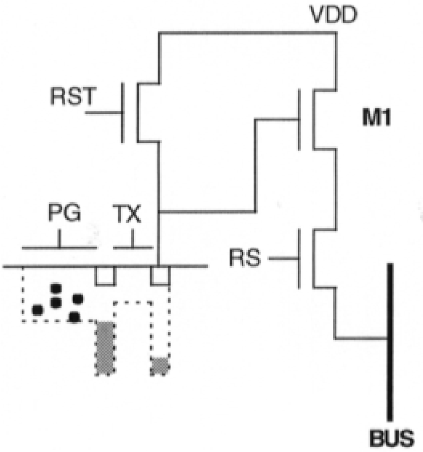


Figure 8.9 Circuit of photogate: type APS [10].

k is Planck's constant, T is the absolute temperature, and V_{dd} is the supply voltage. The advantage of this circuit is not involving time integration during the voltage output process, which makes it operate faster. The disadvantage of using this circuit is that when the intensity of incident light is low, the swing of the output voltage decreases. When that happens, the output voltage is susceptible to noise distortion and consequently causes the SNR to be lower.

In addition to using the CMOS photodiode to realize the sensor circuits discussed above, it can also be used with photogates. The photogate-type APS using linear integration is a circuit structure developed at the Jet Propulsion Laboratory (JPL), shown in Figure 8.9. This circuit structure has the characteristics of the CCD where it stores sensed charges underneath the photogates (PGs). When signals are ready to be read, the TX will turn on. Charges then flow to the other floating diffusion node of the TX to be amplified by the source–follower and outputted. If the differential amplifier based on the correlated double-sampling method was adapted here, we would be able to lower the reset noise, $1/f$ noise, and fixed-pattern noise caused by the changes in the threshold voltage. With a photodiode and five transistors, the fill factor and quantum efficiency of this circuit structure are not as good as those of the photodiode. However, the photogate circuit is able to reduce the noise effects to minimal, which a photodiode cannot do.

8.3.1.2 A Read-Out Circuit After the sensor circuit sends the signals, a read out circuit is required to readout these signals. An active pixel sensor, being most frequently used, is very different from the CCD sensor in that there is an amplifier in each sensor unit. When we consider the array of the entire image sensor units, the gains and offsets of the amplifiers are different. This

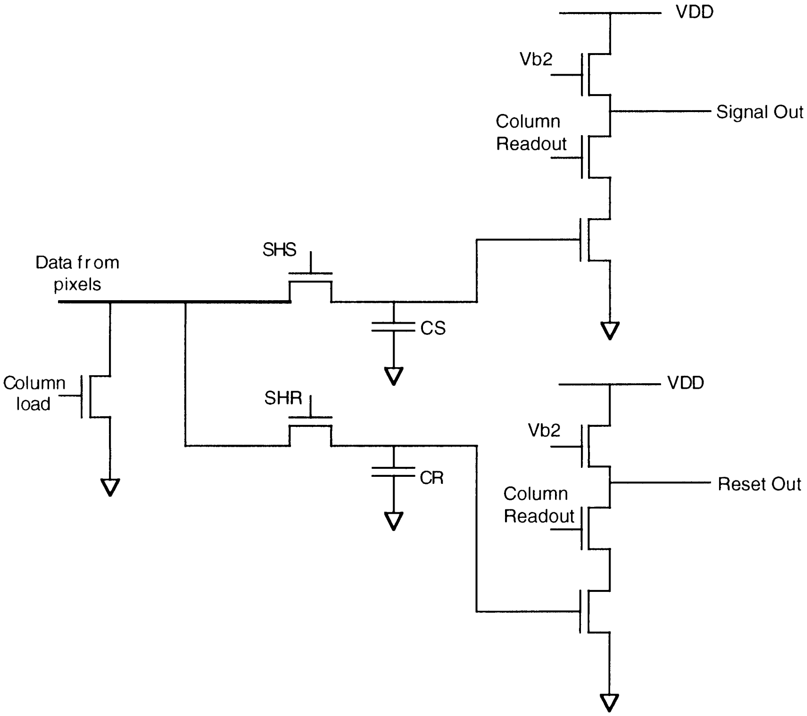


Figure 8.10 Correlated double-sampling circuit [10].

difference results in the largest noise in the APS circuit, the source of fixed pattern noise (FPN). To reduce FPN, it is necessary to add some circuits, such as correlated double sampling (CDS), as shown in Figure 8.10. Output from each column would receive two sample-and-hold capacitance values, one to capture the signal swing and the other to capture the reset signal swing. The difference between the two swing values is the adequate voltage sensed by each sensor circuit, with most of the offset impact from the pixel amplifier being eliminated in the process. This design greatly reduces FPN, from 5 to 0.1%. In addition, there is another method that stores the offset swing of each column. When signals are selected, the circuit takes out the offset swing before outputting the voltage signal. This type of design allows the APS to sense images without the impact of light and broadens its usage in many areas; however, it is much more complicated in circuit design and hardware realization.

Figure 8.11 shows the timing diagram of each signal from Figure 8.10. When the sample-and-hold signal (SHS) is on, the sensed signal is stored on the CS capacitor and then resets the pixel. When the sample-and-hold reset (SHR) is on, its reset signals are stored on the CR capacitor. After a series of this action, the column read-out signal starts to read out the signals in two sets. One is the

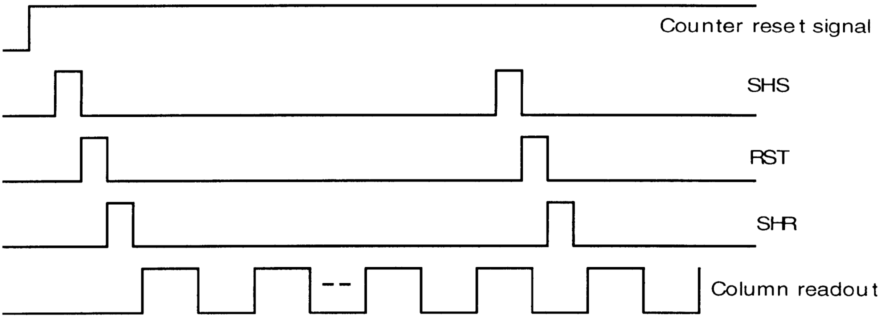


Figure 8.11 Timing of signals in read-out operation.

reset signals and the other is the sensor signals. The backend processing circuit takes the difference between these two signal values to be the output signal. This method also efficiently reduces FPN.

8.3.2 Optical Interconnection Systems

With the development of integrated circuits and advancement in manufacture technologies, transistors are required to have smaller and smaller channel lengths, but at the same time there are more and more transistors that need to fit on a chip. From small- to *large-* to very large to ultralow-scale integration (SSI, LSI, VLSI, ULSI), the number of integrated transistors grows from hundreds to thousands to millions. In the process, integrated circuits have become more functional, more powerful, and faster, which means the connection between chips is high speed and more complicated, thus making the interconnection design between chips ever more important. In a conventional circuit system, electric signals are used as the connection means. Even though it is better in signal amplification, processing, and calculation, the electric connection cannot provide a wider frequency band and reduce noises more efficiently. Most importantly, with the connection between chips becoming high speed, the complexity of using electricity as connection means between chips is increasing exponentially. This makes using the photoelectronic technology to design the connection means between chips a better solution. Advantages of using light as connection means are (1) no crosstalk problems, (2) broader frequency bandwidth, (3) better immunity against noises, (4) higher transmission speed, and (5) better ability against intersymbol interference (ISI). Thus, we see that in high-speed operations the optical interconnection system is simpler in structure and more efficient in operation than the conventional electrical interconnection system.

In recent years, optical interconnections have been successfully implemented by using various technologies such BiCMOS, GaAs, and CMOS [11–14]. Especially in the CMOS technology, it has advantages of low-power dissipation, high integration, low cost, and easy access. In addition, most consumer, computer, and communication electronics are realized by this technology. Since the CMOS microprocessors and digital signal processors can reach a clock frequency of 1 GHz, high-speed interconnections among chips become crucial. Hence, optical interconnections instead of electric connections using the printed circuit board are a good approach. In doing so, using the CMOS technology to realize optical interconnections can be a good choice to integrate the other application circuits [15–17].

For a chip-to-chip interconnection system to work, it is crucial to accurately receive and transmit signals. How to effectively use the photoelectronic technology to design the receiver and transmitter then becomes the key issue. A receiver accurately senses and captures light signals from the transmission terminal and converts them into electric signals for calculation and processing by local chips, while a transmitter converts these output signals from local chips back into light signals before transmitting them. In a circuit design, a good receiver is determined by its sensitivity in capturing light signals and its amplification rate, while a good transmitter must have enough power to activate the laser diode to generate light signals. It is important to factor in these points when evaluating the efficiency and power dissipation of an optical interconnection system.

8.3.2.1 Receiver Circuits In designing a receiver, there are two types of circuit diagrams, as shown in Figure 8.12. Figure 8.12*a* is the most basic photoreceiver circuit design. The theory is to only use an inverter and a loading transistor to directly process signals received from a photodiode. The photoreceiver directly senses photocurrents so that output signals are greatly affected

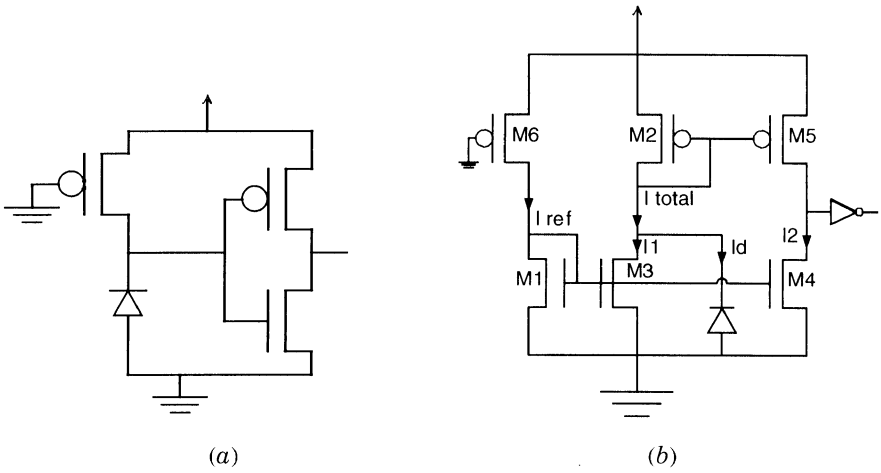


Figure 8.12 Circuit of (*a*) basic photoreceiver and (*b*) improved photoreceiver [18].

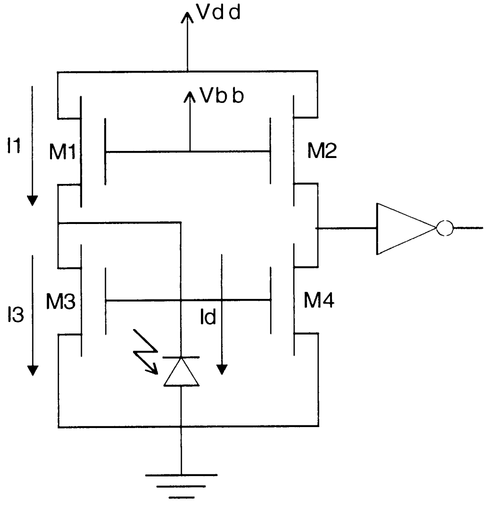


Figure 8.13 Circuit of proposed photoreceiver.

by noise. Although this circuit is very simple, it is not proficient enough in the sensing ability. Figure 8.12*b* shows a better circuit design in which two pairs of current mirrors and a current-to-voltage conversion circuit are utilized to sense photocurrents [18]. This is an indirect way to sense and amplify photocurrents; therefore it has better results in light-sensing ability and against noise disturbance.

The photoreceiver shown in Figure 8.13 improves upon what was shown in Figure 8.12. All circuits include a photodiode, a current mirror of M3 and M4, and two active resistors of M1 and M2 biased by a controllable voltage. When no incident light is present, the photodiode is shut off and the entire circuit is in a stable condition. Currents from M2 will all flow to M4 via an inverter to have an output of logic 1. When incident light is present, the photodiode is turned on. Currents produced by the photodiode, due to the current mirrors, increase and flow through to M2, then to M4. When M4 cannot handle the excess currents, the voltage of the drain node of M4 is accumulated, so when currents go through the inverter, the output becomes logic 0. This design not only retains the advantages of the one shown in Figure 8.12*b*, it also uses fewer transistors, making the chip size smaller than the sensor circuit structure shown in Figure 8.12*b*. The substantial difference between this circuit design and the one shown in Figure 8.12*b* is its control on the operation of the CMOS photodiode. The CMOS photodiode's sensing ability depends upon its fabrication process and the amount of reverse-biased voltage on a photodiode. Under the stable fabrication process, the photoreceiver circuit proposed in this study is able to control the amount of V_{bb} and therefore indirectly controls the amount of reverse-biased voltage of the photodiode. This also allows us to use

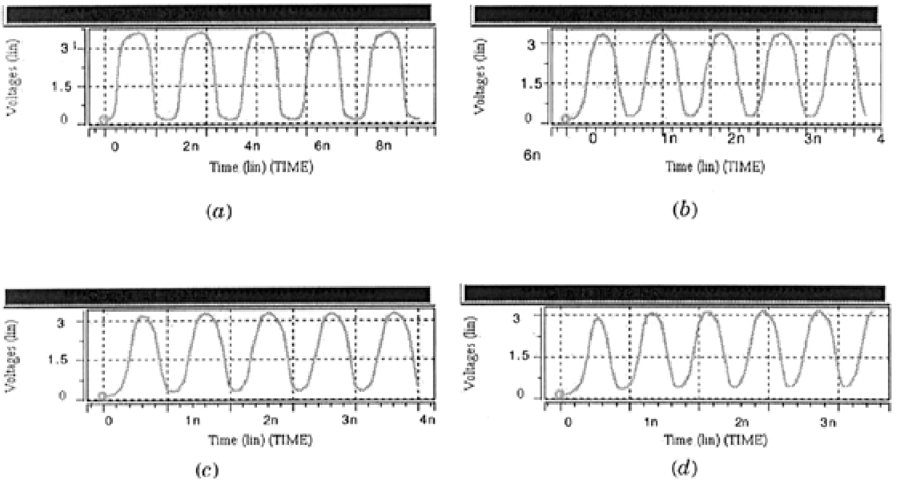


Figure 8.14 Simulated output signals using proposed photoreceiver: (*a*) 400 MHz; (*b*) 800 MHz; (*c*) 1 GHz; (*d*) 1.25 GHz.

the sensing ability of the CMOS photodiode to find the optimized performance for sensing light.

Using the Taiwan Semiconductor Manufacturing Company (TSMC) 0.35-μm CMOS technology, Figure 8.14 shows the simulation results of the proposed photoreceiver in the operating frequencies of 400 MHz, 800 MHz, 1 GHz, and 1.25 GHz where the photodiode is assumed to provide 100 μA and response up to 2.5-GHz signal rates. This photoreceiver can reach an operation frequency of more than 1 GHz. Its power consumption is around 3.8 mW at 1 GHz. Table 8.2 compares our photoreceiver with photodetectors, as shown in Figure 8.12. The proposed photoreceiver is superior to conventional photodetectors.

8.3.2.2 Transmitter Circuits As mentioned earlier in this study, a transmitter requires enough driving power to activate the laser diode to generate

TABLE 8.2 Comparison of Three Photoreceivers

Circuit Diagrams	Maximum Frequencies	Power Supplies (V)	Fabrication Processes
Fig. 8.12*a*	130 MHz	5	0.8-μm CMOS
Fig. 8.12*b*	330 MHz	5	0.8-μm CMOS
Fig. 8.13	1.25 GHz	3.3	0.35-μm CMOS

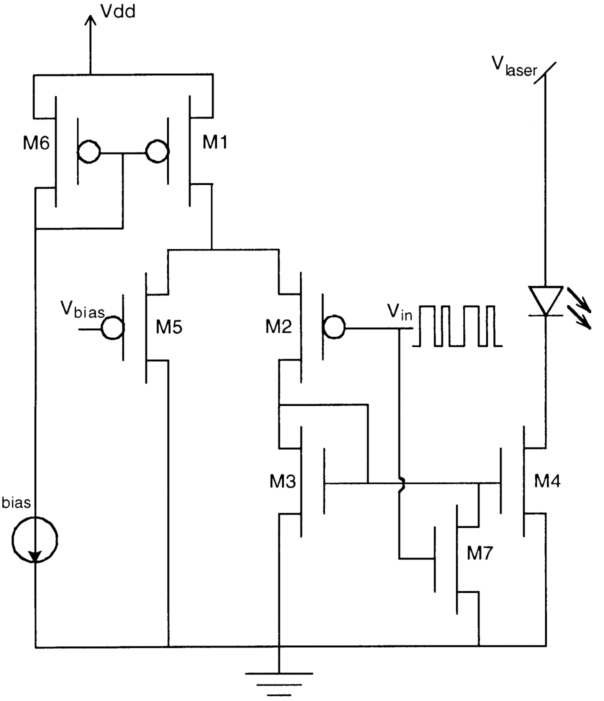


Figure 8.15 Circuit of laser diode driver [19].

light signals. In saying that, the design of the driving circuit needs to consider not only operation efficiency and power consumption but also a circuit that operates under high frequency and has the optimized electron-to-photon conversion capability. Figure 8.15 shows a driving circuit design of a laser diode [19]. The theory behind this circuit is to use a differential pair and two current mirrors to generate the driving currents for a laser diode. This type of circuit design requires many transistors, which means its power consumption is greater. In addition, while the laser diode is not activated, its driving circuit still consumes power.

To improve upon the disadvantage of high power consumption of the circuit design, shown in Figure 8.15, a low-power consumption circuit is proposed, as shown in Figure 8.16. In this circuit design, M1 and M2 form a current mirror. By adjusting transistor sizes of M1 and M2, enough driving currents are produced for the laser diode. The transistor M3 then is used to pass through the driving current to modulate the laser diode for generating light signals. This is done by inputting different control voltages on the gate of M3 to switch the driving current on or off. Looking from the circuit design perspective, the existence of M4 does not make any difference in driving a laser diode. From the power consumption perspective, without M4, M1 would generate currents

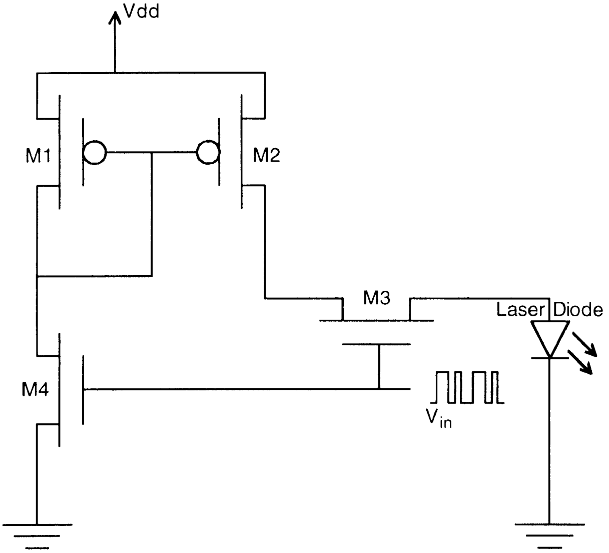


Figure 8.16 Circuit of low-power laser diode driver.

all the way to ground. Therefore the transistor M4 is put in place where currents flow en route to ground and uses the input signal to turn it on or off. This way, when the input signal is low, M4 is turned off. With no currents flowing to ground, power consumption is greatly reduced.

In addition, to ensure that the optical interconnection for chip-to-chip communication could work with any type of light source, a programmable switch is added to the laser diode driver circuit so that it can adjust the amount of driving current outputted. Figure 8.17 shows the transistors of, for example, M2-1, M2-2, and M2-3 that are p-channel metal-oxide-semiconductor (PMOS) combined together with different transistor sizes, and they all form current mirrors with M1. Different transistor sizes produce different volumes of currents. With Mgl, Mg2, and so on, playing the on/off roles, voltages from Vgl, Vg2, and so on, act as switches to generate different values of output currents. This method allows us to easily adjust the driving currents to meet various laser diodes and to overcome distortion effects from laser diode packaging in the optical interconnection system design.

Figure 8.18 shows output signals of the proposed laser diode driver modulated by input signals at clock frequencies of 200 MHz, 500 MHz, 800 MHz, and 1 GHz. Where the VCSEL laser diode can generate 850 nm at a threshold current of 0.6 mA, the proposed laser diode driver can correctly operate at 1 GHz. According to the simulation results of our laser diode driver using the TSMC 0.35-μm CMOS technology, the power consumption at different operating frequencies is listed in Table 8.3 where the supply voltage is 3.3 V. The maximum power consumption that occurred at 1 GHz is around 5.0 mW.

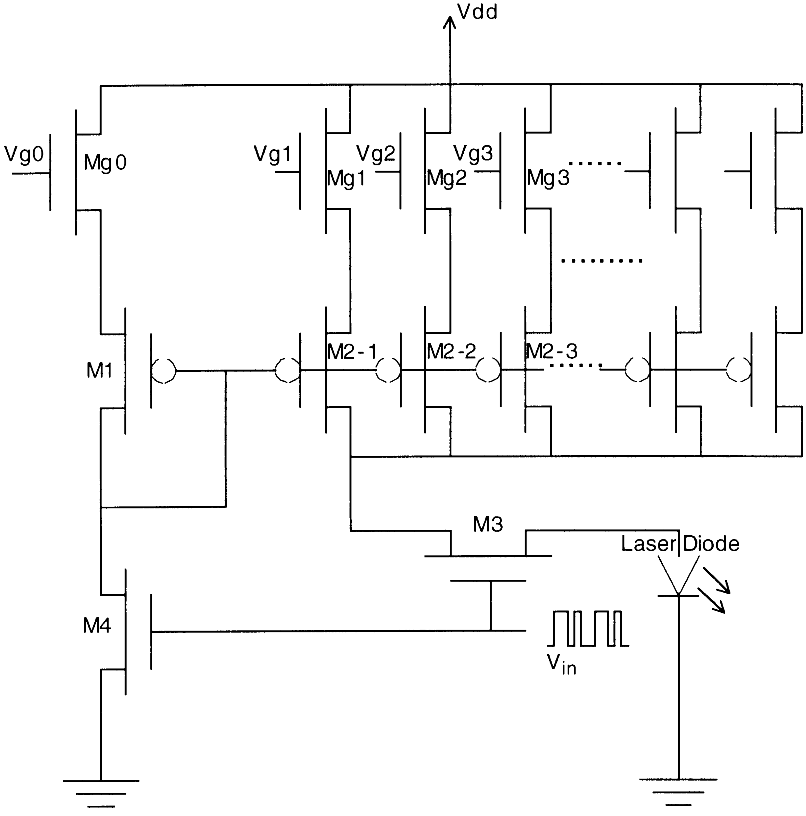


Figure 8.17 Circuit of laser diode driver with adjustable driving currents

8.3.2.3 Optical Chip-to-Chip Interconnection As shown in Figure 8.19*a*, optical communication between two chips is illustrated. Each chip includes photodiodes, photoreceivers, laser diodes, and laser diode drivers. In one chip, the laser diode driver modulated by output signals provides currents to drive the laser diode to emit light. Through free-space or wave-guided light transmission, the photodiode of the other chip induces a current that is sensed and amplified by the receiver circuit. Using the 2 × 2 formation shown in Figure 8.19*a*, four receivers and transmitters are proportionally placed on one chip, and the same is done on the other chip, except the positions of the receivers and transmitters are switched. When this set of chips is connected, the receivers can be directly positioned against the transmitters, thus completing communication of the optical interconnection system. A prototype chip including four receivers and transmitters, as shown in Figure 8.19*b*, was implemented by using the TSMC 0.35-μm CMOS technology. Its die size is 1.8 × 1.5 mm. Internal pads of 250 × 250 μm are designed for flip-chip packaging of laser diodes [20]. In addition, CMOS photodiodes with ring-type electrodes as shown in Section 8.2 were implemented.

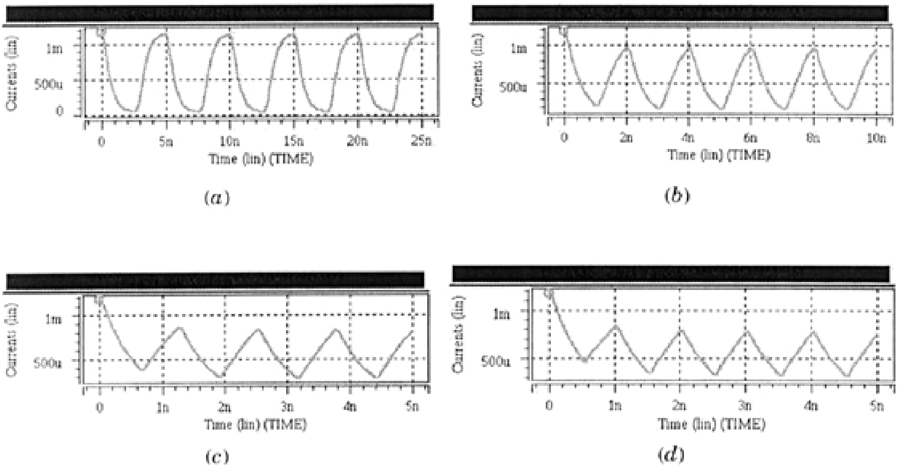


Figure 8.18 Simulated driving currents of proposed laser diode driver: (*a*) 200 MHz; (*b*) 500 MHz; (*c*) 800 MHz; (*d*) 1 GHz.

8.4 MEASUREMENT RESULTS OF PHOTOELECTRONIC COMPONENTS

In the following sections photoelectronic components are compared and the physical characteristics of each are measured and discussed.

8.4.1 Physical Characteristics of CMOS Photodiodes

From the fabrication process of the TSMC 1P3M N-well 0.6-μm CMOS technology, three photodiodes of different P–N contact surfaces are available: N-well P substrate, N-diffusion P-well P substrate, and P-diffusion N-well P substrate.

TABLE 8.3 Power Consumption of Laser Diode Driver at Different Operating Frequencies

Operating Frequencies	Power Consumption (mW)
100 MHz	4.22
200 MHz	4.32
500 MHz	4.40
800 MHz	4.68
1 GHz	5.01

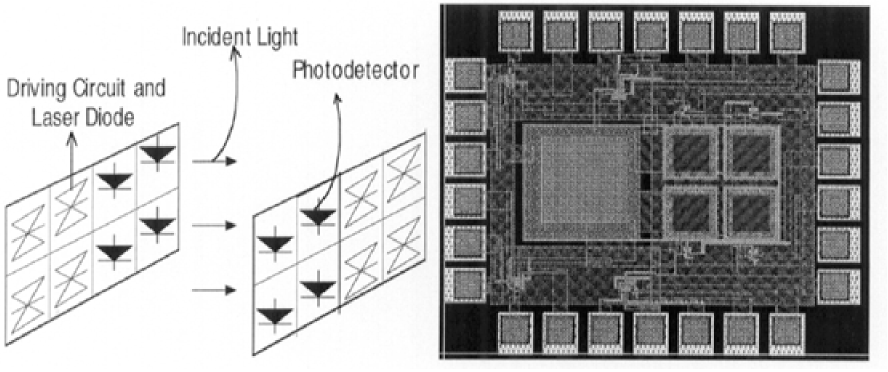


Figure 8.19 (*a*) Optical communication system and (*b*) prototype chip for chip-to-chip interconnection.

8.4.1.1 N-Well P-Substrate Photodiode The N-well P-substrate photodiode is shown in Figure 8.20. In this type of photodiode, the positive node is the P substrate and its negative node is the N well. Because the ion-doping concentration of the N well is a bit higher than that of P substrate, the depletion region formed on the contact surface has a larger area in the P substrate to allow the P-substrate contact window to be within the depletion region to successfully induce currents. In reality, to prevent too much leakage current, we enclose the entire N well with the P-substrate contact window, as shown in Figure 8.20. Because the backend signal processing circuit and photodiode both use the same P substrate, this procedure is done to prevent free carriers from drifting into the backend circuit, which then causes leakage currents. This type of photodiode is more suitable for incident light with longer

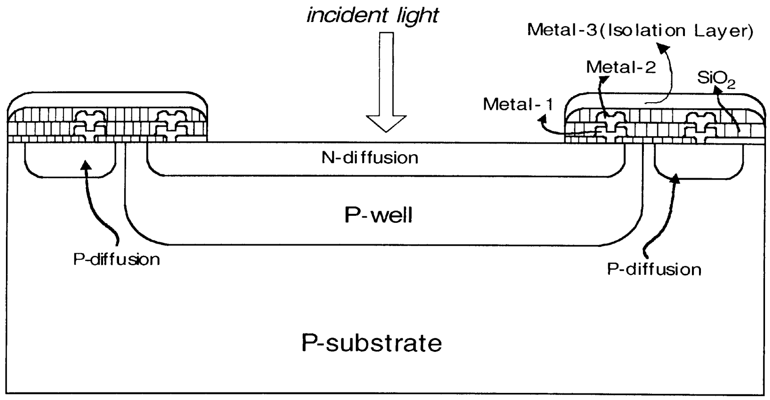


Figure 8.20 N-well P-substrate photodiode.

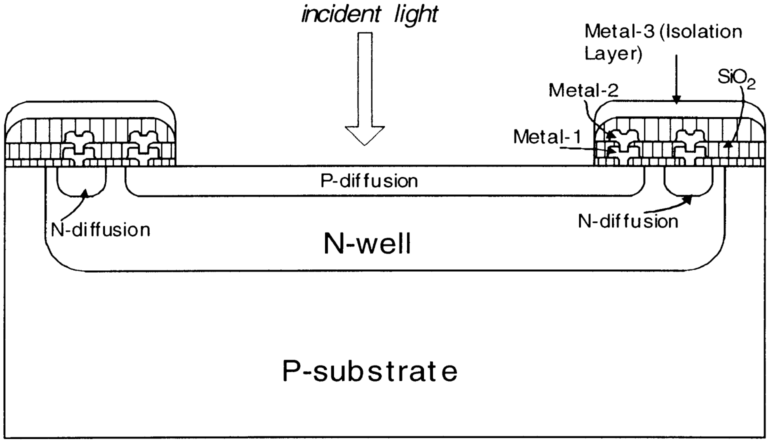


Figure 8.21 N-diffusion P-well P-substrate photodiode.

wavelength due to its deeper doping depth in the N well. Only light with longer wavelength, such as infrared, would be able to reach this contact surface.

8.4.1.2 N-Diffusion P-Well P-Substrate Photodiode Figure 8.21 shows the photodiode with positive and negative nodes using the P substrate and N diffusion, respectively. This photodiode is more suitable for incident light with shorter wavelength. With the doping depth being shallower in the N-diffusion region, lights with shorter wavelength, such as visible light in an image sensor system, can be mostly absorbed on this contact surface.

8.4.1.3 P-Diffusion N-Well P-Substrate Photodiode Figure 8.22 shows a P-diffusion N-well P-substrate photodiode. In this type of photodiode, the positive and negative nodes are P diffusion and N well, respectively. Because the doping concentration of the P diffusion is higher than that of the N well, the depletion region has a larger area in the N well. However, this photodiode has the inevitable problem with leakage currents. This leakage problem is mainly from the N-well P-substrate contact surface below. Whenever incident light projects on the P-type doping region, some light will penetrate through to the N-well P-substrate contact surface and create another photodiode on this contact surface. When this happens, leakage currents become a serious problem, and thus this photodiode is not recommended in this study.

8.4.2 Measurements of CMOS Photodiodes

Figure 8.23 show the measured $I-V$ curves of the N-well P-substrate, P-diffusion N-well P-substrate, and N-diffusion P-well P-substrate CMOS photodiodes. From these graphs, it is evident that currents under reverse-

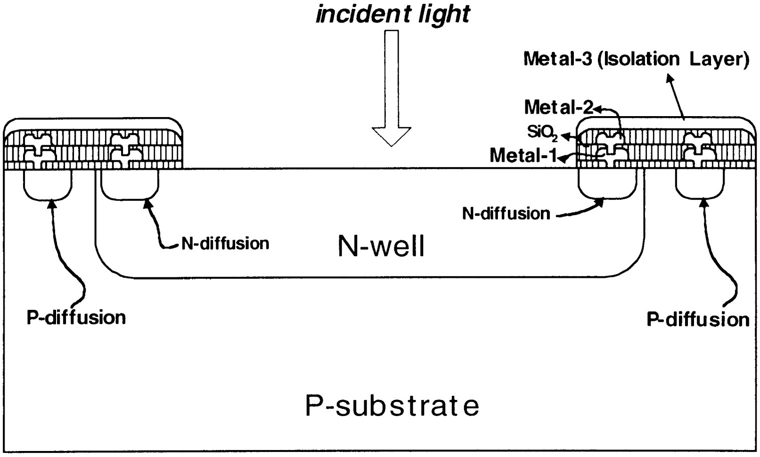


Figure 8.22 P-diffusion N-well P-substrate photodiode.

biased voltage are clearly smaller than those under forward-biased voltage. This is consistent with the principle behind the photodiode.

Figure 8.24 shows the measured dark currents of these three photodiodes at the same area. From Figures 8.24*a* and *b*, the P-diffusion N-well P-substrate photodiode has the most profound dark current problem for reasons explained earlier. The dark current of the N-diffusion P-well P-substrate photodiode is larger than that of the N-well P-substrate photodiode due to process parameters and the spreading effects of different densities between the P well and P substrate. In measuring the response to light, Figure 8.25 shows that the measured currents of N-diffusion P-well P-substrate and N-well P-substrate photodiodes are little enlarged with increasing the reversed-biased voltages at different light wavelengths. Figure 8.26 shows the photoresponses of the

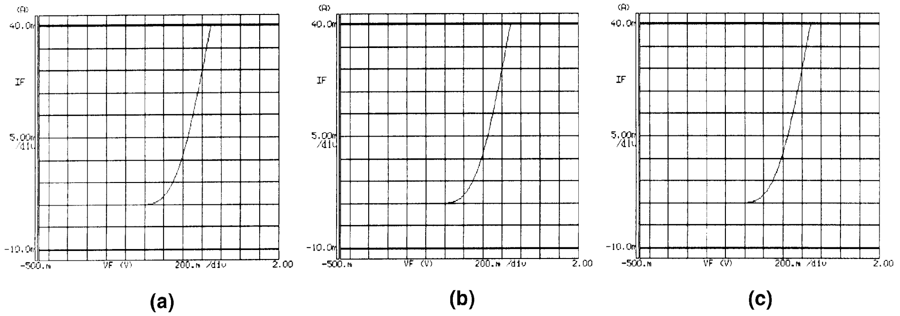


Figure 8.23 Measured I–V curves: (*a*) N-well P substrate; (*b*) N-diffusion P-well P substrate; (*c*) P-diffusion N-well P substrate.

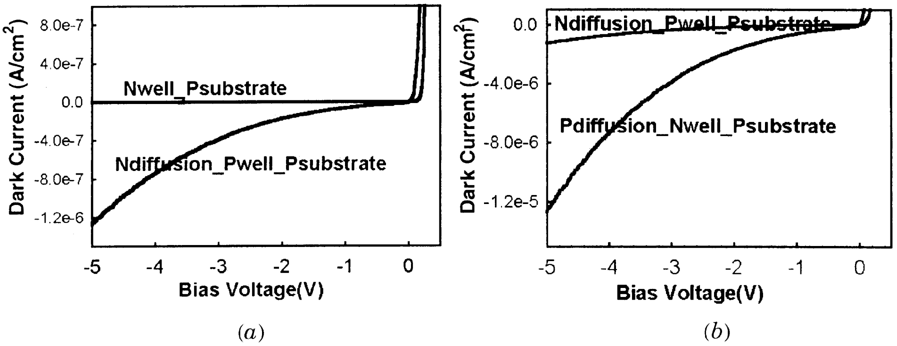


Figure 8.24 Measured dark currents of three photodiodes. (*a*) N-well P substrate versus N-diffusion P-well P substrate; (*b*) N-diffusion P-well P substrate versus P-diffusion N-well P substrate.

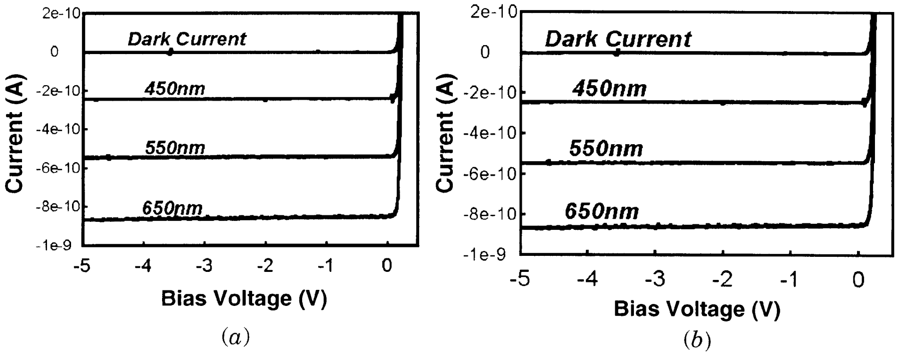


Figure 8.25 Measured currents at different reverse-biased voltages: (*a*) N-diffusion P-well P substrate; (*b*) N-well P substrate.

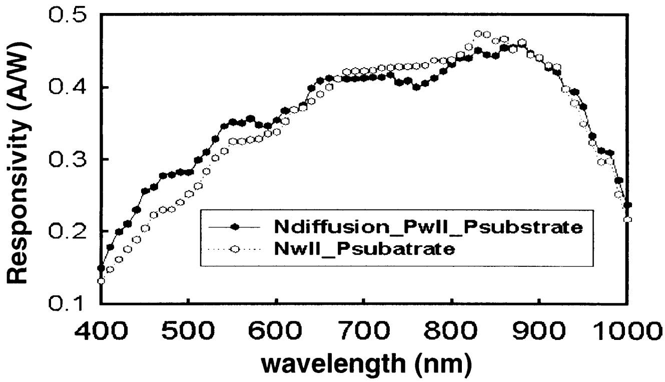


Figure 8.26 Photoresponses of N-diffusion P-well P-substrate and N-well P-substrate photodiodes.

N-diffusion P-well P-substrate and N-well P-substrate photodiodes for light wavelengths from 400 to 1000 nm. When the wavelength of incident light is smaller than 670 nm, the N-diffusion P-well P-substrate photodiode has a better response than the N-well P-substrate photodiode. This effect demonstrates the fact that the N-diffusion region with a shallower doping depth than that of N-well is suitable for sensing light with shorter wavelengths.

REFERENCES

1. J. Wilson and J. Hawkes, *Optoelectronics*, Prentice-Hall, Upper Saddle River, New Jersey, 1998.
2. S. Sze, *Physics of Semiconductor Devices*, John Wiley & Sons, New York, 1981.
3. W. Liu, R. Sheen, J. Hwang, and O. T.-C. Chen, "A low-power and high-frequency CMOS transceiver for chip-to-chip interconnection," *Proceedings of IEEE International Symposium on Circuits and Systems* **3**, 1–4 (May 2000).
4. E. Fossum, "CMOS image sensors: Electronic camera-on-a-chip," *IEEE Trans. Electron Dev.* **44**(10), 1689–1698 (1997).
5. L. Coldren and S. Corzine, *Diode Lasers and Photonic Integrated Circuits*, John Wiley & Sons, New York, 1995.
6. E. Fossum, "Digital camera system on a chip," *IEEE Micro.* **183** 8–15 (1998).
7. M. Loinaz, K. Singh, A. Blanksby, D. Inglis, K. Azadet, and B. Ackland, "A 200mW 3.3V CMOS color camera IC producing 352 × 288 24b video at 30 frames/s," *Dig. IEEE Int. Solid-State Circuits Conf.* 168–169 (Feb. 1998).
8. B. Mansoorian, H. Yee, S. Huang, and E. Fossum, "A 250mW 60 frames/s 1280 × 720 pixel 9b CMOS digital image sensor," *Dig. IEEE Int. Solid-State Circuits Conf.* 312–313 (Feb. 1999).
9. N. Tu, R. Hornesy, and S. G. Ingram, "CMOS active pixel image sensor with combined linear and logarithmic mode operation," *Proc. IEEE Can. Conf. Electrical Comput. Eng.* **2**, 754–757 (1998).
10. S. Mendis, S. Kemeny, R. Gee, B. Pain, C. Staller, Q. Kim, and E. Fossum, "CMOS active pixel image sensors for highly integrated imaging systems," *IEEE J. Solid-State Circuits* **32**(2), 187–196 (1997).
11. J. Choi, B. J. Sheu, and O. T.-C. Chen, "A monolithic GaAs receiver for optical interconnect systems," *IEEE J. Solid-State Circuits* **29**(3), 328–331 (1994).
12. C.-S. Li, "Dense optical interconnects for high speed digital systems," Ph.D. Dissertation, University of California at Berkely (1991).
13. M. S. Elrabaa, M. I. Elmasry, and D. S. Malhi, "A universal 3.3V 1GHz BiCMOS transceiver(driver/receiver)," *Proc. Bipolar/BiCMOS Circuits Technol. Meeting*, 118–120 (1995).
14. H. Zimmermann, T. Heide, and A. Ghazi, "Monolithic high-speed CMOS-photoreceiver," *IEEE Photon. Technol. Lett.* **11**(2), 254–256 (1999).

15. R. Sheen and O. T.-C. Chen, “A 3.3V 600 MHz–1.30 GHz CMOS phase-locked loop for clock synchronization of optical interconnects,” *Proc. IEEE Int. Symp. Circuits Syst.* **4** 429–432 (1998).
16. U. Hilleringmann and K. Goser, “Optoelectronic system integration on silicon: waveguides, photodetectors, and VLSI CMOS circuits on one chip,” *IEEE Trans. Electron Dev.* **42**(5), 841–846 (1995).
17. T. Woodward and A. Krishnamoorthy, “1-Gb/s integrated optical detectors and receivers in commercial CMOS technologies,” *IEEE J. Selected Top. Quant. Electron.* **5**(2), 146–156, (1999).
18. A. Gadiri, Y. Savarin, and B. Kaminska, “An optimized CMOS compatible photoreceiver,” *Proc. IEEE Can. Conf. Electrical Comput. Eng.* **1**, 211–214 (1995).
19. L. Chen, M. Li, C. Chang-Hasnain, and K. Lau, “A low-power 1-Gb/s CMOS laser driver for a zero-bias modulated optical transmitter,” *IEEE Photon. Technol. Lett.* **9**(7), 997–999 (1997).
20. R. Pu, C. Duan, and C. Wilmsen, “Hybrid integration of VCSEL’s to CMOS integrated circuits,” *IEEE J. Select. Topics Quant. Electron.* **5**(2), 201–208 (1999).

Three-Dimensional Holographic Imaging. Edited by Chung J. Kuo, Meng Hua Tsai
Copyright © 2002 John Wiley & Sons, Inc.
ISBNs: 0-471-35894-0 (Hardback); 0-471-22454-5 (Electronic)

CHAPTER 9

Design and Implementation of Computer-Generated Hologram and Diffractive Optical Element

NI Y. CHANG

Department of Electrical Engineering
National Chung Cheng University
Chia-Yi, 62107 Taiwan

CHUNG J. KUO

Graduate Institute of Communication Engineering and
Department of Electrical Engineering
National Chung Cheng University
Chia-Yi, 62107 Taiwan

A conventional hologram is obtained by interfering two different light beams and recording this information through high-resolution photographic film. However, recent applications call for the making of hologram through the semiconductor fabrication process with its surface profile designed by a computer program. This is the so-called computer-generated hologram (CGH). The difference between a CGH and a conventional hologram is that the interference pattern at a hologram plane is calculated in the CGH while conventional hologram physically records the pattern. Therefore, a conventional hologram can only reconstruct a three-dimensional (3D) object that exists while the CGH can reconstruct a 3D object that never exists.

The name CGH sometimes makes people believe that it is only for 3D holographic applications. Indeed, the concept can be used to design an optical element to diffract any input light beam [1–3] to any desired position. In this case, a more appropriate name would be diffractive optical element (DOE) instead of CGH.

The CGH/DOE has been applied to many applications such as 3D display, microlenses, filters, and others. There are four issues in creating a CGH/DOE.

First the propagation of the complex amplitude (which is from the object plane to the hologram plane or from the hologram plane to any output position of the optical system) is difficult to compute. Because of limitations in computation, we can only compute the amplitude at a finite number of sampling points, although this constraint is quite tolerable because the amplitude is a band-limited function, as demonstrated elsewhere [4].

The second issue is the mathematical calculation of the light propagation within the computer. By the scalar diffraction theory, we can simulate the light propagation by taking the Fourier or Fresnel transform of the optical field. Discrete Fourier or Fresnel transform is usually used to achieve this goal, although they require a large amount of computations.

The CGH/DOE is, usually, a thin sheet that provides a phase shift of the incoming light without creating any amplitude attenuation. That is, the transmittance function of the CGH/DOE is a purely phase function. It is thus impossible to design a CGH/DOE with the desired output pattern. An efficient algorithm must be developed to design the desired CGH/DOE to satisfy constraints in the input and output plane and approximate the desired output pattern as much as possible.

The final issue is to transfer the calculated light fields to a transparency. A mask is first generated through a plotting or printing operation that is constrained by the device [such as a pen plotter, laser printer, or electron-beam (e-beam) lithography machine] to achieve this goal. Among these devices, the e-beam lithography machine is the best choice so far for CGH/DOE fabrication. With the mask and a series of fabrication processes, a CGH/DOE is obtained.

This chapter discusses the design and implementation of the CGH/DOE and is divided into six sections. The first section explains the sampling issue in the CGH/DOE. Then, the way to calculate the transmittance function is shown in Section 9.2. Section 9.3 explains the kinoform, a special form of the conventional CGH. The design methods for CGH/DOE are then reviewed in Section 9.4. Section 9.5 shows the fabrication process being used for the CGH/DOE. Finally, brief conclusions are given in Section 9.6.

9.1 SAMPLING ISSUES IN CGH

The function of a CGH is created by an artificial optical wave front from a set of computed data that are an adaquate sampling of the complex wave front amplitude. If a very small point source is placed at each sampling point, then we can obtain the amplitude and phase of the desired output diffraction pattern. If the sampling frequency were high enough, then sampling effects can be ignored and the original wave front can be reproduced; however, large computational efforts are also required. To simplify the CGH design, we can sample the light field and compute the complex values at "necessary" sampling points [5–7].

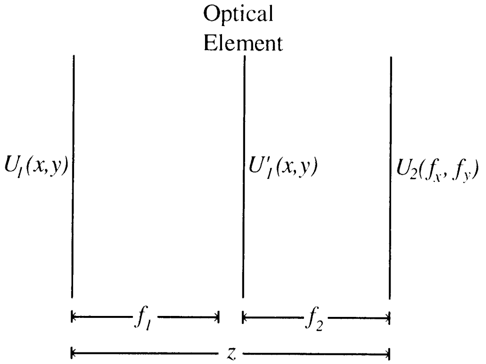


Figure 9.1 Optical architecture being considered for CGH/DOE.

So, how many sampling points of the field should be used? To answer this question, we consider two different situations demonstrated in Figure 9.1. The first situation employs the Fourier transform to model the propagation of the light field, while the second one uses the Fresnel transform.

9.1.1 Fourier Hologram

In the Fourier hologram, the desired output hologram field is the Fourier transform of the input object field, as demonstrated in Figure 9.1. Here, the optical element is a fictitious lens, with focal length f and infinite aperture size between the object and the hologram field. The two fields [input object field $U_1(x, y)$ and output hologram field $U_2(f_x, f_y)$] are located in the front and back focal planes of the lens, respectively, and thus $f_1 = f_2 = f$. According to the scalar diffraction theory, these two fields are related by a Fourier transform and expressed as

$$U_2(f_x, f_y) = \frac{1}{\lambda f} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} U_1(x, y) \exp\left[-j \frac{2\pi}{\lambda f} (x f_x + y f_y) \right] dx\, dy \quad (9.1)$$

According to the Whittaker–Shannon sampling theorem, the bandwidth and required sampling points of the output hologram field are determined by the size of object. Let $L_x \times L_y$ be the dimension of the object; then the spectrum of the field in the hologram plane is constrained within a centered rectangle with dimensions $2B_{f_x} \times 2B_{f_y}$, where

$$2B_{f_x} = \frac{L_x}{\lambda f} \qquad 2B_{f_y} = \frac{L_y}{\lambda f} \quad (9.2)$$

The rectangular sampling grid in the hologram plane thus has spacings

$$\Delta_{f_x} = \frac{1}{2B_{f_x}} = \frac{\lambda f}{L_x} \qquad \Delta_{f_y} = \frac{1}{2B_{f_y}} = \frac{\lambda f}{L_y} \quad (9.3)$$

If the extent of the field in the hologram plane is $L_{f_x} \times L_{f_y}$, then the number of samples required for that plane is

$$N_{f_x} = \frac{L_{f_x}}{\Delta_{f_x}} = \frac{L_{f_x} L_x}{\lambda f} \qquad N_{f_y} = \frac{L_{f_y}}{\Delta_{f_y}} = \frac{L_{f_y} L_y}{\lambda f} \tag{9.4}$$

Similarly, the number of samples required in the object plane is identical to that in the hologram field.

9.1.2 Fresnel Hologram

In the Fresnel hologram, the desired hologram field is the Fresnel transform of the object field. For the configuration shown in Figure 9.1, the transmittance function of the optical element is unitary in the Fresnel hologram or, equivalently, the optical element does not exist. Since there is no lens in the field formation, the hologram field is no longer related to the object field by a simple Fourier transform. Instead, the Fresnel diffraction integral is used to relate the hologram and the object field,

$$U_2(f_x, f_y) = \frac{1}{\lambda f} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \times U_1(x, y) \exp\left[j\frac{\pi}{\lambda z}(x^2 + y^2) \right] \exp\left[-j\frac{2\pi}{\lambda f}(x f_x + y f_y) \right] dx\, dy \tag{9.5}$$

Here, the relation between the bandwidth of the output hologram field and the size of the object is similar to the Fourier hologram. The object function is $U_1(x, y) \exp[j\pi(x^2 + y^2)/\lambda z]$. The intensity distribution $|U_2(f_x, f_y)]$, a quantity we wish to re-create from the output hologram field, is not affected by the presence of a phase distribution $\exp[j\pi(x^2 + y^2)]$ across the object. Therefore, Eq. 9.5 can be considered as the product of the quadratic-phase factor in (f_x, f_y) and the Fourier transform of the modified object field.

The bandwidth that arises from the Fourier transform of the modified object is identical to that in the Fourier hologram (because the quadratic-phase factor in the object plane does not change the object's width). We can use local spatial frequencies to approximate the bandwidth of the quadratic-phase factor with consideration of the finite extent of the output hologram field. Respectively, let $\pm L_{f_x}/2\lambda z$ and $\pm L_{f_y}/2\lambda z$ be the local frequencies in the f_x and f_y directions, respectively. The total bandwidth of the output hologram field is thus the summation result of these bandwidths and those obtained from the Fourier hologram. Hence, the total bandwidth is

$$2B_{f_x} = \frac{L_x + L_{f_x}}{\lambda z} \qquad 2B_{f_y} = \frac{L_y + L_{f_y}}{\lambda z} \tag{9.6}$$

The total bandwidth depends on the bandwidth from the input object and output hologram field. In the hologram plane, the sampling intervals become

$$\Delta_{f_x} = \frac{\lambda z}{L_x + L_{f_x}} \qquad \Delta_{f_y} = \frac{\lambda z}{L_y + L_{f_y}} \tag{9.7}$$

and in each dimension the total number of the samples is

$$N_{f_x} = \frac{L_{f_x}}{\Delta_{f_x}} = \frac{L_{f_x}(L_x + L_{f_x})}{\lambda z} \qquad N_{f_y} = \frac{L_{f_y}}{\Delta_{f_y}} = \frac{L_{f_y}(L_y + L_{f_y})}{\lambda z} \tag{9.8}$$

The number of samples required in the hologram plane are equal to that required in the object plane. Compared with the Fourier hologram, the number of samples required in the Fresnel hologram is larger than that required in the Fourier hologram.

9.2 COMPUTATIONAL ISSUES IN CGH

The relations between the object and holographic field (Eqs. 9.1 and 9.5) involve the Fourier transform. After sampling, the continuous integral in these equations becomes a discrete summation. Since $(\Delta f_x, \Delta f_y)$ and $(\Delta x, \Delta y)$ are the sampling spacings in the hologram and object planes, respectively, the Fourier transform hologram is expressed as

$$U_2(p\,\Delta f_x, q\,\Delta f_y) = \sum_{m=0}^{N_{f_x}-1} \sum_{n=0}^{N_{f_y}-1} U_1(m\,\Delta x, n\,\Delta y)\exp\left[j2\pi\left(\frac{pm}{N_{f_x}} + \frac{qn}{N_{f_y}}\right)\right] \tag{9.9}$$

This relation is the so-called *discrete Fourier transform* and is usually computed through the fast Fourier transform (FFT) algorithm. The most known FFT algorithm [8] only requires $N_{f_x} \times N_{f_y} \log_2(N_{f_x} \times N_{f_y})$ complex multiplications and additions when the number of data points to be calculated is $N_{f_x} \times N_{f_y}$, where N_{f_x}/N_{f_y} is a 2's power number. If the number of data points to be calculated is not 2's power, then a prime-factor FFT [8–10] should be used to reduce the unnecessary computations.

Fresnel and Fourier transforms are closely related. The only difference between these two transforms is the postmultiplication of the quadratic-phase term in the Fresnel transform. Mathematically, the discrete Fresnel transform is expressed as

$$\begin{aligned} U_2(p\,\Delta f_x, q\,\Delta f_y) = \sum_{m=0}^{N_{f_x}-1} \sum_{n=0}^{N_{f_y}-1} U_1(m\,\Delta x, n\,\Delta y)\exp\left\{j\pi\left[\left(\frac{m}{N_{f_x}}\right)^2 + \left(\frac{n}{N_{f_y}}\right)^2\right]\right\} \\ \times \exp\left[j2\pi\left(\frac{pm}{N_{f_x}} + \frac{qn}{N_{f_y}}\right)\right] \end{aligned} \tag{9.10}$$

Apparently, the FFT can be used to calculate the Fresnel transform.

9.3 A SPECIAL FORM OF CGH: KINOFORM

In the CGH/DOE, the complex field existing in the hologram plane should be calculated first. Then an optical element (CGH/DOE) to provide such a complex field (when its input field is known) is fabricated. In practice, it is difficult to fabricate an optical element to provide such a complex field. To simplify the problem, a different method for CGH representation — the kinoform — is proposed. Here we assume the phase of the Fourier coefficient carries most information of a 3D object, while the amplitude information can be ignored. Since the observer is far away from the CGH/DOE, Fourier transform can be used to calculate the field at the observer plane. Therefore, the object can be reconstructed. Although the assumption about ignoring the object's amplitude spectrum appears surprising, it turns out to be accurate if the object is a diffuse one. That is, all the object points have random and independent phase.

A kinoform is divided into $N \times N$ cells, each representing one Fourier coefficient of a 3D object. The amplitude of the Fourier coefficient is set to zero and only the phase is encoded into a large-scale transparency. The encoding process is achieved by linearly mapping the phase value (between 0 and 2π) of the Fourier coefficient into a gray level on a transparency by an output device such as a plotter or printer. Usually, an output device with more gray-level resolution is preferred. Since only some discrete phase values can be encoded due to the finite resolution of the output device, the phase value is subject to quantization error. Finally the large-scale transparencies are photoreduced to a size appropriated for illumination with visible light.

The gray-level transparency after photoreduction is then subjected to photographic bleaching. The bleaching process will remove the darkened grains in photographic emulsion. Not only the bleaching but also the exposure and development of the photoreduction for a kinoform must be performed with care. In particular, the bleaching must be carefully controlled such that the surface profile of the end product is such that light incident on a region of zero phase will be retarded by one wavelength relative to light incident on a region of 2π phase. In this case, all the incident light on the kinoform will appear in the projected image. The drawback of the kinoform is that the error in phase matching the $(0, 2\pi)$ interval results in a single bright spot on the optical axis (middle of the desired image). Figure 9.2*a* shows the gray-level pattern that leads to a kinoform, while Figure 9.2*b* is the image obtained from the kinoform shown in Figure 9.2*a*.

9.4 DESIGN METHODS FOR CGH AND DOE

Because the optical system has enormous complexity, there exist a number of mathematical problems that do not have analytical solutions. The CGH/DOE design clearly belongs to this type of problem [11–18]. If the analytical method

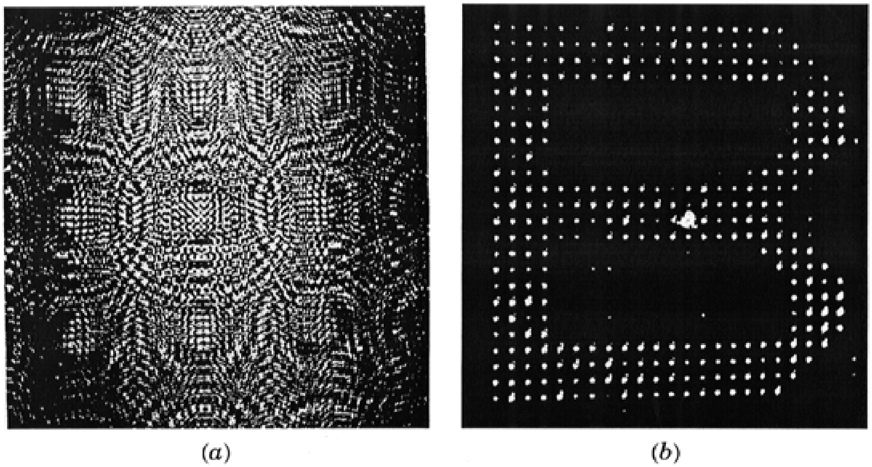


Figure 9.2 (*a*) Photograph of 32-gray-level plot of computed wave-front phase for two-dimensional letter *B*. (*b*) Photograph of two-dimensional kinoform image of letter *B* [20].

fails, then it is often possible to solve the problem by an iterative method. In this section, we discuss three different approaches to solve the CGH/DOE design problem. In summary, the CGH/DOE design problem has the following specific properties:

1. To synthesize a transform pair that has desirable properties in both input and output planes.
2. To reconstruct an object when only partial information is available in each of two planes.

To simplify the problem, we first consider the Fourier transform as the mathematical model for field propagation. Since this type of modeling is incorrect in some applications, we show another design method based on the rigorous coupled-wave analysis.

The synthesis problem arises when the Fourier transform of an input object (e.g., signal, aperture, antenna array) has certain desirable properties (such as uniform spectrum and low side lobes) and the object itself must satisfy certain constraints. However, it is not easy to find a Fourier transform pair (or such a pair does not exist) to satisfy all the constraints in both input and output planes. Therefore, the practical approach is to find a Fourier transform pair that can meet the constraints in both planes as much as possible. In reconstruction problems, only partial information in both planes is measured or few constraints are known a priori. Therefore, the information available is not enough to reconstruct the CGH/DOE transmittance function or the output diffraction pattern of a CGH/DOE. We must find a CGH/DOE transmittance function that can meet the known constraints. In summary, given sets of constraints on the CGH/DOE transmittance function and its output diffraction pattern, the synthesis and reconstruction problems require finding a transform pair to satisfy these constraints.

Is it important that the solution of the problem be unique? For synthesis problems of the CGH/DOE, the solution's uniqueness is not important because there is only one existing solution that can satisfy all constraints in input and output planes. The satisfied constraints may be arbitrary and conflicting, so the more important issue for synthesis problems is whether a solution exists or not. If the solution exists, then the solution must be able to be found by some methods. For reconstruction problems, the uniqueness properties of the solution are important. There may be many different transmittance functions of the CGH/DOE that give rise to the same measured data and satisfy the constraints. So the solution cannot be guaranteed to be the correct one. Fortunately, the uniqueness of the solution is not an issue because the problem is usually solved by the iterative method that could minimize the error in approximation.

The iterative method is shown to be very powerful in solving the problems mentioned above. There are many variations in the iterative method and they are not limited to a single fixed algorithm. The most known approaches are error reduction and the input–output approach. In this section, we first review error reduction and the input–output approach and finally we discuss the rigorous coupled-wave analysis approach to increase the accuracy of the design.

9.4.1 Error Reduction

Gerchberg and Saxton [19] used error reduction to solve the electron microscopy problem. In this approach, both the modulus of an input complex-valued object and the modulus of its output are measured and the the goal is to reconstruct the phase in both input and output planes and the Fourier transform is used to model the propagation of the field. Actually error reduction was invented somewhat earlier by Lesem and Hirsch [20] to solve a synthesis problem for kinoform with a similar set of constraints.

Again, Figure 9.1 (when $f_1 = 0$ and $f_2 \gg 0$) can be used to show the optical architecture of the reconstruction problem. Here, $U_1(x, y)$ is a plane wave and $U_1'(x, y)$ is the field right after the optical element—the CGH/DOE. In the reconstruction problem, the relation between an input $U_1'(x, y)$ and its output $U_2(f_x, f_y)$ is

$$U_2(f_x, f_y) = \mathfrak{F}[U_1'(x, y)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} U_1'(x, y) \exp[j2\pi(x f_x + y f_y)]\, dx\, dy \tag{9.11}$$

where the vector (x, y) represents spatial, angular, or other coordinates and the vector (f_x, f_y) represents spatial, angular, or other frequencies. Depending on the problem itself, the coordinates may be 1-, 2-, or 3D; however, we only consider the 2D case here to simplify the problem. For a reconstruction problem, there is only partial information available in the input–output plane. The problem is to reconstruct $U'_1(x, y)$ and/or $U_2(f_x, f_y)$ by the limited information (or constraints) in the input–output plane. For example, the information may be the measurability of $|U_2(f_x, f_y)|$ but $\theta(f_x, f_y)$ is unknown. For a synthesis problem, the functions $U'_1(x, y)$ and $U_2(f_x, f_y)$ have certain desirable properties (or satisfy certain constraints). For example, in the output plane, it may be desirable to obtain a specified value of $|U_2(f_x, f_y)|$ while simultaneously having a specified value of $|U'_1(x, y)|$ in the input plane.

Error reduction solves the problem of finding a Fourier transform pair to satisfy the input and output constraints. A block diagram of the error reduction approach is shown in Figure 9.3 with $U'_1 \equiv g(x, y)$. The kth iteration of the error reduction approach proceeds as follows. First, the Fourier transform is performed on the transmittance function of the CGH/DOE, $g_k(x, y)$, to yield the solution

$$G_k(f_x, f_y) = |G_k(f_x, f_y)| \exp[j\phi_k(f_x, f_y)] = \mathfrak{F}[g_k(x, y)] \tag{9.12}$$

According to the output constraints, a new function $G'_k(f_x, f_y)$ is obtained by adding the smallest possible change in $G_k(f_x, f_y)$ to make it satisfy the constraints. For example, if the output constraint is that $|U_2(f_x, f_y)|$ is substituted for the modulus of $G_k(f_x, f_y)$, then $G'_k(f_x, f_y)$ is given by

$$G'_k(f_x, f_y) = |U_2(f_x, f_y)| \exp[j\, \Phi_k(f_x, f_y)] \tag{9.13}$$

That is, the Fourier transform has a given (or measured) modulus equal to $|U_2(f_x, f_y)|$, and the phase of $G_k(f_x, f_y)$ is left unchanged. The spatial domain function, $g'_k(x, y)$, is then obtained by taking the inverse Fourier transformed on $G'_k(f_x, f_y)$. A new function $g'_{k+1}(x, y)$ is then obtained from the input constraints and $g'_k(x, y)$.

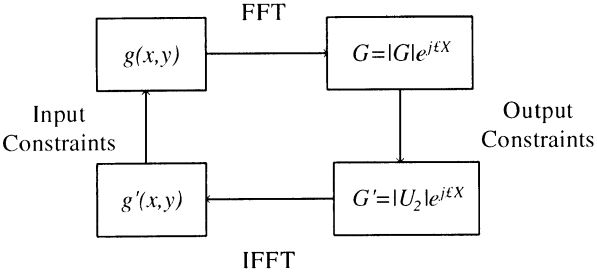


Figure 9.3 Block diagram of error reduction approach.

There are many choices for the initial condition. For example, we may set $g_1(x, y)$ or $\Phi_1(f_x, f_y)$ equal to an array of random numbers. The iterations will then proceed until a Fourier transform pair is found to satisfy all the input and output constraints.

The mean-square error (MSE) is usually used to measure the progress of the iterations and is a criterion to stop the iteration. When the MSE is smaller than a predefined threshold value, an acceptable solution is assumed to be found and the iteration process stops. The definition of the MSE in the output plane is

$$E_O^2 = \frac{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} |G_k(f_x, f_y) - G'_k(f_x, f_y)|^2 \, df_x \, df_y}{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} |G'_k(f_x, f_y)|^2 \, df_x \, df_y} \tag{9.14}$$

while that in the input plane is

$$E_I^2 = \frac{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} |g_k(x, y) - g'_k(x, y)|^2 \, dx \, dy}{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} |g'_k(x, y)|^2 \, dx \, dy} \tag{9.15}$$

When the MSE is zero, the iteration stops. In this case, the all the input and output constraints are satisfied and a solution is found. For most problems, the MSE can only decrease after each iteration but does not go to zero.

For the first few iterations of the error reduction, the error decreases very rapidly. However, the error decreases slowly for the later iterations. The error reduction is successful in finding solutions using a reasonable number of iterations for some optical elements. However, the MSE decreases slowly and requires a large number of iterations for convergence. Of course, if a different initial condition is set, then the required number of iterations for the algorithm to converge is changed.

9.4.2 Input–Output Approach

A faster convergence approach, the input–output approach, was developed to improve the converging rate of the error reduction approach. The input–output and error reduction approaches are identical in the three operations— Fourier transforming $g(x, y)$, satisfying output constraints, and inverse Fourier transforming the result. The difference between the input–output and error reduction approaches is in the input plane operation. If we group these three operations together (as demonstrated in Fig. 9.4), then the iteration becomes a nonlinear system with input $g(x, y)$ and output $g'(x, y)$.

The key property of this system is that its output $g'(x, y)$ when inverse Fourier transformed must satisfy the input constraints. Therefore, when the output $g'(x, y)$ also satisfies the input constraints, then all the constraints are satisfied and a solution to the problem is found. The design problem becomes manipulating the input to force the output $g'(x, y)$ to satisfy the input constraints.

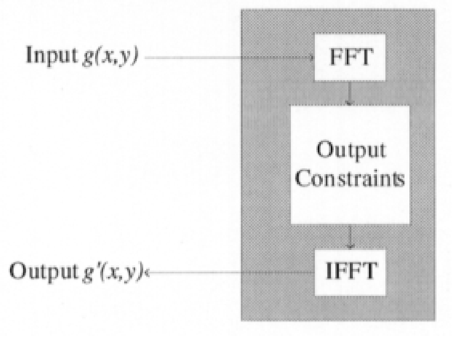


Figure 9.4 Block diagram of input–output approach.

In summary, the next input $g(x, y)$ is chosen to be the current best estimate of the input and to satisfy the input constraints for the error reduction. However, the input does not satisfy the input constraints, and it is unnecessarily an estimate of the input or a modification of the output in the input–output approach. It is just the driving function for the next output $g'(x, y)$.

How the input should be changed in making the output $g'(x, y)$ to satisfy the constraints depends on the particular problem at hand. At the kth iteration, the input is $g_k(x, y)$, which makes the output $g'_k(x, y)$. When the input is changed by adding $\Delta g(x, y)$,

$$g_{k+1}(x, y) = g_k(x, y) + \Delta g(x, y) \tag{9.16}$$

then the new output resulting from input $g_{k+1}(x, y)$ would be

$$\mathrm{g}_{k+1}(x, y) = g'_k(x, y) + \alpha \Delta g(x, y) + \text{additional noise} \tag{9.17}$$

That is, the expected output $g'_{k+1}(x, y)$ is the previous output $g'_k(x, y)$ plus a constant α times the change of the input $\Delta g(x, y)$. Although the system shown in Figure 9.4 is not linear, the change of input is usually small and will result in a small change of output. In this case, we can assume that the change of input and that of output are linearly related, and a logical choice for the new input is

$$g_{k+1}(x, y) = g_k(x, y) + \beta \Delta g_d(x, y) \tag{9.18}$$

where $g_k(x, y) + \Delta g_d(x, y)$ satisfies the input constraints. If $\beta = 1$, then the input–output approach reduces to the error reduction approach. Therefore, the input–output approach is considered as a more general approach to solve the CGH/DOE design problem.

9.4.3 Asymmetric Transform

Design methods of the CGH/DOE shown in previous sections is, actually, based on the theory of projection onto a convex set [21] and the far-field assumption. The Fourier transform is used to calculate the diffraction pattern of the CGH/DOE. At the same time, the transmittance of a CGH/DOE and its diffraction pattern (Fourier transformation) must satisfy both the input and output constraints.

In the Gerchberg–Saxton algorithm, finite-aperture effects of a CGH/DOE are ignored. To remedy this problem, a modified technique called the Yang-Gu algorithm [22–24] is proposed. The Yang-Gu algorithm takes the CGH/DOE finite-aperture effect into account, and thus the field at the output plane of a designed CGH/DOE can be better approximated to the desired one. Since amplitude transmittance of the CGH/DOE designed by the Gerchberg–Saxton or Yang-Gu algorithm is constant, the field at the output plane cannot exactly match the desired one and thus some discrepancies exist.

Both Yang-Gu and Gerchberg–Saxton [25] algorithms calculate the diffraction pattern at the output plane of a CGH/DOE by Fourier transformation. However, the propagation of a monochromatic light wave can only be approximated by Fourier transformation when the distance f_2 (as shown in Fig. 9.1) between the CGH/DOE and its output plane is very large (i.e., the far-field assumption). Some CGHs/DOEs may be used in applications where the above distance is short. In this case, the Yang-Gu (or Gerchberg–Saxton) algorithm does not provide an accurate CGH/DOE design.

To solve this problem, the rigorous coupled-wave analysis [26–31] is integrated into the existing Yang-Gu algorithm for CGH/DOE design. In short, the rigorous coupled-wave analysis is used to calculate the diffraction pattern behind the CGH/DOE at any distance away (i.e., the output plane); however, the inverse Fourier transform is still used to calculate the field [$U_1'(x, y)$ in Fig. 9.1] existing in the CGH/DOC. Since rigorous coupled-wave analysis does not require the far-field assumption, the CGH/DOE can be better designed when the distance between the CGH/DOE and its output plane is short. The rigorous coupled-wave analysis is used to calculate the diffraction pattern of a CGH/DOE; this approach is also named a rigorous coupled-wave analysis algorithm. Since two different approaches (inverse Fourier transform and rigorous coupled-wave analysis) are used, the algorithm is an asymmetric approach.

9.4.3.1 Rigorous Coupled-Wave Analysis Both error reduction and input–output approaches require a symmetric transform (usually a Fourier transform) to model the field propagation. Here, an error reduction–like algorithm based on asymmetric transform is shown. The optical system being considered is demonstrated in Figure 9.1 where the optical element is the CGH/DOE and $U_1(x, y)$ is a plane wave. In this case, the transmittance function of the CGH/DOE becomes $U_1'(x, y)$. This optical system is composed of one CGH/DOE with optical transmittance function [OTF, $g(x, y)$] and an output plane that is a distance f_2 behind the CGH/DOE. [As before, we have $g(x, y) \equiv U'_1(x, y)$.] The wave function right after the CGH/DOE and the diffraction pattern are denoted by $U'_1(x, y)$ and $U_2(x, y)$, respectively, and expressed as

$$U'_1(x, y) = A_1(x, y)\exp[j\theta_1(\mathrm{x}, \mathrm{y})] \tag{9.19}$$

$$U_2(f_x, f_y) = A_2(f_x, f_y)\exp[j\theta_2(\mathrm{f_x}, \mathrm{f_y})] \tag{9.20}$$

The z axis is chosen as the optical propagation axis of the system, and the coordinates of the input and output planes are (x, y) and (f_x, f_y), respectively.

The rigorous coupled-wave analysis is first reviewed in this section. The geometry of the CGH/DOE diffraction problem being treated is demonstrated in Figure 9.5. The incident electromagnetic wave upon the CGH/DOE produces both forward- and backward-diffracted waves. The input region is a homogeneous dielectric with a relative permittivity ε_1. Likewise, the output region is homogeneous with a complex permittivity ε_3. The forward-diffracted waves are absorbed as they propagate in the lossless output region. The CGH/DOE region consists of a periodic distribution of two types of homogeneous materials. The permittivity in region 2 may be expanded in a Fourier series as

$$\varepsilon(x, z) = \varepsilon(x + \Lambda, z) = \sum_p \varepsilon'_p(z)\exp(jpkx) \tag{9.21}$$

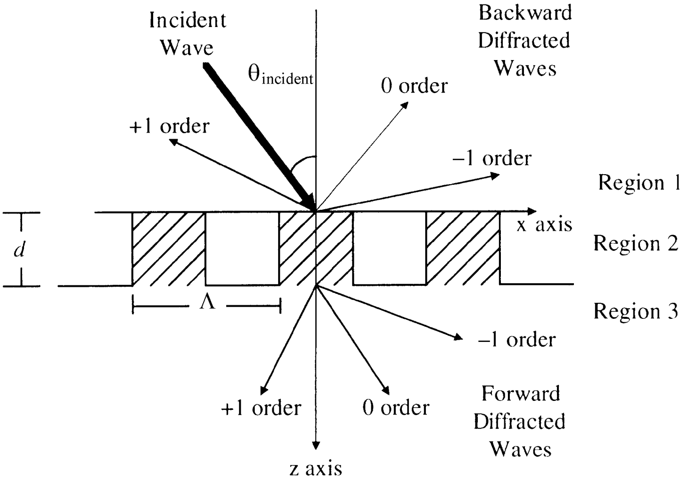


Figure 9.5 Geometry of CGH/DOE diffraction problem.

where Λ is the grating period, $k = 2\pi/\lambda$, and the quantity ε'_p is the pth Fourier component of the complex relative permittivity that is a function of z.

In the input region, the backward-diffracted wave exists because the finite depth of the CGH/DOE creates the finite diffracted light in all orders within the CGH/DOE volume. The diffracted light in all orders includes both forward and backward waves from the boundary at $z = 0$. The total electric field in the input region is the sum of the input and the backward-diffracted waves. The normalized total electric field in the input region is

$$E_1 = \exp[-j(k_{x0}x + k_{z0}z)] + \sum_i R_i \exp[-j(k_{xi}x + k_{ziI}z)] \qquad (9.22)$$

where $k_{x0} = k\varepsilon_1^{0.5} \sin\, \theta_{\text{incident}}$, $k_{z0} = k\varepsilon_1^{0.5} \cos\, \theta_{\text{incident}}$, $k = \pi/\lambda$, θ_{incident} is the angle of incidence, λ is the optical wavelength in free space, R_i is the amplitude of the ith-order backward-diffracted wave, and k_{xi} is determined from the vector Floquet condition. Likewise, the total electric field in output region 3 is

$$E_3 = \sum_i T_i \exp\{-j[k_{xi}x + k_{zi3}(z - d)]\} \qquad (9.23)$$

where T_i is the amplitude of the ith transmitted wave in the output region with wave vector k and d is the groove depth. In a region of CGH/DOE that has groove depth between 0 and d, the electric field may be expressed as

$$E_2 = \sum_i S_i \exp[-j(k_{xi}x + k_{zi2}z)] \qquad (9.24)$$

where S_i is the amplitude of the ith wave field at any point within the modulated region. For a given value of i, the wave field inside the CGH/DOE is not a simple plane wave and may be expressed as a superposition of infinite plane waves. The superposition includes forward-traveling waves and corresponding backward-traveling waves. The amplitudes are to be determined by solving the modulated region wave equations (Maxwell's equations) are

$$\nabla^2 E_2 + k^2 \varepsilon(x, z) E_2 = 0 \qquad (9.25)$$

On substitution of $S_i(z)$ from Eq. 9.24 into Eq. 9.25, the wave equation becomes

$$\frac{d^2 S_i(z)}{dz^2} - 2jk_{z0}\frac{dS_i(z)}{dz} = (k_{xi}^2 + k_{x0}^2)S_i(z) - k^2 \sum_p \varepsilon'_p(z) S_{i-p}(z) \qquad (9.26)$$

where $i = -\infty, \cdots, 0, \cdots, \infty$.

Equation 9.26 can be solved by using the state variable method, which originated from linear system analysis. The state equation may be expressed concisely as $\dot{S} = \mathbf{A}S$, where $\dot{S}$ and S are column vectors and $\mathbf{A}$ is the total coefficient matrix. Equation 9.26 is unsolvable because there are infinite numbers of equations coupled together. To solve this equation, we must purposely ignore some equations when their index $|i|$ is greater than M (where M is an integer). In this case, the $S_i(z)$ term (where $|i| > M$) is discarded. Thus $S_i(z)$ (where $|i| \leqslant M$), it can be expressed as

$$S_i(z) = \sum_{m=-\infty}^{z} C_m \omega_{im} \exp(\lambda_m z) \approx \sum_{m=-M}^{M} C_m \omega_{im} \exp(\lambda_m z) \tag{9.27}$$

where λ_m and ω_{im} are the eigenvalues and eigenvectors, respectively, of the $\mathbf{A}$ matrix. The qualities of C_m are unknown constants that are determined from the boundary condition.

The observable phenomenon of the electromagnetic wave requires that the tangential electric and magnetic fields be continuous, that is, the so-called boundary conditions. By the theory shown above and boundary conditions, we have the following four equations to determine the R_i and T_i if the transmittance function of the optical element is known:

$$\delta_{i0} + R_i = \sum_{m=-\infty}^{\infty} C_m \omega_{im} \approx \sum_{m=-M}^{M} C_m \omega_{im} \tag{9.8}$$

$$(\delta_{i0} - R_i)\kappa_{zi} = \sum_{m=-\infty}^{\infty} C_m \omega_{im}(\kappa_{z0} + j\lambda_m) \approx \sum_{m=-M}^{\infty} C_m \omega_{im}(\kappa_{z0} + j\lambda_m) \tag{9.29}$$

$$T_i = \sum_{m=-\infty}^{\infty} C_m \omega_{im} \exp(\lambda_m d) \approx \sum_{m=-M}^{M} C_m \omega_{im} \exp(\lambda_m d) \tag{9.30}$$

$$\begin{aligned} T_i \kappa_{zi3} &= \sum_{m=-\infty}^{\infty} C_m \omega_{im}(\kappa_{z0} + j\lambda_m) \exp(\lambda_m d) \\ &\approx \sum_{m=-M}^{M} C_m \omega_{im}(\kappa_{z0} + j\lambda_m) \exp(\lambda_m d) \end{aligned} \tag{9.31}$$

9.4.3.2 Rigorous Coupled-Wave Analysis Algorithm The error reduction and input-output approaches discussed before are based on the far-field assumption. That is, the Fraunhofer diffraction pattern of a CGH/DOE is approximated by its Fourier transformation. For this approximation to be valid, both the scalar diffraction theory and the far-field assumption must be satisfied.

However, Fourier transformation does not account for multiple reflections within a CGH/DOE. When the depth-to-wavelength ratio (d/λ) of a CGH/

DOE is large or the period-to-wavelength (Λ/λ) ratio is small, the Fourier approximation to the light propagation is inaccurate. In addition, the far-field assumption does not hold when the desired output plane is very close to the CGH/DOE. Therefore the Gerchberg–Saxton algorithm is not a good approach to design a CGH/DOE when the scalar diffraction theory and far-field assumption do not hold.

To better design a CGH/DOE, the rigorous coupled-wave analysis algorithm can be used. In this algorithm, the concept of projection onto a convex set is still employed, but the rigorous coupled-wave analysis is used to calculate the diffraction pattern at any distance f_2 behind the CGH/DOE. In this case, the diffraction pattern at the output plane can be accurately calculated. However, for simplicity, the inverse Fourier transform is still used to calculate the CGH/DOE function from the output plane. The algorithm converges when the diffraction patterns from two consecutive iterations are similar and satisfy the convergence condition. Since the rigorous coupled-wave analysis is used to calculate the output diffraction pattern of a CGH/DOE, transmittance of the CGH/DOE can be better designed compared with the Gerchberg–Saxton algorithm, which is only applicable under the far-field assumption. In addition, the forward and backward transforms are different; the algorithm is also named the asymmetric approach. The proposed asymmetric algorithm is shown below:

Asymmetric Algorithm

1. Given a threshold value ε, repeat the procedure below until convergence. Then gradually decrease ε to find the minimum ε for the algorithm to converge.
2. Start with a desired output diffraction pattern $U_2(f_x, f_y)$; the OTF $g_0(x, y)$ (at zero iteration) of the desired CGH/DOE is the inverse Fourier transformation of $U_2(f_x, f_y)$.
3. Set the OTF of $g_0(x, y)$ to 1 (input constraint) for every set of (x, y) to account for the property of the CGH/DOC. Let the CGH/DOE new transmittance be $g'_0(x, y)$.
4. Use the rigorous coupled-wave analysis to calculate the diffraction pattern $U'_2(f_x, f_y)$ of $g'_0(x, y)$ at a distance f_2 behind the CGH/DOE plane.
5. For every set of (f_x, f_y),

$$U_2(f_x, f_y) = \begin{cases} U_2(f_x, f_y) + \varepsilon & \text{if } U'_2(f_x, f_y) > U_2(f_x, f_y) + \varepsilon \\ U_2(f_x, f_y) - \varepsilon & \text{if } U'_2(f_x, f_y) < U_2(f_x, f_y) - \varepsilon \\ U_2(f_x, f_y) & \text{otherwise} \end{cases} \tag{9.32}$$

Repeat steps 2–5.

6. If $|U_2(f_x, f_y) - U'_2(f_x, f_y)| < \varepsilon$, the algorithm converges. The desired transmittance of the CGH/DOE (under the given ε) is $g'_k(x, y)$ (at the kth iteration).

The proposed algorithm uses rigorous coupled-wave analysis to calculate the output diffraction pattern of a CGH/DOE that is a distance f_2 away from it. Then the inverse Fourier transform is used to calculate the transmittance of the CGH/DOE from a diffraction pattern. At first glance, the algorithm seems odd because it uses two different transformations to project the diffraction pattern and transmittance function back to each other. Although the inverse Fourier transform is imprecise in obtaining the transmittance function of the CGH/DOE, the rigorous coupled-wave analysis is accurate. In addition, only the convergence of the algorithm is checked when the output diffraction pattern of the CGH/DOE matches the desired one. Therefore, the desired output diffraction pattern is obtained when the algorithm converges.

The convergence of the algorithm depends on the threshold value ε. Here, a large threshold value ε is first set such that the algorithm can easily converge. Then we decrease the threshold value ε gradually to make the algorithm provide a better transmittance function for the desired diffraction pattern. Since this algorithm uses rigorous coupled-wave analysis to calculate the diffraction pattern, the effects of the wavelength λ, depth d, propagation distance f_2, and period Δ can be fully considered during the calculation of the output diffraction pattern. Since these considerations are ignored in the Gerchberg–Saxton algorithm, the asymmetric algorithm shown here can thus provide a better result compared with the conventional approaches.

9.5 FABRICATION OF COMPUTER-GENERATED HOLOGRAM

In this section we briefly describe the fabrication of the CGH/DOE with e-beam lithography. There are several traditional procedures for fabricating the CGH/DOE [17, 32, 33]. After the transmittance of the CGH/DOE is calculated, it is usually transferred to a mask. Traditionally, this has been accomplished by either plotting the CGH/DOE on acetate or exposing photographic film to the light from a cathode-ray tube (CRT) display. In either case, photoreduction is usually necessary to produce the final hologram. Recently, an e-beam lithography technique identical to those used in the fabrication of integrated circuits has been used in creating the CGH/DOE.

Before the CGH/DOE pattern can be written by e-beam lithography, the geometric patterns generated by the design algorithm must be translated to a format that the e-beam will accept. This process, so-called fracturing, consists of taking arbitrary geometric shapes and sectioning them into graphical pattern such as rectangles and trapezoids [32, 34] that the e-beam system will accept.

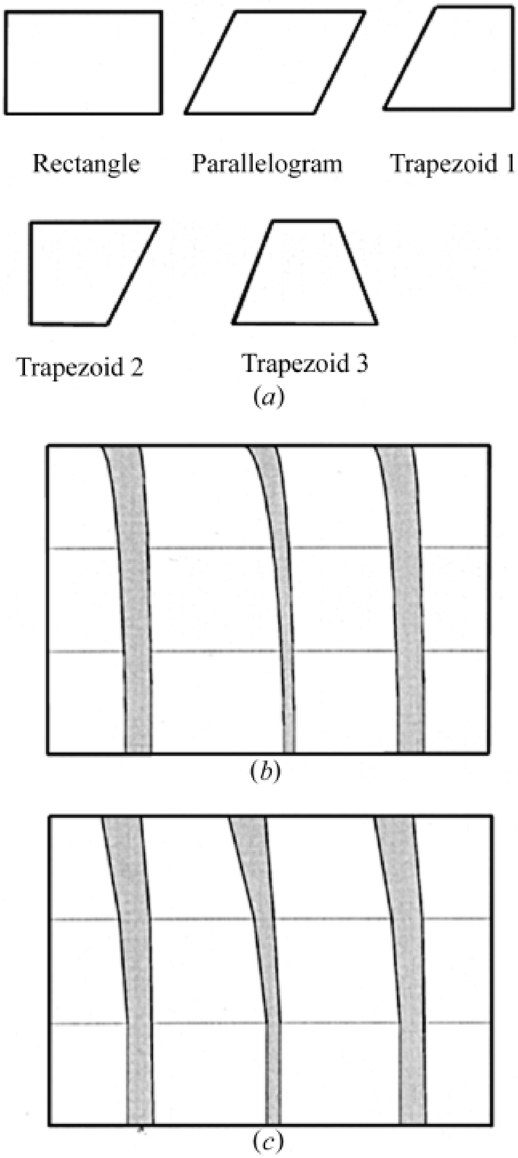


Figure 9.6 (*a*) MEBES figure set. (*b*) Sample image of fringe to be fractured. (*c*) Image fractured by using MEBES figure set.

Graphical data are first fractured and then specified in an e-beam format. Some of the existing formats are MEBES, Cambridge, and JEOL, with MEBES becoming a defactor standard format that many e-beam systems will accept.

Figure 9.6*a* shows the MEBES figure set that consists of acceptable parallelograms and trapezoids, while Figure 9.6*b* shows a sample set of fringes to be fractured. Fracturing consists of an arc-to-chord approximation followed by the formation of acceptable trapezoids, as shown in Figure 9.6*c*. Careful choice of fracturing locations results in data reduction. If the fringes are fractured at the stripe boundaries, additional fracturing, which produces associative trapezoids, needs not be performed. Also, if the fringes are fractured into shapes with horizontal top and bottom boundaries, additional fracturing to produce the acceptable shapes are not necessary.

In MEBES, the pattern to be written is divided into address units (AUs). The e-beam spot size is numerically equal to the distance between AUs that vary from 0.05 μm to 1.1 μm. On some e-beam systems, this spot size variation can be used to save exposure time. If only a small portion of a CGH/DOE requires high resolution, the spot size can be changed during the exposure. The spot size is made smaller by decreasing the current but also increases the exposure time per address and limits the beam spread. By using small spot sizes when absolutely necessary, the overall exposure time is reduced.

In MEBES, any pattern up to 6 in^2 can be written by dividing the pattern into stripes and segments. Each stripe is 32,768 × 256,512 or 32,768 × 1024 AUs. Each segment is one stripe wide and is made up of as many stripes in the *Y* direction as necessary to complete the pattern in the *Y* direction. The pattern is constructed with as many segments as necessary, up to 50, in the *X* direction. The exposure sequence begins with the leftmost segment, writing the stripes from bottom to top within this segment, and then starts at the bottom of the next segment, repeating until all segments have been written.

Due to the exposure sequence of the e-beam, the fractured data must be stored geographically by stripes. In this case, the e-beam receives all the trapezoids within one stripe before receiving the trapezoids in the next stripe. If a large CGH/DOE is being written, the stripes must be stored by segments. If the fracturing algorithm does not pay attention to the stripe boundaries when fracturing, additional fracturing must be performed at the stripe and segment boundaries. The e-beam writes the primitive shapes in the layer of e-beam resist on top of a chrome-coated glass substrate. After development removes the exposed resist, the chrome in the open regions is etched from the substrate to form a binary mask. This mask can be used directly as a thin transmission hologram or it can be used to expose photoresist for etching quartz to create a surface relief phase hologram. Multiple masks, 2^N phase levels can be achieved using N masks.

Figure 9.7 demonstrates the process to achieve a sawtooth thickness function by using four phase levels (two masks). The process has several discrete steps. Each process consists of a photeresist application, exposure through a binary mask, photoresist removal, and etching. Masks are made by e-beam writing shown above. Figure 9.7*a* demonstrates a substrate overcoated with photoresist. This substrate is exposed through the first binary mask with transparent cells of width equal to 1/4 ($1/2^2$) of the period of the desired final

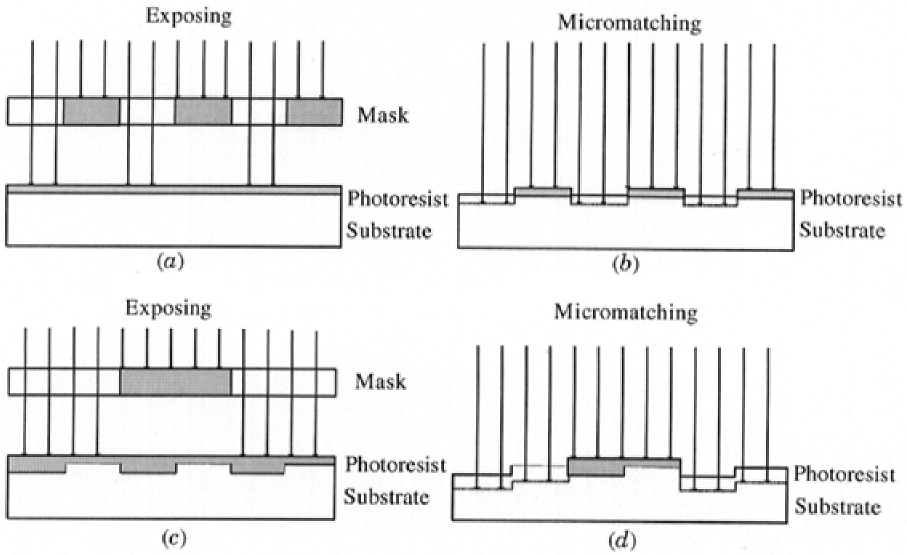


Figure 9.7 Procedures in fabricating four-level CGH/DOE: (*a*) exposing; (*b*) micromatching; (*c*) exposing; (*d*) micromatching. [Note difference in mask shown in (*a*) and (*c*).]

structure. After the exposure process, the photoresist is developed. For a positive photoresist, the development process removes the exposed area and vice versa for negative photoresist. After photoresist development, micromachining (reactive ion etching or ion milling) is applied to remove material from the uncovered portions of the substrate, as demonstrated in Figure 9.7*b*. The first micromachining step removes substrate material to a depth of 1/4 ($1/2^2$) of the desired peak-to-peak depth of the grating. Now photoresist is spun onto the substrate again and is exposed through a second mask that has openings of width equal 1/2 ($1/2^{2-1}$) of the desired final period, as illustrated in Figure 9.7*c*. Micromachining removes the exposed area of the substrate again. However, the depth being etched is 1/2 ($1/2^{2-1}$) of the final desired maximum depth, as shown in Figure 9.7*d*. A prototype of two masks is shown is Figure 9.8.

For the 2^N-level, N different masks, exposures, development, and etching processes are required. The last etching process must be a depth that is half of the total desired peak-to-peak depth. Many different materials, such as silicon and glass, can be used for the substrate. We can also make reflective optical devices by overcoating the etched profile with a thin layer of metal. With the use of e-beam writing, the accuracy of the mask is about 0.1 μm. However, the accuracy is reduced if the profile is complex or alignment of several masks is required. Usually, the diffraction efficiency of the CGH/DOE is about 80%.

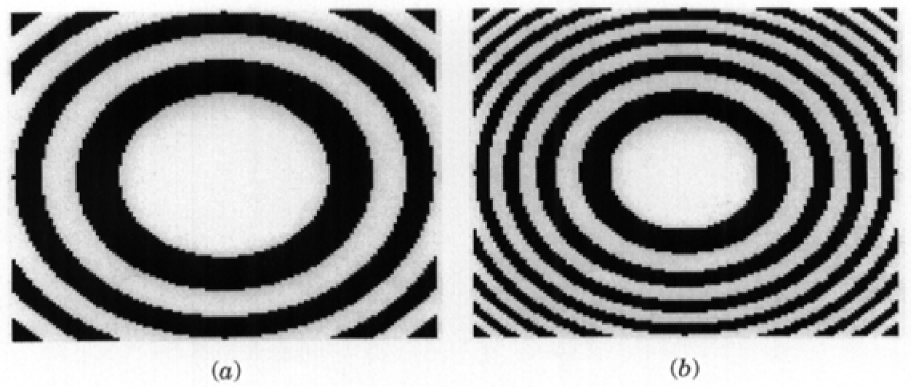


Figure 9.8 Two masks to achieve focusing CGH/DOE: (*a*) first mask; (*b*) second mask.

9.6 CONCLUSIONS

The design methods and fabrication process of the CGH/DOE is reviewed in this chapter. We explained several important issues related to the CGH/DOE, including sampling, computational, design, and fabrication issues. Although the design methods reviewed in this chapter only deal with constant-amplitude transmittance, the idea can be directly applied to design the CGH/DOE with arbitrary amplitude and phase transmittance. At the same time, high-energy beam-sensitive gas technology [35, 36] is now available to fabricate a CGH/DOE with amplitude and phase variations in its transmittance. Therefore, a CGH/DOE with desired output diffraction pattern is obtainable. Since the CGH/DOE can now be easily designed and fabricated, it becomes more and more important in many applications.

ACKNOWLEDGMENT

This work was supported by National Science Council, Taiwan under contracts NSC NSC-87/88/89-2736-L-194-001/002.

REFERENCES

1. L. Eng, K. Bacher, W. Yuen, J. Harris, and C. Chang-Hasnain, “Multiple-wavelength vertical cavity laser arrays on patterned substrates,” *IEEE J. Quant. Electron.* **1**, 624–628 (1995).
2. C. C. Hasnain, J. Harbison, C. Zah, M. Maeda, L. Stoffel, and T. Lee, “Multiple wavelength tunable surface-emitting laser arrays,” *IEEE J. Quant. Electron.* **27**, 1368–1376 (1991).

3. W. Yuen, G. Li, and C. Chang-Hasnain, “Multiple-wavelength vertical-cavity surface-emitting laser arrays with a record wavelength span,” *IEEE Photon. Technol. Lett.* **8**, 4–6 (1996).
4. A. W. Lohmann and D. P. Paris, “Binary Fraunhofer holograms, generated by computer,” *Appl. Opt.* **6**, 1739–1748 (1967).
5. J. W. Goodman, *Introduction to Fourier Optics*, McGraw-Hill, New York, 1968.
6. D. Mendlovic and H. M. Ozaktas, “Fractional Fourier transforms and their optical implementation: I,” *J. Opt. Soc. Am., Part A* **10**, 1875–1881 (1993).
7. H. M. Ozaktas and D. Mendlovic, “Fractional Fourier transforms and their optical implementation: II,” *J. Opt. Soc. Am., Part A* **10**, 2522–2531 (1993).
8. E. O. Brigham, *The Fast Fourier Transform*, Prentice-Hall, Englewood Cliffs, NJ, 1974.
9. R. C. Gonzalez and R. E. Woods, *Digital Image Processing*, Addison-Wesley, New York, 1992.
10. A. V. Oppenheim and R. W. Schafer, *Discrete-Time Signal Processing*, Prentice-Hall, Englewood Cliffs, NJ, 1992.
11. B. R. Brown and A. W. Lohomann, “Computer-generated binary holograms,” *IBM J. Res. Devel.* **13**, 160–168 (1969).
12. W. H. Brown, “Binary computer-generated hologram,” *Appl. Opt.* **18**, 3661–3669, (1979).
13. D. C. Chu, J. R. Fienup, and J. W. Goodman, “Multi-emulsion, on-axis, computer generated hologram,” *Appl. Opt.* **12**, 1386–1388 (1973).
14. W. J. Dallas, “Phase quantization — A compact derivation,” *Appl. Opt.* **10**, 673–674 (1971).
15. W. J. Dallas, “Phase quantization in holograms — A few illustrations,” *Appl. Opt.* **10**, 674–676 (1971).
16. J. R. Fienup, “Interactive method applied to image reconstruction and to computer-generated hologram,” *Opt. Eng.* **19**, 297–305 (1980).
17. R. Nordin, A. Levi, R. Nottenburg, J. Tanbun-Ek, and R. Logan, “A system perspective on digital interconnection technology,” *IEEE J. L ight Wave Technol.* **10**, 811–827 (1992).
18. F. Wyrowski and O. Bryngdahl, “Speckle-free reconstruction in digital hologram,” *J. Opt. Soc. Am., Part A* **6**, 1171–1174 (1989).
19. R. W. Gerchberg and W. O. Saxton, “A practical algorithm for the determination of phase from image and diffraction plane pictures,” *Optik* **35**, 227–246 (1972).
20. L. B. Lesem and P. M. Hirsch, “The kinoform: A new wavefront reconstruction device,” *IBM J. Res. Devel.* **13**, 150–158 (1969).
21. G. Strang, *Introduction to Applied Mathematics*, Wellesley-Cambridge Press, Wellesley, MA, 1986.
22. G. Yang and B. Gu, “On the amplitude-phase retrieval problem in the optical system,” *Acta Phys. Sinica* **30**, 410–413 (1981).
23. B. Gu and G. Yang, “On the phase retrieval problem in optical and electronic microscopy,” *Acta Optica Sinica* **1**, 517–522 (1981).
24. B. Gu, G. Yang, and B. Dong, “General theory for performing and optical transform,” *Appl. Opt.* **25**, 3197–3206 (1986).

25. R. W. Gerchberg and W. O. Saxton, “Phase determination for image and diffraction plane pictures in the electron microscope,” *Optik* **34**, 275–284 (1971).

26. M. G. Moharam and T. K. Gaylord, “Rigorous coupled-wave analysis of planar grating diffraction,” *J. Opt. Soc. Am.* **71**, 811–818 (1981).

27. M. G. Moharam and T. K. Gaylord, “Chain-matrix analysis of arbitrary-thickness dielectric reflection gratings,” *J. Opt. Soc. Am.* **72**, 187–190 (1982).

28. M. G. Moharam and T. K. Gaylord, “Planar dielectric grating diffraction theories,” *Appl. Phys. B* **28**, 1–14 (1982).

29. M. G. Moharam and T. K. Gaylord, “Diffraction analysis of dielectric surface-relief grating,” *J. Opt. Soc. Am.* **72**, 1385–1392 (1982).

30. M. G. Moharam and T. K. Gaylord, “Three-dimensional vector coupled-wave analysis of planar-grating diffraction,” *J. Opt. Soc. Am.* **73**, 1105–1112 (1983).

31. M. G. Moharam and T. K. Gaylord, “Rigorous coupled-wave analysis of metallic surface-relief gratings,” *J. Opt. Soc. Am., Part A* **3**, 1780–1787 (1986).

32. K. S. Urquhart, S. H. Lee, C. C. Guest, M. R. Feldman, and H. Farhoosh, “Computer aided design of computer generated holograms for electron beam fabrication,” *Appl. Opt.* **28**, 3387–3396 (1989).

33. F. Koyama, Y. Hayashi, N. Ohnoki, N. Hatori, and K. Iga, “Two-dimensional multiwavelength surface emitting laser arrays fabricated by nonplanar MOCVD,” *Electron. Lett.* **30**, 1947–1948 (1994).

34. H. Farhoosh, M. R. Feldman, S. H. Lee, C. Guest, Y. Fainman, and R. Eschbach, “Comparison of binary encoding schemes for electron-beam fabrication of computer generated hologram,” *J. Opt. Soc. Am.* **58**, 533–537 (1968).

35. W. Daschner, P. Long, and R. Stein, “Cost effective mass fabrication of multilevel Gluch diffractive optical elements by use of signal optical exposure with a gray-scale mask on high-energy beam-sensitive gas,” *Appl. Opt.* **36**, 4675–4680 (1997).

36. M. R. Wang and H. Su, “Laser direct-write gray-level mask and one-step etching for diffractive microlens fraction,” *Appl. Opt.* **37**, 7568–7576 (1998).

Three-Dimensional Holographic Imaging. Edited by Chung J. Kuo, Meng Hua Tsai
Copyright © 2002 John Wiley & Sons, Inc.
ISBNs: 0-471-35894-0 (Hardback); 0-471-22454-5 (Electronic)

CHAPTER 10

Is Catastrophe Analysis the Basis for Visual Perception?

IGOR TERNOVSKIY* and TOMASZ JANNSON

Physical Optics Corporation
2545 West 237th Street
Torrance, California 90505

H. JOHN CAULFIELD

Department of Physics, Fisk University
100 18th Avenue North
Nashville, Tennessee 37208

10.1 INTRODUCTION

When a three-dimensional (3D) scene is collapsed into a 2D, much information is lost, but the information loss mechanisms leave information about the depth dimension in the 2D image. A means to recover and use that 3D information from the 2D scene is called analysis by catastrophes (ABC). It segments the perceived world into meaningful wholes and their relationships. In the process, it would dramatically reduce the processing load on the visual cortex by offering a complete description of the scene in terms of only two primitives, the fold and cusp catastrophes, along with their particular locations, scales, and orientations. The completeness of this catastrophe description is illustrated along with proof that 3D information can be extracted from the 2D scene. These and other observations point to the possibility that something like ABC may be part of mammalian early vision.

10.2 IN SEARCH OF PERCEPTUAL ELEMENTS

Human vision has a number of remarkable features that need to be accounted for in any theory of visual processing. Among those are the fact that we

*Now with Extreme Teknologies, 5762 Bolsa Avenue, Suite 215, Huntington Beach, CA 92649.

perceive objects, not pixels, and that we can see 3D even without stereo, as when we view objects 10 m or more away, foveation search [1, 2], stereoscopic integration of features, which are different in the two retinal images by definition to determine to give depth information [1, Sec. 7; 3, 4], and the unequaled ability to recognize objects essentially independently of their size, orientation, translation, and illumination. A theory that could account for all of these phenomena and others as well would be valuable if it were also biologically plausible. But such a general model would have to go far beyond "objective" photometry [5] that cannot account for shape and Fourier methods [6], even in their modern extension to wavelets, which do not account for nonlinearity, an essential feature of early vision. Some hybrid geometric–physical theory is required, but all prior work seems to deal with one or the other but not the reality, which is both mixed together.

That vision as the result of the brain's computation is universally accepted, as is the fact that the brain is a very slow electrochemical processor. To reconcile the complexity of the task with the limitations of the processor, it is useful to reduce the scene description to the minimum number of primitive elements (both the number of element types and the number of primitives required to describe the scene). The first step in early vision should be to describe the scene in such terms. Almost certainly this accounts for the *saccadic* eye movement to bring the fovea onto "interesting" scene features two to three times a second, apparently fixating [1, 2] on, for example, *extremal boundaries* [7] (where the object surface turns smoothly away from the viewer [7]) and T-junctions of a stationary scene. Along with a few other authors [8, 9], we identify those characteristic or *singular* points with mathematical entities called *catastrophes*. For this reason, we turn to the branch of advanced analytical geometry called catastrophe theory [10, 11] as the mathematical starting place for an analytical model of how visual perception might occur.

10.2.1 Photogeometrical Manifold Description

Rigorous catastrophe theory (CT) is quite sophisticated mathematically and is a purely geometric theory, devoid of physical concepts such as illumination. Both factors weigh against it as a model of early vision. In this chapter, both objections will be met. An argument will be presented that modern neural networks offer a biologically plausible way to do something very like ABC. In addition, classical CT [10–12] will be broadened to become a photogeometric theory and no longer purely geometric. We turn to that extension of CT now.

Starting with the 2D retinal image with coordinates (u, v) of the 3D scene, we add a third dimension, W (more properly, we should use three new dimensions R, G, and B). This 3D space is now abstract and no longer purely geometric. We also extend the input space from (x, y, z) to the abstract 4D space $(x, y, z; B)$, where B is brightness. We now describe the collapse of the 4D input space into the 3D image space by ABC's nonlinear transformation (see Section 10.3). Next we describe each physical object as a manifold in terms of its own body-centered *normal* coordinates (ξ, h) as described in Ref. 12.

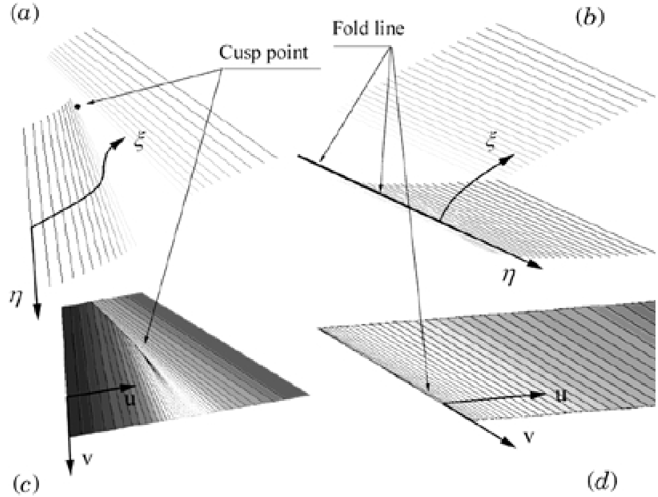


Figure 10.1 Illustration of Whitney's two stable catastrophes: cusp and fold; first one described in normal coordinates as $u = \xi^3 + \eta \cdot \xi$, $v = \eta$; the second one described by $u = \xi^2$, $v = \eta$. We could observe that *fold* catastrophe can be identified with external boundary [7]. Both catastrophes can be presented either as (*a*, *b*) 3D in (ξ, η) space or (*c*, *d*) 2D intensity in (u, v) space.

These lead to ABC singularities or catastrophes in their *canonical* or normal form. In ABC, we may deal with only the two stable Whitney [11] catastrophes from among the 14 catastrophes listed by Thom [10] and Arnold [12]. Figure 10.1 shows how both of those catastrophes, *cusp* and *fold*, have their 3D local structure projected into the 3D retinal space.

The ABC mapping from object space, in normal coordinates, into retina space has the form

$$u = F_1(\xi, \eta) \tag{10.1a}$$

$$v = F_2(\xi, \eta) \tag{10.1b}$$

$$W(u, v) = F_3(\xi, \eta; B) \tag{10.1c}$$

where Eqs. 10.1a and 10.1b are CT geometric projections, while Eq. 10.1c is a new physical formula, describing photometric relations between the object surface ($\xi, \eta; B$) and the retina space ($u, v; W$). Here ξ and η are the coordinates on each manifold, u and v are coordinates along the retina; B is brightness, and W is intensity. After a number of rigorous mathematical steps, described briefly in Section 10.3, we can rewrite formula 10.1c as a sum of two parts:

$$W(u, v) = \underbrace{M(B)}_{\text{(I)}} + \underbrace{g(\xi, \eta)}_{\text{(II)}} \tag{10.2}$$

where M is the regular (Morse [12]) form, representing standard photometric projection [5], while g is a new singular form, representing all object surface mapping singularities (i.e., cusp and fold; see Fig. 10.1).

Being the generalization of Thom's geometric lemma [12], formula 10.2 is the main result of the ABC model. It demonstrates a surprising result that, in addition to the regular photometric term (I), we also obtain the second, singular term (II) which *does not depend* on luminance B. To be in agreement with vision perception, which allows objects to be recognized independently of illumination and surface color, we decided to apply only term II (i.e., completely ignoring the regular term I) for image reconstruction. To our even greater surprise we obtained quite good full scene synthesis (Fig. 10.2).

10.2.2 Image Synthesis

In scene synthesis, the only free parameter is what we have called the *minimum catastrophe size*, a, which is a monotonically growing function of amount of scene compression or *data reduction* (DR) factor. Here DR is the ratio of the number of bits in the original scene to the number of bits needed to give the catastrophe description at scale a. If more detail is necessary, we can go to smaller a. Figure 10.2a shows an original scene, while Figure 10.2b shows its reconstruction from only the singular (catastrophic) term (II) altogether with linear approximation between singularities. *Thus, in ABC, virtually the entire image structure can be restored using only two primary features* (*the cusp and fold catastrophes*) *and their mutual location, orientation, and size.* Since any scene can be described well by those two primary elements (cusp and fold), they constitute the *primary element complete set* (PECS).

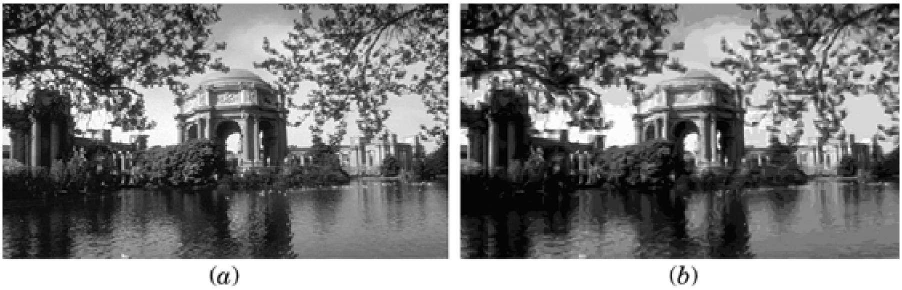


Figure 10.2 Demonstration of ABC-simulated retinal image of typical scene, including (a) original ($\mathrm{DR} = 1:1$) and (b) ABC-reconstructed image with $a = 0.016$ and $\mathrm{DR} = 20:1$ Catastrophe-resolving element, a, is defined as fraction of total scene horizontal size. Data-reduced image (b) has been obtained by applying only second (singular) term of Eq. 10.2, with linear approximation between singularities. Photometric term contribution can be recognized here as difference: (a) − (b). This means that ABC algorithm filters image entirely on basis of two catastrophes: *cusp* and *fold*. This seems analogous to foveal fixation points. Red square in corner illustrates size of catastrophe-resolving element, a.

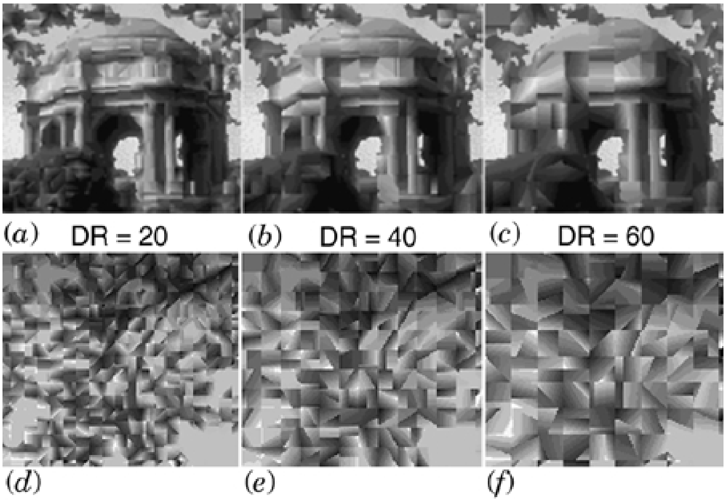


Figure 10.3 Similar to Figure 10.2*b*, with characteristic fragments of scene, such as *dome* and *leaves* for DR values 20, 40, and 60, and *a* values 0.016, 0.023, and 0.031. We can observe that image quality of *dome* is significantly better than that of *leaves* for corresponding *a* value [compare: (*a*) with (*d*) or (*b*) with (*e*)]. This means that ABC *nonlinear* filtering provides automatic segmentation process by extracting "object of interest" (here *dome*) by analogy to good paintings.

This result demonstrates that foveation could be a simple catastrophe location. This is perfectly adequate both for scene description and for massive data reduction (down to 1–5%). The catastrophes remain catastrophes of the same type and location when viewed from multiple "frames" such as the two eyes for stereo (or two time periods for optic flow; see Section 10.3). Analysis by catastrophe appears to meet these requirements for a model of early vision processing quite well.

Note that the information loss is not uniform across all parts of the scene. The scale factor has influenced the reconstructed image. The *dome* in Figure 10.2*b* is well preserved, while the background *leaves* are less well described. This is shown in more detail in Figure 10.3, where those specific scene portions are shown at a high DR. To give some sense of the quality of reproduction, we measure the *peak signal-to-noise ratio* or (PSNR) for those same two scene portions and various *a* or DR values, as in Figure 10.4. Objects become clearer as we attend closer to them. Remember that we did not define any object a priori. The *dome* arose as an object in Figure 10.2 fully unsupervised. The *leaves* emerge as distinct objects as we decrease *a*. This is precisely the way human vision seems to work.

Impressionist painters achieved the effect of backgrounds with "too-high *a*" routinely. Unlike nature that is somewhat fractal with new information

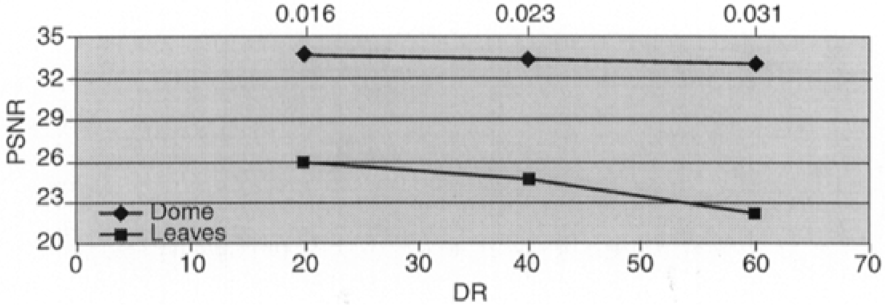


Figure 10.4 Peak signal-to-noise ratio (PSNR) as function of data reduction (DR), or catastrophe-resolving element's a value (proportional to DR value), for two characteristic fragments of scene (Fig. 10.1), as in Figures 10.3*a–c* (*dome*) and 10.3*d–f* (*leaves*). We see that PSNR values for *dome* are always significantly higher than corresponding PSNR values for *leaves*. The PSNR values are defined as $\text{PNSR} = 10\log_{10}\{255^2/(1/NM)\Sigma_{i=1}^{N}\Sigma_{j=1}^{M}(d_{ij} - f_{ij}]^2]\}$, where d_{ij} is pixel gray value (averaged over color) for data-reduced image and f is its corresponding value for original image, summarized over $N \times M$ number of all pixels of frame.

greeting every increase in resolution, impressionist paintings look realistic only from a particular distance.

In Figure 10.5, we show what we promised in the introduction: the extraction of 3D information from a 2D image. We do that by showing two views — the original and one rotated by a computer using the known local properties of the catastrophes. These two images demonstrate the 3D extraction but also suggest a simple way to allow the human viewer to see the 2D scene in 3D, namely a stereo pair for the original scene.

10.3 DISCUSSION

The presented analytic modeling of image analysis/synthesis demonstrates that high data reduction is possible with only two primary elements — catastrophes. The primary elements are basic building blocks for any scene, any image, and any object. A question arises if such modeling can, in principle, be a basis for analysis of visual perception. In this context, we can observe that the ABC algorithm reduces membership of primary elements to an absolute minimum, an optimum situation from an informational point of view (a membership with only single primary element is rather an unrealistic scenario). Moreover, within the ABC system, entire 3D image geometry is reduced to the 2D intensity retina pattern, and such mapping is locally *isomorphic* (or even *homeomorphic*). This means that stereopsis would be a rather local phenomenon, well observable at larger distances. Indeed, some neurological studies of primary visual cortex have identified *ocular* cells that operate locally, up to 2° angular

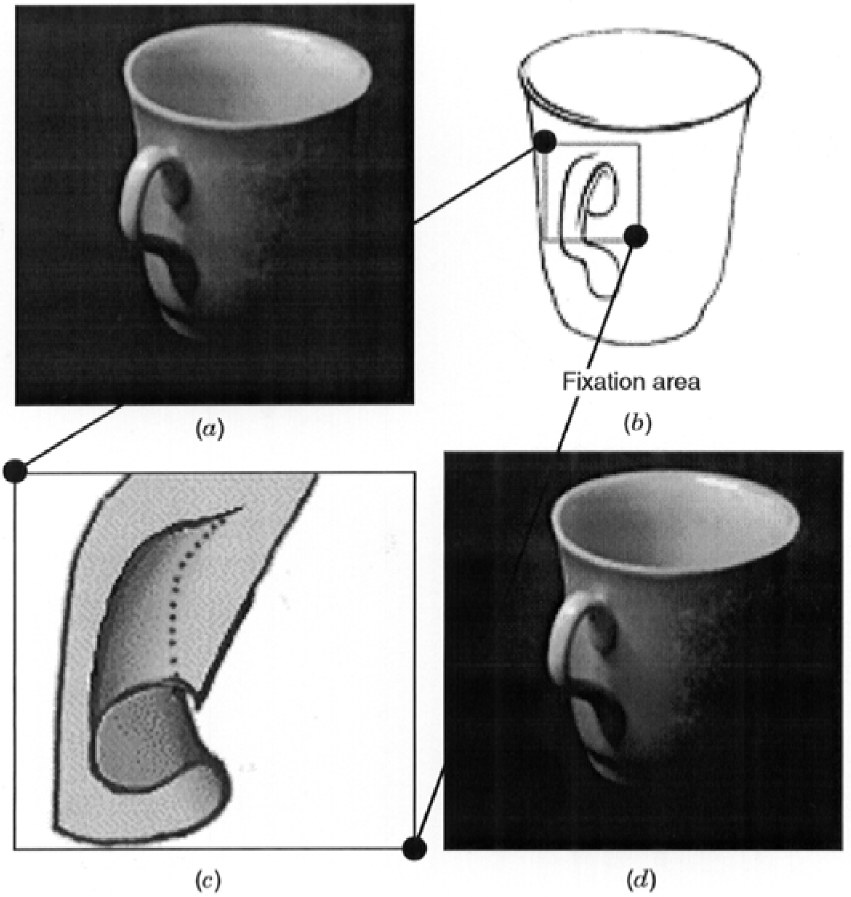


Figure 10.5 Demonstration of 3D nature of catastrophes on basis of simple photographic object, a *cup*, including (*a*) data-reduced cup (DR = 20 : 1); (*b*) cup's fixation area: an ear; (*c*) detailed illustration of catastrophes as 3D objects; *cusp* area is shown only from one side; second hidden side is shown as broken line; (*d*) using 3D profile of both catastrophes, a cup, coded as in Fig. 10.2*a*, has been automatically rotated by 2°, thus demonstrating local 3D dimensionality of monoscopic image in catastrophic representation (Fig. 10.2*a*).

separation [1, p. 147]. Also, the geometric structure of the visual cortex is vertically uniform [1, p. 112]. The ABC system operates with two stable catastrophes — *cusp* and *fold* — which are 3D in nature but with retina mapping that is 2D *soft-edge* intensity distribution, the latter replacing the third dimension (see Fig. 10.1). Therefore, we should search for specialized line cells, corner cells, or rather their groupings, representing the catastrophes' 3D nature in the form of 2D intensity distribution, such as in Figures 10.1*c* and 10.1*d*.

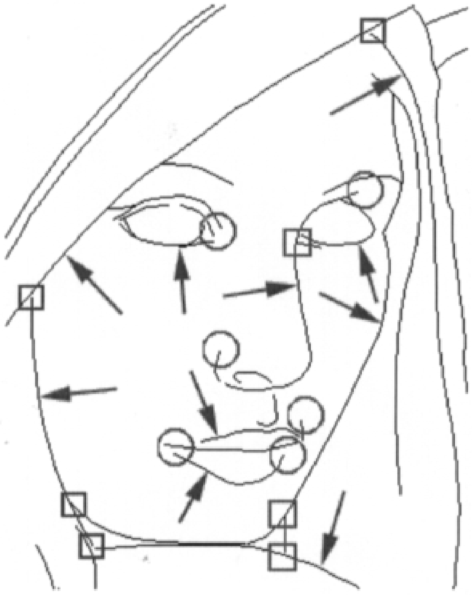


Figure 10.6 Demonstration of ABC's fixation area, roughly equivalent of saccadic points in primary vision. We see that all fixation areas are also primary element areas: fold (arrows) and cusp (circles). Also, "T-shapes" (squares) can be identified; once, as a combination of two independent fold's (red color) and the second as reconstruction of unstable catastrophe called "swallowtail" and related to cusp (blue color). In general, all fold-related and cusp-related features are in blue and red, respectively.

Pinwheel cell groupings [13–15] should be such characteristic groupings for cusp detection (in fact, one of the block elements of the ABC algorithm for cusp detection is similar to pinwheel grouping of logic elements). It should be emphasized that it is not the line itself but rather the characteristic parabolic asymmetric intensity distribution in the vicinity of this line (see Fig. 10.1) that makes this structure a *fold* catastrophe. Similarly, the end of the line [1, pp. 81–85] can only be anticipated as a *cusp*, assuming that we will be able to see a complimentary view from the other side. At the same time, a *T shape* can be interpreted as a combination of two folds as shown in Figure 10.6. In addition, if a *cube* consists of hard edges, then there is an ambiguity between 3D and 2D interpretation, as shown in Ref. 16, Fig. 1*b*. This ambiguity will be canceled, however, if we will be able to draw soft-edge-style or photograph an object with *diffuse* illumination (to avoid hard shadowing that arises as a result of direct, or collimated, illumination). Our every-day experience tells us that we have a tendency to interpret line termination as a *cusp* and line connection between cusps, as well as an external object line, as a *fold*, as shown in Figure 10.6. Good artists understood this a long time ago (see Ref. 17, Fig. 1). Yet, only an asymmetric soft edge in the form of a square root type intensity dependence [7], is a true fold. Therefore, such anticipation can sometimes be wrong, leading to familiar ambiguities and/or illusions. On the other hand, catastrophe mapping can be extremely effective, as illustrated by the example of *Atteave's cat* where very primitive catastrophic graphics is sufficient for cat visualization.

In summary, we have demonstrated that catastrophe-based ABC is a possible model for vision perception, since we could not find any contradiction with equivalent neurobiological results, while the following features of the visual cortex seem to agree with the ABC system: foveation search with singular fixation points; local matching corresponding to parts of two stereoscopic retina images; local 3D features of monoscopic images; 2D structure of visual cortex [1, p. 112]; high data reduction [18]; modular and modestly parallel visual cortex architecture [1, pp. 99–100]; highly nonlinear and hierarchic feature extraction [1, pp. 99–100; 18, 19] leading to neural net models [19, 20]; highly effective pattern recognition, highly independent illumination, shadowing, color, orientation, and scale; and finally, excellent image quality reconstruction (synthesis) from highly disperse singular elements.

It should be noted that a total number of singularities have been determined by only two factors: scale, represented by a values, and their possible types (here are only two: *cusp* and *fold*). The remaining part of the algorithm, including the reconstruction (synthesis) of a scene, has been done automatically, by algorithmic computing.

Humans are known to have the ability to visualize rotated scenes. Analysis by catastrophe shows how this can be done quite simply and accurately. We know humans attend to the foreground before the background and ABC shows how this can be done automatically. Foveation based on catastrophes wherein the information gathered is the type, size, location, and orientation of catastrophes gives enough information for an accurate scene reconstruction. This is almost certainly not what the brain does with that information, but it shows information sufficient for essentially any task. It seems more likely that the brain classifies objects syntactically using the catastrophes, their parameters, and the spatial relationships among them.

Although it seems unlikely that humans use a digital algorithm fully equivalent to ABC, the utility of a neural algorithm performing essentially the same operations would be obvious and would account very economically for foveation, the ability to pair features from different viewpoints in stereo and optic flow, the ability to segregate objects and see them in 3D even without stereo, and the remarkable ability of humans to recognize in a scene what is there even with a very slow computer. The neurological feature detectors could be there for the two catastrophe types as we have shown. Private conversations with H. H. Szu [21] suggest that there are plausible neural algorithms for finding the "independent components" of a scene, which seem to fall into classes very like the two catastrophe types shown here. Finally, ABC makes sense out of what previously seemed counterintuitive concerning the role of edges in images. That is, edges can "be everything (even) if it is so difficult to agree with this" [ref. 1, p. 87].

10.4 METHODOLOGY

The work reported here is primarily algorithmic. It generates but does not resolve numerous psychophysical questions. Are soft boundaries and their locales critical to stereopsis? How are illusions accounted for? From results in this chapter, we see that graphical lines with no soft locality (as in the Necker cube) and photographs with bright specular lighting (and thus hard shadowing) can cause ambiguity and thus illusions. But natural lighting of real scenes gives both hard edges and their “soft surround,” from which we can deduce the needed catastrophic information for correct and unambiguous scene interpretation.

We developed the ABC concept somewhat earlier as a means for image compression and MPEG-4-like integration of graphical and synthetic images [22]. Our interest in human visual perception arose when we observed that actual foveation points corresponded closely to catastrophe locations [23]. Catastrophes have also been observed in other natural phenomena [24].

In this chapter, only stationary objects are discussed. For moving objects, the number of catastrophes needed increases to 14 if we include the other 12 3D-to-2D mapping singularities. The discussion of these complications seemed to us a distraction in an introductory work.

Likewise, we have chosen to omit discussions of, for example, texture and color. The ABC theory extends to these areas, but it seemed better to limit this chapter to the simplest case of a stationary, smooth, black-and-white image. In what we show here, color images are used for ABC algorithm choice of a.

To derive formula 10.2 from Eq. 10.1, we apply Arnold’s nonlinear, purely geometric local transformation [12], generalized into a photogeometric domain, obtained by introducing a new photometric coordinate, luminance B. First, in the search of singular terms, we expand $W = F(\xi, \eta, B)$ into a local infinite Taylor series in terms of (ξ, η) and B in the vicinity of singular point (ξ_0, η_0, B_0). In strict analogy to Thom’s lemma [10], proven by Arnold et al. [12], we use reduction to the normal form procedure [12], which includes nonlinear substitution of coordinates. Linear and quadratic terms are obviously regular (Morse) ones. Therefore, we can prove that the luminance B physical coordinate does not introduce new singularities. As a result of this, the second term of Eq. 10.2 does not contain B. This procedure allows one to determine the local singular term (II) in a unique way, allowing for isomorphic reconstruction.

The ABC algorithm has been based on ABC algorithmic theory, briefly discussed above; the singular position has been found by least squares method application, with an a priori defined a scale value. Therefore, each singularity has been defined as a hierarchical data digital stream, such as, *type of singularity; location; angular position;* and *value.* As a result, a whole scene has been digitized in a hierarchical way. The connecting points have been approximated by planes and the image synthesis was still very good.

ACKNOWLEDGMENTS

The work of I. T and T. J. was partially supported by the National Institute of Standards and Technology, and the work of H. J. C. was partially supported by the National Aeronautics and Space Administration.

REFERENCES

1. D. H. Hubel, *Eye, Brain, and Vision*, Scientific American Library, New York, 1995, p. 80.
2. B. Fisher and B. Breitmayer, "Mechanisms of visual attention revealed by saccadic eye movements," *Neuropsychologia* **25**, 73–82 (1987).
3. B. Farrel, "Two-dimensional matches from one-dimensional stimulus components in human stereopsis," *Nature*, **395**, 688–693 (1998).
4. J. J. Koenderink and A. J. Van Dorn, "Geometry of binocular vision and model of stereopsis *Biol. Cybernet.* **21**, 29–35 (1976).
5. M., Born and E. Wolf, *Principles of Optics*, Pergamon, New York, 1980.
6. J. W. Goodman, *Introduction to Fourier Optics*, McGraw-Hill, New York, 1988.
7. H. G., Barron and J. M. Tenenbaum, "Interpreting line drawings as three dimensional surfaces," *Artificial Intelligence* **17**, 75–116 (1981).
8. W. L. Hansen, *Geometry in Nature*, A. K. Peters, Wellesley, MA, 1993.
9. J. L., Koenderink and A. J. Van Dorn, "The singularities of the visual mapping," *Biol. Cybernet.* **24**, 51–59 (1976).
10. R. Thom, "Topologic models in biology," *Topology* **8**, 313–335 (1969).
11. H. Whitney, "On singularities of mapping in Euclidean spaces," *Ann. Math.* **62**, 374–410 (1955).
12. V. I. Arnold, S. M., Guisein-Zade, and A. N. Varchenko, *Singularities of Differentiable Maps*, Birkhåuser, New York, 1985.
13. T. Bonhoefter and A., Grinwald, "Iso-orientation domains in cat visual cortex are arranged in pinwheel-like patterns," *Nature* **353**, 429–431 (1991).
14. A. Das and C. D. Gilbert, "Topography of contextual modulations mediated by short-range interaction in primary visual cortex," *Nature* **399**, 655–661 (1999).
15. U. Eysel, "Turning a corner in vision research," *Nature* **399**, 641–644 (1999).
16. J., Sun and P. Perona, "Early computation of shape and reflectance in the visual system," *Nature* **379**, 165–168 (1996).
17. P. Sinha and T. Poggio, "Role of learning in three-dimensional form perception," *Nature* **384**, 460–463 (1996).
18. D. C. Van Essen, C. H., Anderson, and D. J. Felleman, "Information processing in the primary visual system. An integrated system perspective," *Science* **255**, 419–422 (1992).
19. T. Poggio and S. Edelman, "A network that learns to recognize three-dimensional objects," *Nature* **343**, 263–266 (1990).

20. S. Grossberg., "3-D vision and figure-ground separation by visual cortex," *Percept. Psychophys.* **55**, 48–53 (1994).

21. H. H. Szu, Naval Surface Warfare Ctr., Dahlgren, VA, private communication, August 1999.

22. I. V. Ternovskiy and T. Jannson, "Singular manifold extraction as a novel approach to image decomposition," *SPIE Proc.* **3455**, 209–213 (1998).

23. T., Jannson, A., Kostrzewski, and I. V. Ternovskiy, "Fuzzy logic genetic algorithm for hypercompression," *SPIE Proc.* **3165**, 298–306 (1997).

24. J. F. Nye, "Rainbow scattering from spherical drops—An explanation of the hyperbolic unbilic foci," *Nature* **316**, 543 (1984).

INDEX