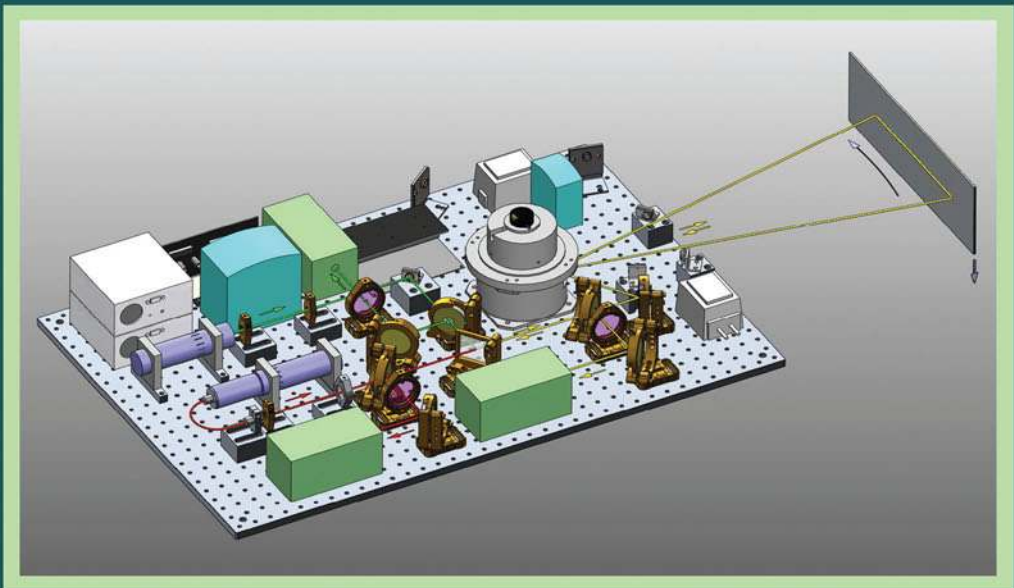# Handbook of Optical and Laser Scanning

Second Edition

Edited by
Gerald F. Marshall and Glenn E. Stutz

Handbook of Optical and Laser Scanning

Second Edition

OPTICAL SCIENCE AND ENGINEERING

Founding Editor
Brian J. Thompson
University of Rochester
Rochester, New York

RECENTLY PUBLISHED

*Please visit our website **www.crcpress.com** for a full list of titles*

Handbook of Optical and Laser Scanning

Second Edition

Edited by
Gerald F. Marshall and Glenn E. Stutz

In gratitude to God for my career and my jobs that I've enjoyed,

such that I didn't have to work!

Gerald F. Marshall

To my wife Christine and daughter Erica

To the memory of my parents Ed and Eileen

Glenn E. Stutz

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

Contents

Preface

Optical and laser scanning is the controlled deflection of light, visible or invisible. The aim of *Handbook of Optical and Laser Scanning, Second Edition*, is to provide engineers, scientists, managerial technologists, and students with a resource to be used as a reference for understanding the fundamentals of optical scanning technology. This text has evolved from three previous books, *Laser Beam Scanning* (1985), *Optical Scanning* (1991), and *Handbook of Optical and Laser Scanning* (2004). Since their publication, many advances have occurred in optical scanning, requiring updating of previous material and introduction of additional scanning technologies. This new edition also adds a few chapters on scanning applications illustrating the practical use of scanning technology.

Optical and laser scanning is a topic that is extremely broad in scope. It encompasses the mechanisms that control the deflection of light, optical systems that work with these mechanisms to perform scanning functions, and factors that affect the fidelity of the images generated or obtained from the scanning systems. Each of these subtopics is addressed in this book from a variety of perspectives.

A scanning system can be an input or output system or a combination of both. Input systems acquire images in either two or three dimensions. These systems can operate at a fixed wavelength or over a broad spectrum. They can reacquire the original light source by gathering either the specular or diffuse reflection or by fluorescing the image and acquiring the fluoresced light. Output systems direct light to produce images for applications such as marking, visual projection, and hard copy output. Ladar and many inspection systems use the same optical path to both illuminate the scene and acquire the image. A scan system requires not only optics but disciplines such as mechanics, electronics, magnetics, fluid dynamics, material science, acoustics, image analysis, firmware, software, and a host of others. This book brings together the knowledge and experience of 26 authors from England, Japan, and the United States.

The continuous and rapid changes in technological developments preclude the publication of a definitive book on optical and laser scanning. The contributors have accomplished their tasks painstakingly well, and each could have written a volume on his own particular subject. This book can be used as an introduction to the field and as a reference for persons involved in any aspect of optical and laser beam scanning.

Chapters 1 through 3 cover three basic scanning systems topics: Gaussian laser beam characterization, optical systems for laser scanners, and scanned image quality. Chapters 4 through 7 cover aspects of monogonal (single mirror facet) and polygonal scanning system design, including bearings. Chapters 8 and 9 discuss aspects of galvanometric and resonant scanning systems, including flexure pivots. Chapters 10 through 12 cover holographic, acousto-optical and electro-optical scanning systems. Chapters 13 and 14 cover piezoelectric scanners and scanning of optical disks. Chapters 15 and 16 cover two applications of optical scanning technology namely computer to plate (CTP), and underwater scanning. These chapters have been inserted to illustrate the significance of scanning in society today.

Gerald F. Marshall
Glenn E. Stutz

Preface to Laser Beam Scanning (1985)

To the memory of my parents Albert and Ethelena.

The aim of this volume is to provide engineers, scientists, and students with a guideline to the fundamentals of laser beam scanning. It brings together the knowledge and experience of seven specialists in the field, from England, Germany, Scotland, and the United States.

   The book covers the recently developed holographic scanners, the well-established polygonal scanners, and the galvanometric and resonant scanners. It includes complementary chapters on gas bearings for rotating scanning devices, the aerodynamic considerations of polygonal scanners, Gaussian laser beam diameters, and the optical design of components and systems relating to data storage on optical disks.

Gerald F. Marshall

Preface to Optical Scanning (1991)

To Irene, my wife; children, Clare Margaret and Mark Peter; Guy Nicholas and Maria Elizabeth, with love.

The aim of this volume is to provide application-oriented engineers and technologists, scientists, and students, with a guideline and a reference to the fundamentals of input and output optical scanning technology and engineering. It brings together the knowledge and experience of 16 international specialists from England, Japan, Scotland, and the United States. Brief biographies of the contributors are included. The Foreword and Afterword by Leo Beiser unify the selected topics of the 13 chapters, and give an overview evaluation of the technologies within the field of optical scanning engineering. Optical scanning technology is a comprehensive subject that encompasses not only the mechanics of controlling the deflection of a light beam but also all aspects that affect the imaging fidelity of the output data that may be displayed on a screen or recorded on paper.

A scanning system may be an input scanner, an output scanner, or a scanner that combines both of these functional attributes. A system's imaging fidelity depends on, and begins with, the reading of the input information and ends with the writing of the output data. Optical scanning intimately involves a number of disciplines: optics, material science, magnetics, acoustics, mechanics, electronics, and image analysis, with a host of considerations.

The book covers Gaussian laser beam diameters and divergence, optical and lens design for scanning systems, and scanned image quality. It deals with rotary scanning devices and systems, namely, holographic scanners for bar code readers and graphic arts, polygonal scanners, windage (i.e., the aerodynamic aspects), bearings, motors, and control systems associated with high-performance polygonal scanners. *Optical Scanning* treats oscillatory devices and systems; specifically, galvanometric and resonant low-inertia scanners, acousto-optical, and electro-optical scanners, and modulators. It closes with optical disk scanning technology.

The dream is to produce a definitive book on optical scanning, but this is an impossible task to accomplish in this ever more rapidly changing era of technological developments. All the authors have done his best; each of them could have written a volume on his own special subject. The book is complete as an introduction to the field. With the common thread of the subject title, the disparate chapters are brought into perspective in the Afterword.

To assist the reader, measured quantities are expressed in dual units wherever possible and appropriate; the secondary units are in parentheses. The metric system takes precedence over other systems of units, except where it just does not make good sense.

A strong effort has been made for a measure of uniformity in the book with respect to terminology, nomenclature, and symbology. However, with the variety of individual styles of the 16 contributing authors who are scattered across the Northern Hemisphere, I have placed greater importance on the unique contributions of the authors rather than on form.

I extend my thanks to the following persons: Brian J. Thompson, Provost of the University of Rochester, for his patient confidence in inviting me to produce this additional volume

on the subject of scanning in this series; my 16 contributing co-authors for their splendid material; and the reviewers of the manuscripts and typescripts, namely,

Robert Basanese	Rofin-Sinar, Inc.
Leo Beiser	Leo Beiser, Inc.
John H. Carosella	Speedring Systems, Inc.
Duane Grant	IBM Corporation
Michael J. Hayford	Optical Research Associates
Ron Hooper	Hooper Engineering Company
Charles S. Ih	University of Delaware
David B. Kay	Eastman Kodak Company
Kathryn A. McCarthy	Tufts University
Robert J. Schiesser	Charles Stark Draper Lab, Inc.
David Strand	Energy Conversion Devices, Inc.
William Taylor	Kollmorgan Corporation
Stanley W. Thomas	Lawrence Livermore Laboratory
Daniel Vukobratavich	University of Arizona
David L. Wright	Spectra-Physics Lasers, Inc.
Francis Yu	The Pennsylvania State University
Ross Zelesnick	RCA, Inc.

Each gave his or her time to critique a script, made helpful comments, and provided excellent suggestions. I thank John H. Carosella of Speedring Systems, Inc., for his indirect support, which I much appreciate. I am also grateful for the generous help and time given, especially in proofreading and organizing the index, by my wife, Irene.

I am pleased to be the coordinator of these works and value the privilege of being the one to share this treatise with my colleagues in the field.

Read, scan, study, and enjoy.

Gerald F. Marshall

Preface to Handbook of Optical and Laser Scanning (2004)

With gratitude to my wife, Irene, colleagues, and friends.

To the memory of my parents, Ethelena and Albert, brothers, Donald and Edward and sisters, Andrée and Kathleen.

Optical and laser beam scanning is the controlled deflection of a light beam, visible or invisible. The aim of *Handbook of Optical and Laser Scanning* is to provide application-oriented engineers, managerial technologists, scientists, and students with a guideline and a reference to the fundamentals of input and output optical scanning technology and engineering. This text has its origin in two previous books, *Laser Beam Scanning* (1985) and *Optical Scanning* (1991). Since their publication, many advances have occurred, which has made it necessary to update and include the changes of the past decade. This book brings together the knowledge and experience of 27 international specialists from England, Japan, and the United States.

Optical and laser scanning technology is a comprehensive subject that encompasses not only the mechanics of controlling the deflection of a light beam, but also all aspects that affect the imaging fidelity of the output data that may be recorded on paper or film, displayed on a monitor, or projected onto a screen. A scanning system may be an input scanner, an output scanner, or one that combines both of these functional attributes. A system's imaging fidelity begins with, and depends on, the accurate reading and storage of the input information—the processing of the stored information—and ends with the presentation of the output data. Optical scanning intimately involves a number of disciplines: optics, material science, magnetics, acoustics, mechanics, electronics, and image analysis, with a host of considerations.

The continuous and rapid changes in technological developments preclude the publication of a definite book on optical and laser scanning. The contributors have accomplished their tasks painstakingly well, and each could have written a volume on his own particular subject. This book can be used as an introduction to the field and as an invaluable reference for persons involved in any aspect of optical and laser beam scanning.

To assist the international scientific and engineering readership, measured quantities are expressed in dual units wherever possible and appropriate; the secondary units are in parentheses. The metric system takes precedence over other systems of units, except where it does not make good sense. A serious effort has been made for a measure of uniformity throughout the book with respect to terminology, nomenclature, and symbology. However, with the variety of individual styles from 27 contributing authors who are scattered across the Northern Hemisphere, I have placed greater importance on the unique contributions of the authors than on form.

The chapters are arranged in a logical order beginning with the laser light source and ending with a glossary. Chapters 1 through 3 cover three basic scanning systems topics: Gaussian laser beam characterization, optical systems for laser scanners, and scanned image quality. Chapters 4 through 7 cover aspects of monogonal (single mirror-facet) and polygonal scanning system design, including bearings. Chapters 8 and 9 discuss aspects

of galvanometric and resonant scanning systems, including flexure pivots. Chapters 10 through 14 cover holographic, optical disk, acousto-optical, electro-optical scanning systems, and thermal printhead technology. A useful glossary of scanner terminology follows Chapter 14.

Gerald F. Marshall

Cover Image

Start of scan detector

Detector 2

End of scan detector

Polygon scanner

Telescope 2

Target plane

Laser 2
λ = 532 nm

Kinematic mirror

Detector 3

Laser 1
Fiber coupled
λ = 635 nm

Dichroic beam combiner

Detector 1

Telescope 1

Acknowledgments

Editors

Gerald F. Marshall

Gerald F. Marshall is a retired Consultant in Optical Design and Engineering, specializing in optical scanning and display systems; he resides in Niles, Michigan. His extensive experience includes senior positions with Kaiser Electronics, San Jose, California; Energy Conversion Devices and Axsys Technologies (formerly Speedring Systems), both in Rochester Hills, Michigan; Medical Lasers, Burlington and Diffraction Limited, Bedford, Massachusetts. Previously he was engaged as a Senior Research and Development Engineer for airborne navigational display systems at BAC, plc. (formerly Ferranti Ltd.), Edinburgh, Scotland, and earlier as a physicist with Morganite International Ltd., London, England. The author of many papers, he holds a number of patents and is the editor and contributor of three internationally recognized reference books, *Laser Beam Scanning* (Marcel Dekker, 1985)*, Optical Scanning* (Marcel Dekker, 1991), and the first edition of *Handbook of Optical Scanning* (Marcel Dekker, 2004), on which subjects he regularly gave short courses for the SPIE-The International Society for Optical Engineering and at the University of Wisconsin-Madison. He is a Fellow of The Institute of Physics (IOP), The Optical Society of America (OSA), and The SPIE, of which he is a former director (1991–93). He received a BSc degree from the University of London, England.

Glenn E. Stutz

Glenn E. Stutz is the COO and CTO for Lincoln Laser Company. He has spent more than 25 years in the laser scanning business. He has been involved in the design and manufacturing transition of systems for retinal scanning, agricultural inspection, film printing, large scale metrology, laser projection, microscopy, and pwb inspection. Prior to joining Lincoln Laser Company, he was involved with high energy lasers while with TRW. Stutz holds a BS degree in Optics from the University of Rochester and an MS in Optical Sciences from the University of Arizona. He also holds an MBA from Arizona State University.

Contributors

Andreas Blume, BS, Dipl. Ing. (FH)
Piezosystem Jena, Inc.
Hopedale, Massachusetts, USA

David C. Brown, PhD
Cambridge Technology Inc.
Lexington, Massachusetts, USA

Frank M. Caimi, BSc, PhD
Ocean Visibility and Optics
    Laboratory
Harbor Branch Oceanographic Institute at
    Florida Atlantic University
Fort Pierce, Florida, USA

Carl E. Conti
Consultant
Hammondsport, New York, USA

**Fraser R. Dalgleish, B.Eng (Hons),
MSc, PhD**
Ocean Visibility and Optics Laboratory
Harbor Branch Oceanographic Institute at
    Florida Atlantic University
Florida, USA

Timothy K. Deis, BS, MS
Consultant
Pittsburgh, Pennsylvania, USA

LeRoy D. Dickson, PhD
Wasatch Photonics, Inc.,
Logan, Utah, USA

Emery Erdelyi, BSEE
Axsys Technologies, Inc.
San Diego, California, USA

Chris Gerrard, BSc
Westwind Air Bearings Division,
    a GSI Group company
Poole, Dorset, United Kingdom

Timothy A. Good, MS
Metrologic Instruments, Inc.
Blackwood, New Jersey, USA

Milton Gottlieb, BS, MS, PhD
Consultant, Carnegie Mellon University
Pittsburgh, Pennsylvania, USA

Thomas F. Johnston, Jr., PhD
Optical Physics Solutions
Grass Valley, California, USA

Donald R. Lehmbeck, BS, MS
Xerox Corporation
Webster, New York, USA (retired)
College of Imaging Arts and Sciences,
    Rochester Institute of Technology
Rochester, New York, USA (Adjunct
    Faculty)
Imaging Quality Technology Consulting
Penfield, New York, USA
Torrey Pines Research
Fairport, New York, USA

James Litynski, BS
Piezosystem Jena, Inc.
Hopedale, Massachusetts, USA

Gerald F. Marshall, BSc, F.Inst.P.
Consultant in Optics
Niles, Michigan, USA

Jean Montagu, MS
Engineering Consultant
Cambridge, Massachusetts, USA

Mark T. Montgomery, BS, MS
SkyCross, Inc.
Viera, Florida, USA

Gregory Mueller, BS
MacDermid Printing Solutions
San Marcos, California, USA

Gerald A. Rynkowski
Axsys Technologies, Inc.
Rochester Hills, Michigan, USA

Stephen F. Sagan, MS
NeoOptics
Lexington, Massachusetts, USA

Tetsuo Saimi, MD
Matsushita Electric Industrial Co., Ltd
Kadoma, Osaka, Japan

Michael W. Sasnett, MSEE
Optical System Engineering
Los Altos, California, USA

Daniel D. Stancil, PhD
Carnegie Mellon University
Pittsburgh, Pennsylvania, USA

Glenn E. Stutz, BS, MS, MBA
Lincoln Laser Company
Phoenix, Arizona, USA

John C. Urbach, BS, MS, PhD[†]
Portola Valley, California, USA

Reeder N. Ward, BS, MS
Noah Industries, Inc.
Melbourne, Florida, USA

[†] Deceased

1

Characterization of Laser Beams: The M^2 Model

Thomas F. Johnston, Jr.

Optical Physics Solutions
Grass Valley, California, USA

Michael W. Sasnett

Optical System Engineering
Los Altos, California, USA

CONTENTS

1.1   INTRODUCTION

The M^2 model is currently the preferred way of quantitatively describing a laser beam, including its propagation through free space and lenses; specifically, as ratios of its parameters with respect to the simplest theoretical gaussian laser beam. The present chapter describes the model and measuring techniques for reliably determining—in each of the two orthogonal propagation planes—the key spatial parameters of a laser beam; namely, the beam waist diameter $2W_0$, the Rayleigh range z_R, the beam divergence Θ, and waist location z_0.

1.2 HISTORICAL DEVELOPMENT OF LASER-
BEAM CHARACTERIZATION

In 1966, six years after the first laser was demonstrated, a classic review paper[1] by Kogelnik and Li of Bell Telephone Laboratories was published, which served as the standard reference on the description of laser beams for many years. Here the $1/e^2$ diameter definition[1,2] for the width of the fundamental-mode gaussian beam was used. The more complex transverse irradiance patterns, or transverse modes, of laser beams were identified with sets of eigen-function solutions to the wave equation, including diffraction, describing the electric fields of the beam modes. These solutions came in two forms: those with rectangular symmetry were described mathematically by Hermite–Gaussian functions, those with cylindrical sym-metry by Laguerre–Gaussian functions. So with the appropriate basis set, any beam could be decomposed into a weighted sum of the electric fields of these modes, at least in principle. Mathematically, for this expansion to be unique the phases of the electric fields must be known. This is difficult at optical frequencies. Irradiance measurements alone, where the phase information is lost in squaring the E-fields, does not allow determination of the expan-sion coefficients. This "in principle but not in practice" description of light beams was all that was available and seemed to be all that was needed for several succeeding years.

Workers often measured beam diameters by scanning an aperture across the beam to detect the transmitted power profile. Apertures used were pinholes, slits, or knife-edges, and the beam diameters were (and still are) defined based on the measurement effect that would be produced on a fundamental-mode beam. Commercial laser beams were speci-fied as being pure fundamental mode, the lowest order or zero–zero transverse electro-magnetic wave eigenfunction, "TEM_{00}."

In 1971, Marshall[3] published a short note introducing the M^2 factor, indicating M ($=\sqrt{M^2}$) as the multiplying factor by which the diameter of a beam is larger than that of the funda-mental mode of the same laser resonator. Marshall's interest lay with the effects produced by industrial lasers and since they depend on focused spot size, he pointed out that they depend on M^2. No discussion was given of how to measure M^2 and the concept languished thereafter for several years.

From the late 1970s and into the 1980s, Bastiaans,[4] Siegman,[5,6] and others developed the-ories of bundles of light rays at narrow angles to an axis based on the Fourier transform relationship between the irradiance and the spatial frequency (or ray-angle) distributions to account for the propagation of the bundle. Such a bundle of rays is a beam. The beam diameter was defined as the standard deviation of the irradiance distribution (now called the second-moment diameter, when multiplied by four), and the square of this diameter was shown to increase as the square of the propagation distance—an expansion law for the diameter of hyperbolic form. These theories could be tested by measuring just the beam's irradiance profile along the propagation path.

In about 1987, one of us designed a telescope to locate a beam waist for an industrial CO_2 laser at a particular place in the external optical system. The design was based on measure-ments showing where the input beam waist was located and on blind faith that the laser datasheet claim for a "TEM_{00}" beam was correct. This telescope provided nothing like the expected result. Out of despair and disorientation came the energy to make more beam measurements and from these measurements came the realization that the factor that lim-ited the maximum distance between the telescope and the beam waist it produced was exactly the same factor by which actual focus-spot diameter at the work surface exceeded

the calculated TEM_{00} spot diameter. That factor was M^2 and when used in modified Kogelnik and Li equations, design of optical systems for multimode beams became possible.[7] This ignited some interest in knowing more about laser beams than had previously been considered sufficient. Laser datasheets that claimed "TEM_{00}" were no longer adequate.

In the 1980s, commercial profilers[8] reporting a beam's $1/e^2$ diameter became ubiquitous. By the end of the 1980s, experience with commercial profilers and these theories converged with the development[6] of the theoretical M^2 model and a commercial instrument[9] to measure the beam quality based on it, which first became available in 1990. The time to determine a beam's M^2 value dropped from half a day to half a minute.

With high accuracy M^2 measurements more readily available in the early 1990s, the reporting of a beam's M^2 value became commonplace, and commercial lasers with good beams were now specified[10] as having $M^2 < 1.1$. The International Organization for Standards began committee meetings to define standards for the spatial characterization of laser beams, ultimately deciding on the beam quality M^2 value based on the second-moment diameter as the standard.[11] This diameter definition has the best theoretical support, in the form of the Fourier transform theories of the 1980s, but suffers from being sensitive to noise on the profile signal, which often makes the measured diameters unreliable.[12,13] That led to the development in 1993 of rules[14] to convert diameters measured with the more forgiving methods into second-moment diameters for a large class of beams.

The M^2 model as commercially implemented does not cover beams that twist as they propagate in space, those with general astigmatism.[15,16] The earlier Fourier transform theories and their more recent extensions do, however, and allow for ten constants[17] needed to fully characterize a beam (adding to the six used in the M^2 model). Recently, in 2001, the first natural beam[18] (as opposed to a test beam artificially constructed) was measured by Nemes et al. that required all ten constants for its complete description.

Several recommendations can be made for characterizing a beam. Model the beam only to the level of complexity appropriate to your needs: three constants suffice if the beam spot is round at all propagation distances; six constants cover beams with simple astigmatism, divergence asymmetry, or waist asymmetry; ten constants are needed for beams with elliptical spots whose orientation twist in space (general astigmatism). Measure your beams with a reliable method, and when required, convert those values at the end into ISO standard units. Lastly, stay appraised of developments in instrumentation that may meet your need with more convenience, speed, and accuracy.

1.3 ORGANIZATION OF THIS CHAPTER

Section 1.2 provides an historical introduction to the field, outlining how the field developed to its present state.

The technical discussion begins in Section 1.4 by explaining the M^2 model. This mathematical model built around the quantity M^2 (variously called the beam quality, times-diffraction-limit number, or the beam propagation factor) describes the real, multimode beams that lasers produce and how their properties change when propagating in free space.

This discussion is continued in Section 1.5 covering the transformation of a beam through a lens. Section 1.6 explains the different methods used to define and measure beam diameters, and how measurements made with one method can be converted into the values measured with one of the other methods. This includes the standard diameter definition

adopted by the International Organization for Standardization (ISO), the second-moment diameter, and the experimental difficulties encountered with this method.

The technical development continues in Section 1.7 where the logic and precautions needed in measuring the beam quality M^2 are presented. Thoroughly discussed is the "four-cuts" method (a cut is a measurement of a beam diameter), the simplest way to obtain an accurate M^2 value.

Section 1.8 discusses the common and possible types of beam asymmetry that may be encountered in three dimensions when the propagation constants for the two orthogonal (and usually independent) propagation planes are combined. The concept of the "equivalent cylindrical beam" is introduced to complete the technical development of the M^2 model. Propagation plots for beams with combinations of asymmetries are illustrated. A short discussion follows of "twisted beams," those with general astigmatism, which are not covered in the M^2 model, and require a beam matrix of ten moments of second order for their complete description. This second-order beam matrix theory is a part of the underpinnings of the ISO's choice of the noise-sensitive second-moment diameter as the "standard."

Section 1.9 applies the M^2 model to an analysis of a stereolithography laser-scanning system. Using results of earlier sections, by working backward from assumed perturbations or defects in the scanned beam at the work surface, the deviations in beam constants at the laser head that would produce them are found. An overview of the M^2 model, in Section 1.10, concludes the text.

A glossary follows explaining the technical terms used in the field, with the references ending the chapter.

1.4 THE M^2 MODEL FOR MIXED-MODE BEAMS

In laser beam-scanning applications, the main concern is having knowledge of the beam spot-size—the transverse dimensions of the beam—at any point along the beam path. The mixed-mode ($M^2 > 1$) propagation equations are derived as extensions of those for the fundamental mode, so pure modes and particularly the fundamental mode are the starting point.

1.4.1 Pure Transverse Modes: The Hermite–Gaussian and Laguerre–Gaussian Functions

Lasers emit beams in a variety of characteristic patterns or transverse modes that can occur as a pure single mode or more often, as a mixture of several superposed pure modes. The transverse irradiance distribution of a pure mode is the square of the electric field amplitude versus the transverse distance from the beam axis, which when measured is termed a transverse profile. This amplitude is described mathematically by Hermite–Gaussian functions if it has rectangular symmetry, or by a Laguerre–Gaussian function if it has circular symmetry.[1,2,5,19] These functions when plotted reproduce the familiar spot patterns—the appearance of a beam on an inserted card—first photographed in Reference 20 and shown in References 1 and 19. Computed spot patterns are displayed here in Figure 1.1. The computations were done in Mathematica for the first six cylindrically symmetric modes, in order of increasing diffraction loss for a circular limiting aperture. These modes are the solutions to the wave equation for a bundle of rays propagating at small angles (paraxial rays) to the z-axis, under the influence of diffraction and are of the general forms[1,2,7]

$$U_{mn}(x,y,z) = H_m(x/w)H_n(y/w)u(x,y,z) \qquad (1.1a)$$

(a)   Lowest Diffraction Loss Modes   (b)   Lowest Diffraction Loss Modes
                      First three   Next three

                                                    00   10
              Cylindrical symmetry   Cylindrical symmetry
              Laguerre-Gauss modes   Laguerre-Gauss modes
                   TEM$_{pl}$ TEM$_{pl}$

Synthesis of   Synthesis of
starred modes   starred modes

    01   Rotate 01 by 90°   01*   03   Rotate 03 by 90°   03*

              +   =   +   =

    02   Rotate 02 by 90°   02*   11   Rotate 11 by 90°   11*

              +   =   +   =

    2w   2w   2w   2w   2w   2w
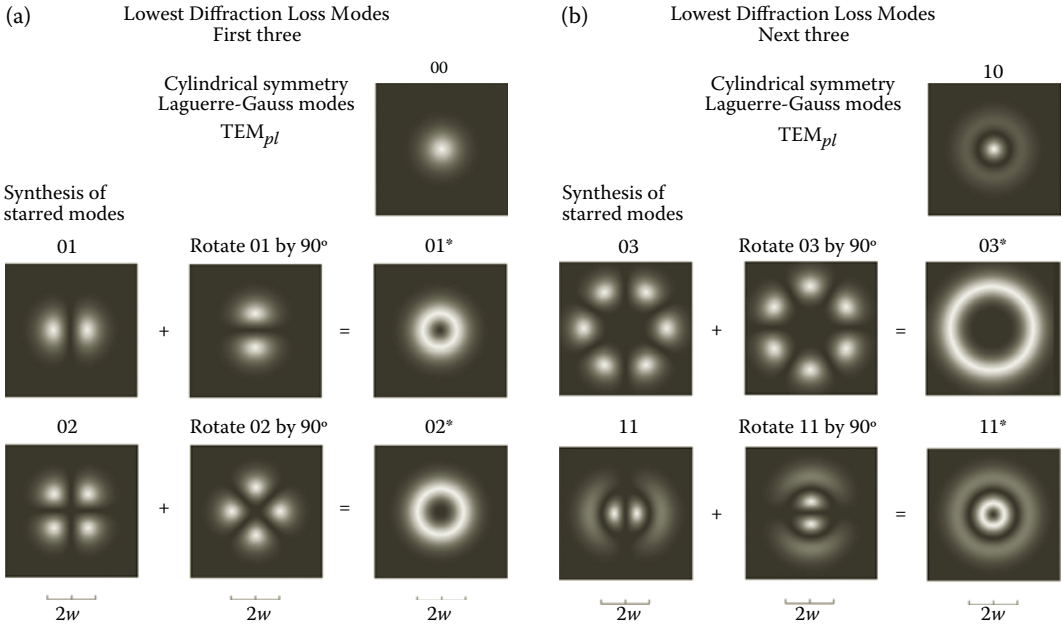
FIGURE 1.1
Computed spot patterns for cylindrically symmetric modes in order of increasing diffraction loss for a circular lim-
iting aperture. The subscript numbers *pl* above each image indicate the mode order. Starred modes are constructed
as shown, as the sum of a pattern with a copy of itself rotated by 90°: (a) First three modes. (b) Next three modes.

or

$$U_{pl}(r, j, z) = L_{pl}(r/w, j) u(r, z). \tag{1.1b}$$

In Equation 1.1a, $H_m(x/w)H_n(y/w)$ represents a pair of Hermite polynomials, one a function
of x/w, the other of y/w, where x, y are orthogonal transverse coordinates and w is the radial
scale parameter. In Equation 1.1b, $L_{pl}(r/w, \varphi)$ represents a generalized Laguerre polynomial,
a function of the r, φ transverse radial and angular coordinates. These polynomials have
no dependence on the propagation distance z other than through the dependence $w(z)$ in
x/w, y/w, or r/w. The $w(z)$ dependence describes the beam convergence or divergence. The
other function u is the gaussian

$$u = \left(\frac{2}{p}\right)^{1/2} \exp\left[\frac{-(x^2 + y^2)}{w^2}\right] = \left(\frac{2}{p}\right)^{1/2} \exp\left[\frac{-r^2}{w^2}\right]. \tag{1.2}$$

Because the radial gaussian function splits into a product of two gaussians, one a function
of x, the other of y, the Hermite–Gaussian function splits into the product of two functions,
one in x/w only and the other in y/w only, each of which is independently a solution to the
wave equation. This has the consequence that beams can have independent propagation
parameters in the two orthogonal planes (x, z) and (y, z).

These functions of the transverse space coordinates consist of a damping gaussian fac-
tor, limiting the beam diameter, times a modulating polynomial that pushes light energy
out radially as polynomial orders increase. The order numbers m, n of the Hermite poly-
nomials, or p, l of the Laguerre polynomial of the pure mode also determine the number
of nodes in the spot pattern, for which the modes are named. They are designated as

transverse electromagnetic modes, or $TEM_{m,n}$ for a mode with m nodes in the horizontal direction and n nodes in the vertical direction, or $TEM_{p,l}$ for a mode with p nodes in a radial direction—not counting the null at the center if there is one—and l nodes in going angularly around half of a circumference. Figure 1.2a through f, show the theoretical beam irradiance profiles for the six pure modes from Figure 1.1. Because these are the six lowest loss modes,[21,22] they are commonly found in real laser beams. The modes as shown all originate in the same resonator—they all have the same radial scale parameter $w(z)$. The addition of an asterisk to the mode designation—a "starred mode"—signifies a composite of two degenerate (same frequency) Hermite–Gaussian modes or as here, Laguerre–Gaussian modes in space and phase quadrature to form a mode of radial symmetry. This is explained in Reference 20, discussed in Reference 5, p. 689, and shown in Figure 1.1 for a mode pattern with an azimuthal variation ($l \neq 0$) as the addition of the mode with a copy of itself after a 90° rotation, to produce a smooth ring-shaped pattern.

The simplest mode is the TEM_{00} mode, also called the lowest order mode or fundamental mode of Figures 1.1 and 1.2a, and consists of a single spot with a gaussian profile (here L_{pl}
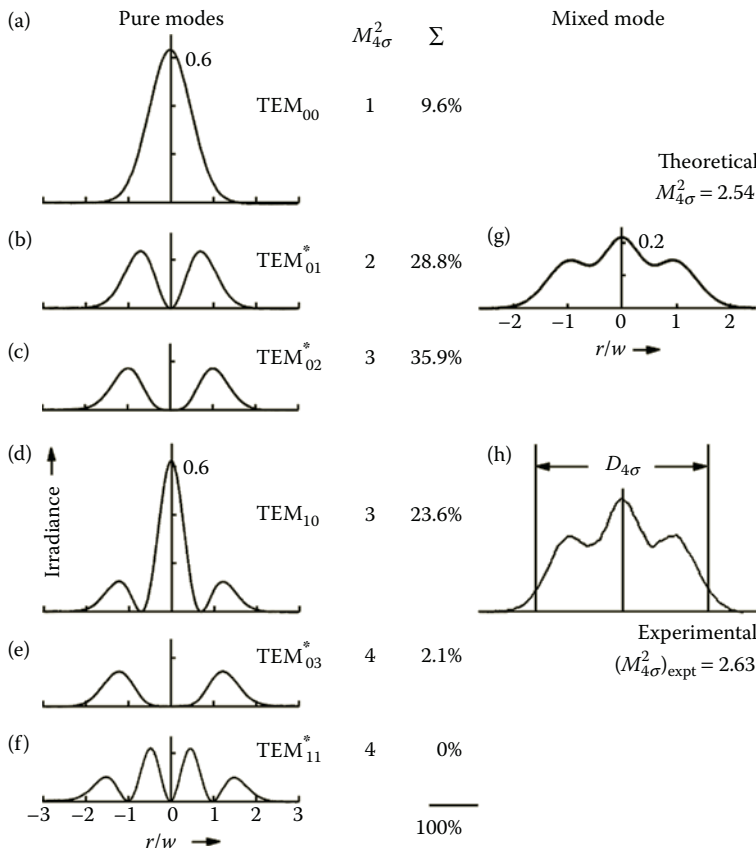
FIGURE 1.2

Synthesis of a mixed-mode as the weighted sum of pure modes. The theoretical pinhole profiles (a) to (f) for the six pure radial modes from Figure 1.1, shown in the first column, are summed with weighting fraction Σ of the third column to produce the mixed-mode profile (g). The beam qualities $M^2_{4\sigma}$ for each mode, in the second column, are similarly summed with weight Σ to produce the mixed-mode beam quality also shown in (g). The matching experimental pinhole profile is shown in (h).

is unity). The next higher-order mode has a single node (Figures 1.1 and 1.2b) and is appropriately called the "donut" mode, symbol TEM_{01}^*. The next two "starred" mode spots look like a donut with larger holes, the spot pattern of the TEM_{10} mode looks like a target with a bright center, and the TEM_{11}^* mode spot looks like a target with a dark center (Figures 1.1 and 1.2). All higher-order modes have a larger beam diameter than the fundamental mode. The six pure modes of Figure 1.2 are shown with the vertical scale normalized such that when integrated over the transverse plane, each contains unit power.

The physical reason that Hermite–Gaussian and Laguerre–Gaussian functions describe the transverse modes of laser beams is straightforward. Laser beams are generated in resonators by the constructive interference of waves multiply reflected back and forth along the beam axis. For this interference to be a maximum, permitting a large stored energy to saturate the available gain, the returned wave after a round trip of the resonator should match the transverse profile of the initial wave. The functions that do this are the eigenfunctions of the Fresnel–Kirchhoff integral equation used to calculate the propagation of a paraxial rays with diffraction included.[5,19] In other words, these are precisely the beam irradiance profiles that in propagating and diffracting maintain a self-similar profile, allowing after a round trip, maximum constructive interference and gain dominance.

1.4.2 Mixed Modes: The Incoherent Superposition of Pure Modes

While a laser may operate in a close approximation to a pure higher-order mode, for example, by a scratch or dust mote on a mirror forcing a node and suppressing a lower-order mode with an irradiance maximum at that location, actual lasers tend to operate with a mixture of several high-order modes oscillating simultaneously. The one major exception is lasing in the pure fundamental mode in a resonator with a circular limiting aperture, where the aperture diameter is critically adjusted to exclude the next higher-order (donut) mode. Each pure transverse mode has a unique frequency different from that for adjacent modes by tens or hundreds of MHz. This is usually beyond the response bandwidth of profile measuring instruments so any mode interference effects are invisible in such measurements.

Figure 1.2g shows a higher-order mode synthesized by mixing the five lowest order modes of Figure 1.2a through e in a sum with the weightings shown in the column labeled Σ. These weights—also called mode fractions—were chosen by a fitting program to match the result to the experimental pinhole profile (see Section 1.6.4.2) of Figure 1.2h. In the experiment[14] the number of transverse modes oscillating and their orders were known (by detecting the radio-frequency transverse mode beat notes in a fast photodiode). This information was used in the fitting procedure. The laser was a typical 1-m-long argon ion laser operating at a wavelength of 514 nm, except that a larger than normal intracavity limiting aperture diameter was used to produce this mode mixture.

Because the polynomials of Equation 1.1 have no explicit dependence on z, the profiles and widths of the modes in a mixture remain in the same ratio to each other and specifically to the fundamental mode as the beam propagates. This means that however the diameter $2W$ of a mixed-mode beam is defined (several alternatives are discussed in Section 1.6), if this diameter is M times larger than the fundamental-mode diameter at one propagation distance, it will remain so at any distance:

$$W(z) = Mw(z). \tag{1.3}$$

This equation introduces the convention that upper case letters are used for the attributes of high-order and mixed modes and lower case letters used for the underlying fundamental mode.

1.4.3 Properties of the Fundamental Mode Related to the Beam Diameter

The attributes of the simplest beam, a fundamental mode with a round spot (a cylindrically symmetric or stigmatic beam) are reviewed in Figures 1.3 and 1.4. The beam profile varies as the transverse irradiance distribution and is given by the function of gaussian form[1,2] (Figure 1.3a):

$$I\left(\frac{r}{w}\right) = I_0 \exp\left[-2\left(\frac{r}{w}\right)^2\right]. \tag{1.4}$$

The symbol I denotes a detector signal proportional to irradiance (and by using I instead of E, the recommended symbol for irradiance, avoids confusion with the electric field of the beam). The peak irradiance is I_0, and the radial scale parameter w introduced in Equation 1.1 can now be identified as the distance transverse to the beam axis at which the irradiance value falls to $1/e^2$ (13.5%) of the peak irradiance. This $1/e^2$ diameter definition, introduced[1,2] in the early 1960s, has been universally used since with one exception. (The one exception is in the field of biology where the fundamental-mode diameter is defined as the radial distance to drop to $1/e$ (36.8%) of the central peak value, making beams in biological references a diameter $2w' = \sqrt{2}w$ instead of $2w$.) Many different beam diameter definitions have been used subsequently for higher-order modes (these are discussed in Section 1.6) but they all share one common property: when applied to the fundamental-mode, they reduce to the traditional $1/e^2$ diameter.

Tables of the gaussian function are usually listed under the heading of the normal distribution, normal curve of error, or Gauss distribution and are of the form (see p. 763 of Reference 23)

$$I(x) = \left[\frac{1}{s(2p)^{1/2}}\right] \exp\left(\frac{-x^2}{2s^2}\right) \tag{1.5}$$
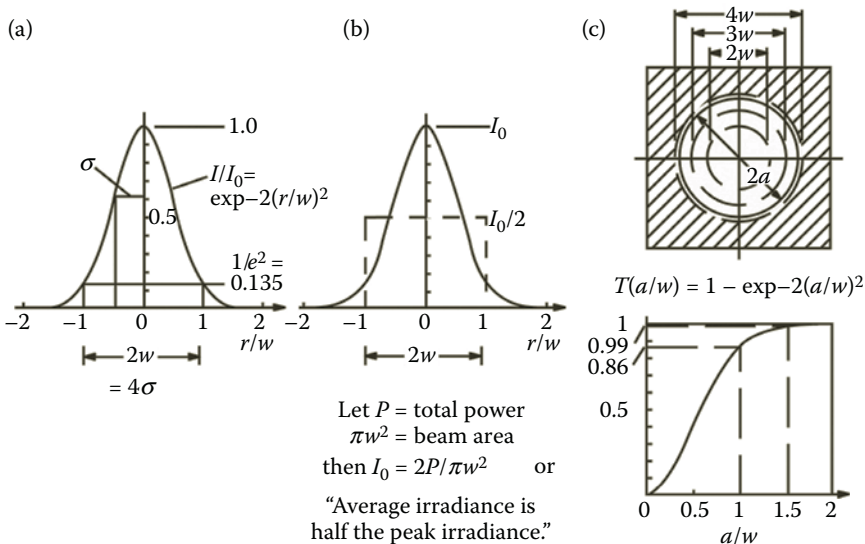
FIGURE 1.3
Properties of the fundamental mode related to the beam diameter, explained in the text; (a) definition of the $1/e^2$ diameter as the distance between the 13.5% levels on the pinhole profile; (b) relation between the peak irradiance and average irradiance; (c) transmission fraction through a circular aperture.
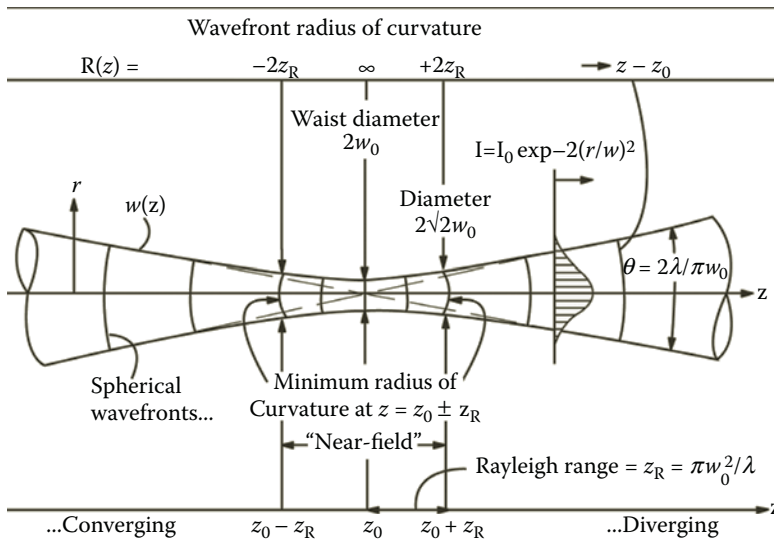
FIGURE 1.4
Propagation properties of the pure gaussian, fundamental-mode beam. The wavefront curvatures are exaggerated to show their variation with propagation distance.

where σ is the standard deviation of the gaussian distribution. Comparing Equation 1.4 and Equation 1.5 shows that the $1/e^2$ diameter is related to the standard deviation σ of the irradiance profile, as defined in Equation 1.5, as

$$2w = 4s. \tag{1.6}$$

For a beam of total power P, the value of the peak irradiance I_0 is found[5] by integrating Equation 1.4 over the transverse plane (yielding I_0 times an area of $\pi w^2/2$) and equating this to P. The result

$$I_0 = \frac{2P}{pw^2} \tag{1.7}$$

is easily remembered by noting that "the average irradiance is half the peak irradiance." This is a handy, often-used simplification allowing the actual beam profile to be replaced by a round flat-topped profile of diameter $2w$ for back-of-the-envelope conceptualizations (see Figure 1.3b).

If the gaussian beam is centered on a circular aperture of diameter $2a$ the transmitted fraction $T(a/w)$ of the total beam power is given by a similar integration[5] over the cross-sectional area as (see Figure 1.3c):

$$T\left(\frac{a}{w}\right) = 1 - \exp\left[-2\left(\frac{a}{w}\right)^2\right]. \tag{1.8}$$

This gives a transmission fraction of 86.5% for an aperture of diameter $2w$, and 98.9% for one of diameter $3w$. In practice, a minimum diameter for an optic or other aperture to pass the beam and leave it unaffected is $4.6w$ to $5w$ to reduce the sharp edge diffraction ripples overlaid on the beam profile to an amplitude of <1%.[5] It is interesting to note that for a low power, visible, fundamental-mode beam, the spot appears to be a diameter of about $4w$ to the human eye viewing the spot on a card.

The transmission of a fundamental-mode beam past a vertical knife-edge is also readily computed. The knife-edge transmission function is $T(x/w) = 0$ for $x' \leq x$, $T = 1$ for $x' > x$, where x is the horizontal distance of the knife-edge from the beam axis and x' is the horizontal integration variable. In Equation 1.4, the substitution $r^2 = x^2 + y^2$ is made, the integration over y yields multiplication by a constant, and the final integration over x' is expressed in terms of the error function as:

$$T\left(\frac{x}{w}\right) = \left(\frac{1}{2}\right)\left[1 \pm \text{erf}\left(\frac{\sqrt{2}\, x}{w}\right)\right], \quad + \text{ if } x < 0, \quad - \text{ if } x > 0. \tag{1.9}$$

The error function of probability theory in Equation 1.9 is defined (see p. 745 of Reference 23) as

$$\text{erf}(t) = \left(\frac{2}{\sqrt{p}}\int_0^t \exp(-u^2)\text{du}\right) \tag{1.10}$$

and is tabulated in many mathematical tables. The $1/e^2$ diameter of a fundamental-mode beam is measured with a translating knife-edge by noting the difference in translation distances of the edge ($x_1 - x_2$) that yield transmissions of 84.1% and 15.9%. By Equation 1.9 this separation equals w, and the beam diameter is twice this difference.[*]

1.4.4 Propagation Properties of the Fundamental-Mode Beam

The general properties expected for the propagation of a gaussian beam can be outlined from simple physical principles. As predicted by solving the wave equation with diffraction, a bundle of focused paraxial rays converges to a *finite* minimum diameter $2w_0$, called the waist diameter. The full angular spread θ of the converging and, on the other side, diverging beam is proportional to the beam's wavelength λ divided by the minimum diameter,[10] $\theta \propto \lambda/2w_0$. A scale length z_R for spread of the beam, is the propagation distance for the beam diameter to grow an amount comparable to the waist diameter, or $z_R\theta \sim w_0$, giving $z_R \propto w_0^2/\lambda$. Because the rays of the bundle propagate perpendicularly to the wavefronts (surfaces of constant phase), at the minimum's location the rays are parallel by symmetry and the wavefront there is planar. At large distances $z - z_0$ from the waist diameter location at z_0—the propagation axis is z—the wavefronts become Huygen's wavelets diverging from z_0 with wavefront radii of curvature $R(z)$, and eventually become plane waves. Since the wavefronts are plane at the minimum diameter at the waist and at large distances on either side, but converge and diverge through the waist, there must be points of maximum wavefront curvature (minimum radius of curvature) to either side of z_0.

The actual beam propagation equations describing the change in beam radius $w(z)$ and radius of curvature $R(z)$ with z, are derived[1,2,5] as solutions to the wave equation in the complex plane and show all of these features. They are (see Figure 1.4):

$$w(z) = w_0\sqrt{1 + \frac{(z - z_0)^2}{z_R^2}} \tag{1.11}$$

$$R(z) = (z - z_0)\left[1 + \frac{z_R^2}{(z - z_0)^2}\right] \tag{1.12}$$

$$z_R = \frac{pw_0^2}{l} \tag{1.13}$$

$$q = \frac{2l}{pw_0} = \frac{2w_0}{z_R} \tag{1.14}$$

---

[*] The knife-edge transmission function is illustrated later in Figure 1.8c and 1.8f of Section 1.6.

and

$$y(z) = -\tan^{-1}\left(\frac{z}{z_R}\right).$$ (1.15)

In these equations, the minimum beam diameter $2w_0$ (the waist diameter) is located at z_0 along the propagation axis z. A plot of $w(z)$ versus z, beam radius versus propagation distance [Equation 1.11] is termed the axial profile or propagation plot and is a hyperbola. The scale length for beam expansion, z_R, is termed the Rayleigh range [Equation 1.13] and has the expected dependence on λ and w_0. The radius of curvature $R(z)$ of the beam wavefront, as given by Equation 1.12, has the expected behavior. At large distances from the waist—the region termed the "far-field"—and where $|z - z_0| \gg z_R$ the radius of curvature first becomes $R \to (z - z_0)$ and then becomes plane when $|R| \to \infty$ as $|z - z_0| \to \infty$, and also is plane at $(z - z_0) = 0$. By differentiating Equation 1.12 and equating the result to zero the points of minimum absolute value of the radius of curvature are found to occur at $z - z_0 = \pm z_R$ and have the values $R_{min} = \pm 2z_R$. The full divergence angle θ develops in the far-field, the beam envelope is asymptotic to two straight lines crossing the axis at the waist location (Figure 1.4). Finally, $\psi(z)$ is the phase shift[5,24] of the laser beam relative to that of an ideal plane wave. It is a consequence of the beam going through a focus (the waist), the gaussian beam version of the Gouy phase shift.[24]

By Equation 1.11, the diameter $2w(z)$ of the beam increases by the factor $\sqrt{2}$ (and for a round beam the cross-sectional area doubles) for a propagation distance $\pm z_R$ away from the waist (Figure 1.4). This condition is often used to define the Rayleigh range z_R,[5,25] but another significant condition is that at these two propagation distances the wavefront radius of curvature goes through its extreme values ($|R| = R_{min}$). The Rayleigh range can be defined as half the distance between these curvature extremes. The region within a Rayleigh range of the waist is defined as the "near-field" region. Within this region wavefronts flatten as the waist is approached and outside they flatten as they recede from the waist. A positive lens placed in a diverging beam and moved back towards the source waist will encounter ever-steeper wavefront curvatures so long as the lens remains out of the near-field. On the lens output side, the transformed waist moves away from the lens, moving qualitatively as a geometrical optics image would. When the lens enters the near-field region still approaching the source waist, ever-flatter wavefronts are encountered and then the transformed waist *also* approaches the lens. The laser system designer who misunderstands this unusual property of beams will have unpleasant surprises. Many laser systems have undergone emergency redesign when prototype testing revealed this counter-intuitive focusing behavior! In many ways, Rayleigh range is the single most important quantity in characterizing a beam (notice that this is a factor in all of Equations 1.11 through 1.15). It will be shown in the next section that measurement of a beam's Rayleigh range is the basis for measuring the beam quality M^2 of a mixed-mode beam.

As the lowest order solution to the wave equation, the fundamental-mode with a gaussian irradiance profile of a given waist diameter $2w_0$ is the beam of lowest divergence, at the limit set by diffraction,[10] of any paraxial bundle with that minimum diameter. Confining a bundle to a smaller diameter proportionally increases—by diffraction—the divergence angle of the bundle, and the product $2w_0\theta$ is an invariant for any mode. The smallest possible value, $4\lambda/\pi$, is achieved only by the fundamental mode. This is just the Uncertainty Principle for photons—laterally confining a photon in the bundle increases the spread of its transverse momentum and correspondingly the divergence angle of the bundle. This limit cannot be achieved by real-world lasers but sometimes it is closely approached. Helium–neon lasers, especially the low-cost versions with internal mirrors (no Brewster windows),

are wonderful sources of beams within 1% or 2% of this limit. Aside from the wavelength, which must be known to specify any beam, the ideal, round, (stigmatic) fundamental-mode beam is specified by only two constants: the waist diameter $2w_0$ and its location z_0 (or equivalents such as z_R and z_0). This will no longer be true when mixed modes are considered.

As noted at the beginning of this section the propagation constants for the (x, z) and (y, z) planes are independent and can be different. In each plane, the rays obey equations exactly of the same form[6] as Equations 1.11 through 1.15 with subscripts added indicating the x or y plane. For beams with pure (but different) gaussian profiles in each plane, two more constants are introduced for a total of four required to specify the beam. If $z_{0x} \neq z_{0y}$ (different waist locations in the two principal propagation planes) the beam exhibits simple astigmatism; if $2w_{0x} \neq 2w_{0y}$ (different waist diameters) the beam has asymmetric waists.[*]

1.4.5 Propagation Properties of the Mixed-Mode Beam: The Embedded Gaussian and the *M²* Model

In Section 1.4.2 a mixed mode was defined as the power-weighted superposition of several higher-order modes originating in the same resonator, each with the same underlying gaussian waist radius w_0 determining the radial scale length $w(z)$ in their mode functions [Equations 1.1 and 1.2]. This underlying fundamental mode, with w_0 fixed[5] by the radii of curvature and spacing of the resonator mirrors, is called the embedded gaussian for that resonator regardless of whether or not the mixed mode actually contains some fundamental-mode power. To treat the mixed-mode case, use is made[7] of the fact that its diameter is everywhere (for all z) proportional to the embedded gaussian diameter. From Equation 1.3 the substitution $w(z) = W(z)/M$ in Equations 1.11 through 1.15 yields the mixed-mode propagation equations:

$$W(z) = W_0 \sqrt{1 + \frac{(z - z_0)^2}{z_R^2}} \tag{1.16a}$$

$$R(z) = (z - z_0)\left[1 + \frac{z_R^2}{(z - z_0)^2}\right] \tag{1.17a}$$

$$Z_R = \frac{p W_0^2}{M^2 \mathit{l}} = z_R \tag{1.18}$$

and

$$\Theta = \frac{2M^2 \mathit{l}}{p W_0} = \frac{2W_0}{z_R} = M\varphi. \tag{1.19}$$

The mixed mode, a sum of transverse modes with different optical frequencies, no longer has a simple expression for the Gouy phase shift analogous to Equation 1.15. The convention followed here is that upper case quantities refer to the mixed mode and lower case quantities refer to the embedded gaussian.

Also useful are the inverse forms of Equation 1.16a and Equation 1.17a expressing the waist radius W_0 and waist location z_0 in terms of the beam radius $W(z)$ and wavefront curvature $R(z)$ at propagation distance z:

$$W_0 = \frac{W(z)}{\sqrt{1 + [p W(z)^2 / M^2 \mathit{l} R(z)]^2}} \tag{1.16b}$$

---

[*] These beam asymmetries are illustrated later in Figure 1.15 of Section 1.8.

and

$$z_0 = \frac{R(z)}{1 + [M^2 \, \lambda R(z) / p W(z)^2]^2} \tag{1.17b}$$

These forms are obtained from Reference 1, with the substitution $w = W/M$ in their Equations 24 and 25.

Many of the properties of the fundamental-mode beam carry over to the mixed-mode one (Figure 1.5). Since $W_0 = Mw_0$, substitution of this in the middle part of Equation 1.19 yields the last part, the mixed-mode divergence is M times that of the embedded gaussian. Similarly, the beam propagation profile $W(z)$ also has the form of a hyperbola (one M times larger) with asymptotes crossing at the waist location. The Rayleigh ranges are the same for both mixed and embedded gaussian modes as substituting $W_0 = Mw_0$ in the middle of Equation 1.18 shows, so the radii of curvature and the limits of the near-field region are the same for both. The mixed-mode beam diameter still expands by a factor of $\sqrt{2}$ in a propagation distance of z_R away from the waist location z_0, the starting diameter W_0 is just M times larger.

In considering propagation in the independent (x, z) and (y, z) planes, there are now two new constants needed to specify the beam, M_x^2 and M_y^2, for a total of six required constants. In making up the mixed mode, the Hermite–Gaussian functions summed in the two planes need not be the same or have the same distribution of weights, making $M_x^2 \neq M_y^2$ a possibility. In this case the beam is said to have divergence asymmetry since $\Theta \propto M^2$ by the first part of Equation 1.19.

It might be asked, why are these Equations 1.16 through 1.19 termed the "M^2 model" (and not the "M model")? There are two reasons. The first is that the embedded gaussian is buried in the mixed-mode profile, and cannot be measured independently, making it difficult to directly determine M. The mixed-mode diameter still grows by $\sqrt{2}$ in a
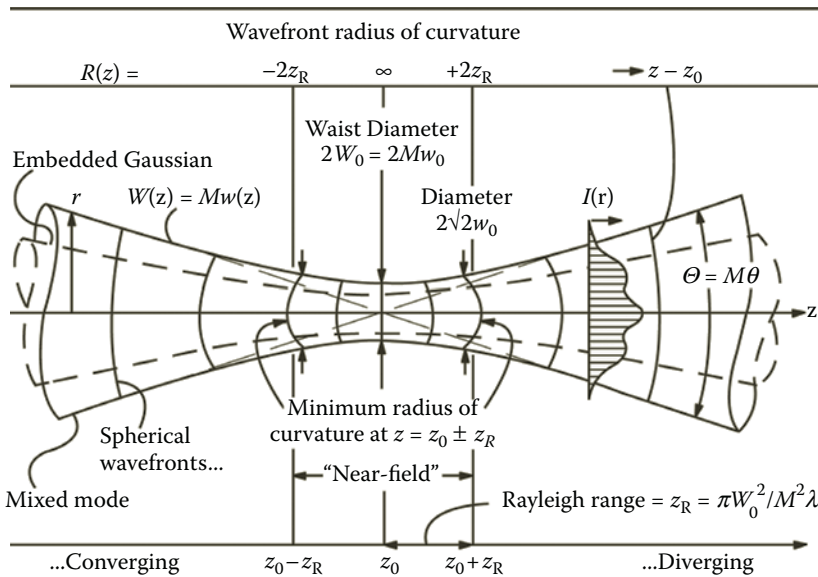
FIGURE 1.5
Propagation properties of the mixed-mode beam drawn for $M^2 = 2.63$. The embedded gaussian is the fundamental-mode beam originating in the same resonator. The wavefront curvatures are exaggerated to show their variation with propagation distance.

propagation distance z_R from the waist location, so z_R can be found from several diameter measurements fitted to a hyperbolic form. The waist diameter $2W_0$ can also be measured, thus giving directly, by Equation 1.18,

$$M^2 = \frac{p W_0^2}{1 z_R}.$$ (1.20)

This is how M^2 is in fact measured, the practical aspects of which will be discussed in Section 1.7. (As an aside, notice that Equation 1.20 shows that M^2 scales as the square of the beam diameter; this is used later in the discussion of conversions between different diameter definitions in Section 1.6.4.)

The second reason is the more important one: M^2 is an invariant of the beam, and is conserved[26] as the beam propagates through ordinary nonaberrating optical elements. Like the fundamental-mode beam whose waist diameter-divergence product was conserved, the same product for the mixed-mode beam is

$$(2W_0)\Theta = \frac{(2W_0)2M^2 1}{p W_0} = M^2 \frac{4 1}{p}.$$ (1.21)

This is larger by the factor M^2 than the invariant product for a fundamental mode.

Equation 1.21 can be rearranged to read

$$M^2 = \frac{\Theta}{(2 1 / p W_0)} = \frac{\Theta}{q_n}.$$ (1.22)

Here $\theta_n = 2\lambda/\pi W_0$ is recognized as the divergence of a fundamental-mode beam with a waist diameter $2W_0$, the same as the mixed-mode beam. This is called the normalizing gaussian; it has an M times larger scale constant $W_0 = M w_0$ in its exponential term than the embedded gaussian and it would *not* be generated in the resonator of the mixed-mode beam. It does represent the diffraction-limited minimum divergence for a ray bundle constricted to the diameter $2W_0$. Thus by Equation 1.22 the invariant factor M^2 can be seen to be the "times-diffraction-limit" number referred to in the literature.[5] This also identifies M^2 as the inverse beam quality number, the highest quality beam being an idealized diffraction-limited one with $M^2 = 1$, while all real beams are at least slightly imperfect and have $M^2 > 1$.

The value of the M^2 model is twofold. Once the six constants of the beam are accurately determined (by fitting propagation plot data for each of the two independent propagation planes) they can be applied by the system designer to accurately predict the behavior of the beam throughout the optical system before it is built. The spot diameters, aperture transmissions, focus locations, depths of field, and so forth can all be found for the vast majority of existing commercial lasers. The second value is that there are commercial instruments available that efficiently measure and document a beam's constants in the M^2 model. This permits quality control inspection of the lasers at final test, or whenever there is a system problem and the laser is the suspected cause. Defective optics can introduce aberrations in the beam wavefronts. If inside the laser, they increase M^2 by forcing larger amounts of high-divergence, high-order modes in the mixed-mode sum. If outside the resonator, they also adversely affect M^2. Measurement of the beam quality during system assembly, after each optic is added to detect a downstream increase in M^2, can aid in quality control of the overall optical system.

Beams excluded from the model as described are those whose orthogonal axes rotate or twist about the propagation axis (called beams with general astigmatism[15,16,27]) such as might come from lasers with nonplanar ring or out-of-plane folded resonators. The

symmetry of the beam is determined by the symmetry of the resonator. Fortunately, few commercial lasers produce beams having these characteristics. An overview of the full range of symmetry possibilities for laser beams is discussed in Section 1.8.3.

The fact that M^2 is not unique, that is, that a given value of M^2 can be arrived at by a variety of different higher-order modes or mode weights in the mixed mode is sometimes stated to be a deficiency of the M^2 model. This is also its strength. It is a simple predictive model that does not require measurement and analysis to determine the mode content in a beam. In the evolution of beam models, the original discussion[1,2] pointed out that as eigen-functions of the wave equation, the full (infinite) set of Hermite–Gaussian or Laguerre–Gaussian functions (Equation 1.1) describing the electric field of the beam modes form an orthonormal set. As such they could model an arbitrary paraxial light bundle with a weighted sum. This is true only if the phases of the E-fields are kept in the sum, and mea-suring the phase of an optical wave generally is a difficult matter. Summing the irradiances (the square of the E-fields) breaks the orthonormality condition and for years it was not obvious that a simple model relying only on irradiance measurements was possible. Then in the 1980s, methods based on Fourier transforms of irradiance and ray angular distribu-tions of light bundles were introduced,[4,6] which showed that as far as predictions of beam diameters in an optical system were concerned, irradiance profile measurements would (usually) suffice. The M^2 model was born, and commercial instruments[10] for its application soon followed. Later we realized that modes "turn on" in a characteristic sequence as dif-fraction losses are reduced in the generating resonator. This makes a given M^2 correspond to a unique mode mix in many common cases after all (see Section 1.6.4).

1.5 TRANSFORMATION BY A LENS OF FUNDAMENTAL AND MIXED-MODE BEAMS

Knowledge of how a beam is transformed by a lens is not only useful in general, but in particular, a lens is used to gain an accessible region around the waist for the measure-ments of diameters that are analyzed to produce M^2 (see Section 1.7). This transformation is discussed next.

In geometrical optics a point source at a distance s_1 from a thin lens produces a spheri-cal wave whose radius of curvature is R_1 at the lens (and whose curvature is $1/R_1$), where $R_1 = s_1$. In traversing the lens, this curvature is reduced by the power $1/f$ of the lens (f is the effective focal length of the lens) to produce an exiting spherical wave of curvature $1/R_2$ according to the thin lens formula:

$$\frac{1}{R_2} = \frac{1}{R_1} - \frac{1}{f}. \tag{1.23}$$

An image of the source point forms at the distance R_2 from the lens from convergence of this spherical wave. Note that the conventions used in Equation 1.23 are the same as in Equation 1.17, namely, the beam always travels from left to right, converging wavefronts with center of curvature to the right have negative radii, and diverging wavefronts with centers to the left have positive radii. [The usual convention in geometrical optics[28] is that converging wavefronts leaving the lens are assigned positive radii, which would put a minus sign on the term $1/R_2$ of Equation 1.23.]

The quantities used in the beam-lens transform are defined in Figure 1.6. Following Kogelnik[1] the beam parameters on the input side of the lens are designated with a subscript 1 (for "1-space") and on the output side with a subscript 2 (for "2-space"). The principal plane description[28] of a real (thick) lens is used, in which the thick lens is replaced by a thin one acting at the lens principal planes H1, H2. Rays between H1 and H2 are drawn parallel to the axis by convention, and waist locations z_{01} and z_{02} are measured from H1 and H2 respectively (with distances to the right as positive for z_{02} and distances to the left as positive for z_{01}).

A lens inserted in a beam makes the same change in wavefront curvature as it did in geometrical optics [Equation 1.23], but the wavefront R_2 converges to a waist of finite diameter $2W_{02}$ at a distance z_{02} given by Equation 1.17b. For each of the two independent propagation planes, there are three constants required to specify the transformed beam, and three constraints needed to determine them. The lens should be aberration-free (typically, used at $f/20$ or smaller aperture) and, if so, the beam quality is not changed in passing through it, giving the first condition $M_2^2 = M_1^2$. The second constraint is that the wavefront curvatures match, between the input curvature modified by the lens [Equation 1.23], and the transformed beam at the same location as specified by the transformed beam constants through Equation 1.17a. A beam actually has two points with the same magnitude and sign of the curvature, one inside the near-field region of that sign and one outside, which differ in beam diameters. The ambiguity as to which point is matched is removed by the third constraint, that the beam diameter is unchanged in traversing the (thin) lens.

These three constraints define three equations that next are solved for the transformed waist diameter and location. This is facilitated by Equations 1.16b and 1.17b for W_0 and z_0 as functions of $W(z)$ and $R(z)$. The solution[1,29–31] is written in terms of the transformation constant Γ (using the modern symbols from a commercial M^2 measuring instrument[9]) as follows:

$$\Gamma = \frac{f^2}{[(z_{01} - f)^2 + z_{R1}^2]} \tag{1.24}$$

$$M_1^{\,2} = M_2^2 = M^2 \tag{1.25}$$

$$W_{02} = \sqrt{\Gamma} W_{01} \tag{1.26}$$

$$z_{R2} = \Gamma z_{R1} \tag{1.27}$$

$$z_{02} = f + \Gamma(z_{01} - f) \tag{1.28}$$

A set of these equations apply to each of the two principal propagation planes (x, z) and (y, z).
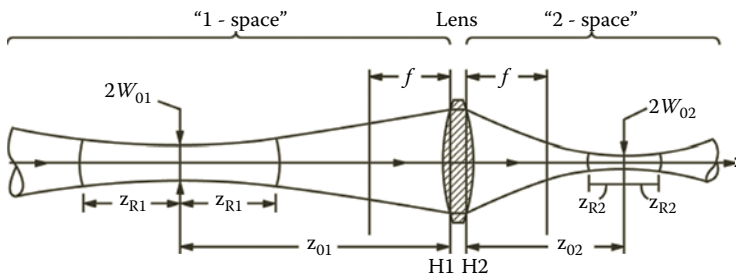
FIGURE 1.6
Definitions of quantities used in the beam-lens transform.
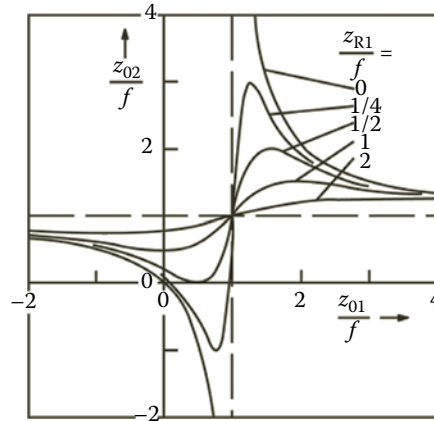
FIGURE 1.7
Parametric plots of the transformed waist location as a function of the input waist location for the beam-lens transform, with f as the lens focal length and the Rayleigh range z_{R1} of the input beam as parameters.

The transform equations [Equations 1.24 through 1.28] are not as simple as in geometrical optics because of the complexity of the way the beam wavefront curvatures change with propagation distance, Equation 1.17a. Like the image and object distances in geometrical optics, the transformed beam waist location depends on the input waist location—but also depends, as does the wavefront curvature, on the Rayleigh range of the input beam. The most peculiar behavior as the waist-to-lens distance varies is when the input focal plane of the lens moves within the near-field of the incident beam, $|z_{01} - f| < z_{R1}$. Then the slope of the z_{02} versus z_{01} curve turns from negative to positive (in geometrical optics the slope of the object to image distance curve is always negative). This sign change can be demonstrated by substituting Equation 1.24 into Equation 1.28 and differentiating the result with respect to z_{01}. As the lens continues to move closer to the input waist, the transformed waist location also moves closer to the lens, exactly opposite to what happens in geometrical optics. In the beam-lens transform, the input and transformed waists are *not* images of each other (in the geometrical optics sense). Despite the intransigence of beam waists, the object–image relationship of beam diameters at conjugate planes on each side of the lens does apply just as in geometrical optics. A good modern discussion of the beam-lens transform is presented in O'Shea's textbook[32] (where his parameter $\alpha^2 = \Gamma$ here).

A pictorial description of the beam-lens transform is given by a figure in Reference 30, redrawn here as Figure 1.7. Variables normalized to the lens focal length f are used to show how the transformed waist location z_{02}/f varies with the input waist location z_{01}/f. The input Rayleigh range z_{R1}/f (also normalized) is used as a parameter and several curves are plotted for different values. The anomalous slope regions of the plot are evident. The geometrical optics thin lens result, Equation 1.23, is recovered when the input Rayleigh range becomes negligible, $z_{R1}/f = 0$ (the condition for a point source), and the slopes of both wings of the curve are then always negative.

1.5.1 Application of the Beam-Lens Transform to the Measurement of Divergence

An initial application of the beam-lens transform equations is to show that the divergence of the input beam Θ_1 in 1-space of Figure 1.6 can be determined by measuring the

beam diameter $2W_f$ at precisely one focal length behind the lens exit plane H2 in 2-space from the equation:

$$\Theta_1 = \frac{2W_f}{f}.$$ (1.29)

This result is independent of where the lens is placed in the input beam. This follows by finding in 2-space the diameter $2W_f$ at $z_2 = f$ [from Equation 1.16a] and substituting Equations 1.19, 1.24, and 1.28:

$$2W_f = 2W_{02}\left[1 + \frac{(f - z_{02})^2}{z_{R2}^2}\right]^{1/2} = 2W_{02}\left(\frac{f}{z_{R2}}\right)\left(\frac{1}{\Gamma^{1/2}}\right)$$

$$= 2W_{01}\left(\frac{f}{z_{R2}}\right)\left(\frac{1}{\Gamma}\right) = 2W_{01}\left(\frac{f}{z_{R1}}\right) = \Theta_1 f$$

which is Equation 1.29. In Figure 3b of Reference 25 there is an illustration showing how the transform equations operate to keep the output beam diameter one focal length from the lens fixed at the value $\Theta_1 f$ despite variations in the input waist location, z_{01}. The measurement method implied by Equation 1.29 is the simplest way to get a good value for the beam divergence Θ_1. Care should be taken to pick a long enough focal length lens that the beam diameter is large enough for the precision of the diameter-measurement method in use.

1.5.2 Applications of the Beam-Lens Transform: The Limit of Tight Focusing

When the aperture of a short focal length lens is filled on the input side, the smallest possible diameter output waist is reached and this is called the limit of tight focusing. This limit is characterized by (1) the beam diameter at the lens being given by $2W_{lens} = \Theta_2 f$; (2) the output waist being near the focal plane $z_{02} = f$; and (3) there being a short depth of field at the focus, $z_{R2}/f \ll 1$. Applying Equation 1.29 in the reverse direction gives the 2-space divergence as the ratio of the beam diameter $2W_{1f}$ at f to the left of the lens, to the focal length, $\Theta_2 f = 2W_{1f}$. By condition (1) this means $2W_{1f} = 2W_{lens}$ or that there is little change in the input beam diameter over a propagation distance f. That makes the first condition characterizing the tight focusing case equivalent to $z_{R1}/f \gg 1$. Then from Equation 1.19,

$$2W_{lens} = \frac{2\,\jmath M^2 f}{p W_{02}}$$

or

$$2W_{02} = 2\,\jmath M^2\left(\frac{f}{p W_{lens}}\right) = 2\,\jmath M^2 (f/\#)$$ (1.30)

for the tight focusing limit. Here Siegman's definition[5] is used that a lens of diameter D_{lens} is filled for a fundamental-mode beam of diameter πW_{lens} (this degree of aperture filling produces <1% clipping of the beam). Thus $f/\pi W_{lens} = f/D_{lens} = (f/\#)$. The depth of field of the focus is $z_{R2} = \pi W_{02}^2/M^2\lambda = \pi M^2\lambda(f/\#)^2$. This generalizes a familiar result[5] for a fundamental-mode beam to the $M^2 \neq 1$ case.

Marshall's point[3] (from 1971) is made by Equation 1.30, that a higher-order mode beam focuses to a larger spot by a factor of M^2, with less depth of field, and therefore cuts and welds less well than a fundamental-mode beam.

1.5.3 The Inverse Transform Constant

The transform equations work equally well going from 2-space to 1-space, with one transformation constant the inverse of the other,

$$\Gamma_{21} = \frac{1}{\Gamma_{12}}. \qquad (1.31)$$

This obviously is true by symmetry but the algebraic proof is left to the reader.

1.6 BEAM DIAMETER DEFINITIONS FOR FUNDAMENTAL AND MIXED-MODE BEAMS

It has been said that the problem of measuring the cross-sectional diameter of a laser beam is like trying to measure the diameter of a cotton ball with a pair of calipers. The difficulty is not in the precision of the measuring instrument, but in deciding what is an acceptable definition of the edges.

Unlike the fundamental-mode beam where the $1/e^2$ diameter definition is universally understood and applied, for mixed modes a number of different diameter definitions[7] have been employed. The different definitions have in common that they all reduce to the $1/e^2$ diameter when applied to an $M^2 = 1$ fundamental-mode beam, but when applied to a mixed mode with higher-order-mode content they in general give different numerical values. As M^2 always depends on a product of two measured diameters, its numerical value changes also as the square of that for diameters. It is all the same beam, but different methods provide results in different currencies; one has to specify what currency is in use and know the exchange rate.

Since the adoption[11] by the ISO committee on beam widths of the second-moment diameter as the standard definition for beam diameters, there has been increasing effort among laser users to put this into practice. This definition, discussed in Section 1.6.3.5, has the best analytical and theoretical support but is difficult experimentally to measure reproducibly because of sensitivity to small amounts of noise in the data. The older methods therefore persist and the best strategy[25] at present is to use the more forgiving methods for the multiple diameter measurements needed to determine M^2. Then at one propagation distance, do a careful diameter measurement by the second-moment definition to provide a conversion factor. This conversion factor can then be applied to obtain standardized diameters at any distance z in the beam. This strategy will likely evolve in the future if and when instrument makers respond to the ISO Committee's choice and devise algorithms and direct methods for ready and accurate computations of second-moment diameters.

1.6.1 Determining Beam Diameters from Irradiance Profiles

Beam diameters are determined from irradiance profiles, the record of the power transmitted through a mask as a function of the mask's translation coordinate transverse to

the beam. A sufficiently large linear power detector is inserted in the beam, with a uniformly sensitive area to capture the total power of the beam. Detection sensitivity should be adequate to measure ~1% of the total power, and response speed should allow faithful reproduction of the time-varying transmitted power. The mask is mounted on a translation stage, placed in front of the detector, and moved or scanned perpendicularly to the beam axis to record a profile. An instrument that performs these functions is called a beam profiler. In a useful version based on a charge-coupled-device (CCD) camera, the masking is done on electronic pixel data under software control.

The beam propagation direction defines the *z*-axis. The scan direction is usually along one of the principal diameters of the beam spot and commercial profilers are mounted to provide rotation about the beam axis to facilitate alignment of the scan in these directions. The principal diameters for an elliptical spot are the major and minor axes of the ellipse (or the rectangular axes for a Hermite–Gaussian mode). The principal propagation planes (*x, z*) and (*y, z*) are defined as those containing the principal spot diameters. The beam orientation is arbitrary and in general may require rotation of coordinates to tie it to the laboratory reference frame. It is assumed this rotation is known, and without loss of generality to give simple descriptive terminology in this discussion, here the *z*-axis is taken to be horizontal, the principal propagation planes as the horizontal and vertical planes in the laboratory, with the scan along the *x*-axis. If the mask requires centering in the beam (e.g., a pinhole) to find the principal diameter, it is mounted on a *y*-axis stage as well and *x*-scans at different *y*-heights taken to determine the widest one at the beam center. Alternatively, a mirror directs the beam onto the profiler and the spot is put at different heights to find the beam center by tipping the mirror about a horizontal rotation axis. If the beam is repetitively pulsed and detected with an energy meter, the stage is moved in increments between pulses. If a CCD camera is the detector, a scan line is the readout of sequential pixels and no external mask is required in front of the camera. A CCD camera generally requires a variable attenuator[33] inserted before the camera to set the peak irradiance level just below the saturation level of the camera for optimum resolution of the irradiance value on the ordinate axis of the profile.

The results of this process are irradiance profiles such as shown in Figure 1.8 for two pure modes, the fundamental mode in the first row and the donut mode in the second, where three scans are calculated for each, one for a pinhole (first column), a slit (second column), and a knife-edge (third column) as masks. The traditional definitions used to extract diameters from these profiles are the same for the pinhole and slit. This is to normalize the scan to the highest peak as 100%, then to come down on the scan to an ordinate level at $1/e^2$ (13.5%) and measure the diameter—or clip width—as the scan width between these crossing points (called clip levels or clip points and shown as dots in Figure 1.8). The symbols D_{pin} and D_{slit} are used for these two diameters. For the knife-edge diameter (symbol D_{ke}) the definition is to take the scan width between the 15.9% and 84.1% clip points and double it, as this rule produces the $1/e^2$ diameter when applied to the fundamental mode.

As shown in Figure 1.8 the diameter results for the donut mode (TEM_{01}^*) are all larger than the $2w$ diameter of the fundamental mode, as expected. However, the answers for the three different methods for the donut mode—and in general, for all higher-order modes—are all different! The ratio of the donut mode to fundamental-mode diameter is 1.51, 1.42, and 1.53 by the pinhole, slit, and knife-edge methods, respectively. The reason, obviously, is that traces of different shapes are produced by the different methods. The pinhole cuts the donut right across the hole and records a null at the center; the slit extends vertically across the whole spot and records a transmission dip in crossing the hole but never reaches zero due to the contribution of the light above and below the hole. Even higher transmission
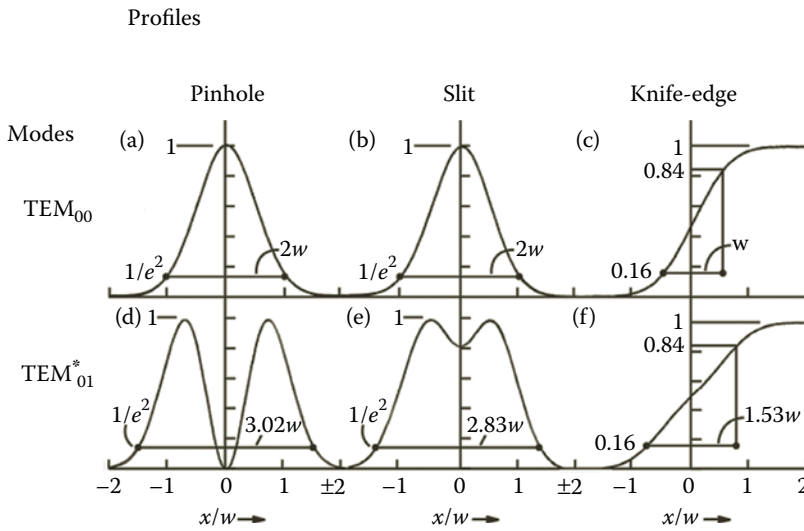
Profiles

FIGURE 1.8
Theoretical beam profiles (irradiance vs. translation distance) from a scanning pinhole (a) and (d), slit (b) and (e), and knife-edge (c) and (f) cutting the fundamental and donut modes, illustrating that different methods give different diameters for higher-order mode beams. The knife-edge diameter is defined as *twice* the translation distance between the 15.9% and 84.1% cut points.

results with the knife-edge and here the donut profile differs from the fundamental one only in being less steeply sloped (the spot is wider) and having slight inflections of the slope around the hole at the 50% clip point, the beam center.

There are two other two common definitions. The first is the diameter of a circular aperture giving 86.5% transmission when centered on the beam. It is variously called the variable-aperture diameter, the encircled power diameter, or the "power-in-the-bucket" method, and designated by the symbol D_{86}. The last is the second-moment diameter, defined as four times the standard deviation of the radial irradiance distribution recorded by a pinhole scan, and designated by the symbol $D_{4\sigma}$. For the ratio of donut mode to fundamental-mode diameters, these definitions give 1.32 and 1.41 respectively, also different from the three other values above.

After the discussion of some common considerations (Section 1.6.2), these five diameter definitions are evaluated in Section 1.6.3 leading to the summary given in Table 1.1.

1.6.2 General Considerations in Obtaining Useable Beam Profiles

Five questions are important in evaluating what beam diameter method is best for a given application:

1. *How important is it to resolve the full range of irradiance variations?* Only a pinhole scan (or its near equivalent, a CCD camera snapshot read out pixel by pixel) shows the full range, but this is not of significance in some applications, for example, where the total dose of light delivered is integrated in an absorber.

2. *How important is it to use a method that is insensitive to the alignment of the beam into the profiler?* If the test technician cannot be relied on to carefully center the beam on the profiler, the slit or knife-edge methods still give reliable results,

TABLE 1.1

Properties of Mixed-Mode Diameter Definitions

Diameter Symbol	Scan Aperture and Name	Diameter Definition	Conversion Constant $c_{i\sigma}$ to $D_{4\sigma}$	Alignment Sensitive?	Resolution of I(r) Peaks	Convolution Error?	Signal-to-Noise Ratio	Comments
D_{pin}	Pinhole (dia. H)	Separation of clip points $1/e^2$ down from the highest peak	0.805	Yes	High	Yes, if $H/2w > 1/6$	Low	Shows best details of irradiance peaks
D_{slit}	Slit (width S)	Separation of clip points $1/e^2$ down from the highest peak	0.950	No	Medium	Yes if $S/2w > 1/8$	Medium	Directly measured diameter is close to $D_{4\sigma}$
D_{ke}	Knife-edge	Twice the separation of 15.9%, 84.1% clip points	0.813	No	Low	None	High	Most robust diameter experimentally (vs. noise and spot structure)
D_{86}	Variable aperture ("power in the bucket")	Diameter of centered circular aperture passing 86.5% of total power	1.136	Yes	Low	None	High	Works well only on round beams. Computed readily on CCD cameras. Used on kW lasers
$D_{4\sigma}$ or $D_{2\sqrt{2}s}$	Second-moment (linear or radial)	Four times the standard deviation of irradiance distribution from a pinhole scan	1	Yes	High	Yes, as for pinhole scan	Low	*ISO standard diameter.* Susceptible to error from noise on wings of the profile. Supported best by theory. Computed readily on CCD cameras
N/A	CCD camera	Various custom algorithms	N/A	No	Medium	Yes	Low	Computes all of above diameter definitions with appropriate software

but not the other methods. With a CCD camera there is a trade-off between alignment sensitivity and accuracy. For best accuracy, a magnifying lens—of known magnification—can be placed in front of the camera to fill the maximum number of pixels, but then the camera becomes somewhat alignment sensitive.

3. *With what accuracy and repeatability is the diameter determined?* The amount of light transmitted by the mask determines the signal-to-noise ratio of the profile and ultimately answers the question. The methods based on a pinhole scan (D_{pin}, $D_{4\sigma}$, and CCD cameras) suffer from low light levels in this regard. On the other hand, a laser beam is generated in a resonator subject to microphonic perturbations, making the beam jitter in position and the profile distort typically by about 1% of the beam diameter, so that a greater instrument measurement accuracy is usually not significant.

4. *Is the convolution error associated with the method significant?* The convolution error is the contribution to the measured diameter due to the finite dimensions of the scan aperture, either the diameter H of a pinhole or width S of a slit. A 10-micron focused spot cannot be accurately measured with a pinhole of 50-micron diameter. The distortion of a pinhole profile of a fundamental mode is shown in Figure 1.9a as a function of the ratio of hole diameter to the mode width $H/2w$. The peak amplitude drops and a slight broadening occurs as $H/2w$ increases. The central 100% peak amplitude point is "washed out" or averaged to a lower value in the profile by the sampling of lower amplitude regions nearby as the finite diameter pinhole scans across the center as Figure 1.9b indicates. The reduction in peak amplitude of the convoluted profile is like lowering the clip level below 13.5% on the original profile: the measured diameter becomes larger. Very similar profile distortions occur with a slit scan as a function of $S/2w$; here S is the silt width. The ratio of the measured width including this convolution error to the correct width is plotted in Figure 1.9c for the pinhole (H) and slit (S). This gives the rule of thumb for pinhole scans: to keep the error in the measured diameter to 1% or less, keep the pinhole diameter H to one-sixth or less of $2w$, that is, $H < w/3$. The corresponding rule[34] for slits is the measured diameter is in error by <1% if the width S is 1/8 or less of $2w$. For modes like TEM_{10} of Figure 1.2d with a feature (the central peak) narrower than that of the fundamental mode, the aperture widths H or S should be no bigger than these same fractions of the narrow feature's width. (Note, McCally[34] uses the biologist's definition of $1/e$ clip points for the fundamental-mode diameter, a factor $1/\sqrt{2}$ smaller than our $1/e^2$ diameter; his results require conversion.)

Distortion of the profile can be a more subtle effect and can give misleading results. When measuring a predominantly TEM^*_{01} focused beam through the waist region, for example, a pinhole profiler will at first show the expected trace, with a dip in the middle like Figure 1.8d or e. This will change to one with a central peak as in Figure 1.8a at the propagation distance along the beam where the pinhole is no longer small compared to the beam diameter. The donut hole can fall through the pinhole!

Convolution errors are a concern normally only when working with focused beams, as when measuring divergence by the method of Section 1.5.1. Generally, however, it is desirable to go to the far-field, reached by working in 2-space at the focal plane behind an inserted lens, to obtain a true (undistorted) profile. The beam coming out of the laser often has "diffractive
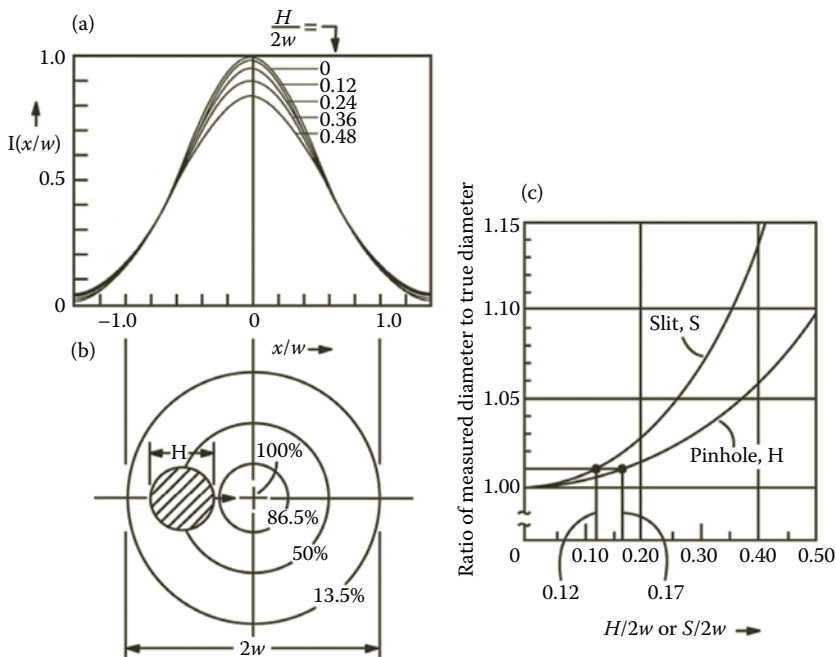
FIGURE 1.9

Convolution of the theoretical fundamental-mode profile in a scan with a pinhole or slit of finite dimensions (*H*, diameter of the pinhole; *S*, width of the slit; 2*w*, the $1/e^2$ diameter of the mode). (a) Distortion of the shape and width of the pinhole profile as *H*/2*w* increases. (b) Plan view of the pinhole scan showing "washout" of the 100% amplitude point. For the pinhole shown, *H*/2*w* = 0.24, corresponding to the third curve down from the top in (a). (c) Convolution error, or ratio of the measured diameter $2w_{meas}$ to the true diameter 2*w*, as a function of *H*/2*w* for the pinhole and *S*/2*w* for the slit.

overlay," low-amplitude high-divergence light diffracted from the mode-limiting internal aperture, overlaid on the main beam. The resulting interference can significantly distort the profile, even at <1% amplitude of the diffracted light. It is the E-fields that interfere; for an irradiance I = E^2 overlaid by a 0.01 E^2 distorting component, the E-fields add and subtract as $E \pm 0.1 E$ at the interference peaks and valleys. The resulting fringe contrast ratio, I_{peak}/I_{valley} = $[(1.1)/(0.9)]^2$ = 1.49 is a significant distortion to the profile even though the power in the diffractive overlay is insignificant. Moving the profiler some distance away from the output end of the laser spreads the diffractive overlay rapidly compared to the beam expansion, but often several meters additional distance is required. This leaves the use of a lens to reach the far-field as the answer, and convolution distortion then must be dealt with.

Aligning a small-diameter (e.g., 10 micron) pinhole to a small (e.g., 100 micron)-focused spot is another problem. The search time to achieve overlap and some transmitted signal for peaking alignment can be very long if done manually, so having a fast update rate— 10 scans a second is good—provided by commercial instruments can be a major aide. Some instruments[9] have electronic alignment systems to facilitate finding the overlap of small pinhole and small beam.

Knife-edges have no convolution error to the extent that they are straight (razor blades are straight[8] to <2 microns deviation over 1000 microns length). The circular aperture of the encircled power method is usually a precision drilled hole and has no convolution

error so long as it is accurately round and made in a material much thinner than the hole diameter (to avoid occultation error).

5. *Are the diameter measurements along the propagation path free of discontinuities and abrupt changes?* Consider making many diameter measurements along the propagation axis, and fitting the data to a hyperbola to find the beam's Rayleigh range and beam quality. Discontinuities in the data will make a poor fit and final result. Such discontinuities can arise[35] with the $1/e^2$ clip-level diameter definitions with mixed modes with low peaks on the edges, as in Figure 1.2g, only lower. As the mode mixture changes to bring the outer peaks near the clip level, the measured diameter can jump from the separation of the outer peaks of the profile to the width of the central peak as amplitude noise perturbs the profile. Similarly, for a mixed mode with rectangular symmetry, as azimuth is continuously changed from the major principal plane direction towards the minor one, the relative amplitude of the outermost peaks of the profile can drop.[35] The clip point then can jump discontinuously with perturbing noise when the height is near the clip level. Only D_{pin} and D_{slit} are subject to this difficulty.

This last question can be rephrased to ask, is the diameter definition readable by a machine? A human observer will notice an outer peak of height near the clip level causing the profiler readout to fluctuate, and correct the situation by adjusting the mode mixture, the azimuth, or the clip level. A machine will take the bad data in, and produce unreliable results. When a lot of diameter data needs to be gathered, as in measuring a propagation plot to determine M^2, automated machine data acquisition is desirable. In this regard, the knife-edge diameter is best, as it always produces an unambiguous monotonic trace for all higher-order and mixed modes.

1.6.2.1 How Commercial Scanning Aperture Profilers Work

Commercial profilers[8] typically use the $1/e^2$ diameter definition with pinhole and slit masks, and occasionally will report an incorrect diameter due to the "not entirely machine readable" defect of these definitions. These profilers use a rotating drum to carry a slit or pinhole mask smoothly and rapidly (typically at a 10 Hz repetition rate) in front of a large area detector inserted into the drum. On the first pass through the laser spot, the electronics remembers the 100% signal level, and on the second pass when the 13.5% clip level is crossed as the signal rises, a counter is started. This counts the angular increments of drum motion from an angular encoder, which when multiplied by the known drum radius, provides the mask translation in spatial increments of 0.2 microns. (In newer, high precision designs discussed in the next paragraph this increment has been reduced to 0.01 microns.) When the clip level is passed as the signal falls, the counter is stopped and the value of the beam diameter—total counts times spatial increment—is reported. Actually, what is reported on the digital readout is an average selected by the user of the last two to 20 measurements, to slow the report rate down to what can be read visually. If a pure donut mode is scanned with the pinhole version of this instrument [the profile of Figure 1.8d], the counter starts at the clip-level dot on the left ($x/w = -1.51$) but stops as the falling clip level is met at the left edge of the donut hole ($x/w = -0.16$). The scan continues and the counter turns on again as the clip level is passed with the rising signal at the right edge of the donut hole ($x/w = +0.16$), because the drum has not completed a revolution to reset the counter for a new measurement. Finally, the counter turns off again at the rightmost clip-level dot

(x/w = +1.51), and the diameter reported is the actual diameter minus the width of the hole at the clip-level height, an error of about –11%. This possible error usually goes unnoticed because the dips in mixed-mode profiles do not often go as low as 13.5%.

In recent years scanning aperture profilers have been mechanically upgraded to provide more precision (0.01 micron spatial resolution) and interfaced with PC controllers to provide more features in addition to beam diameter: full 12-bit digitized profiles and the $D_{4\sigma}$ diameters calculated from them (not just clip widths and analog traces), profile peak position, centroid position, spot ellipticity (with slit or knife-edge profilers carrying two orthogonal apertures), and even absolute power (when so calibrated). With micron-sized apertures and submicron sampling, beam diameters of 5 microns can be measured to 2% accuracy. As before, different detector types (silicon, germanium, or pyroelectric) cover wavelengths from UV to Far IR. Beams pulsed at repetition frequencies down to 1 kHz can be measured with profilers having user-controlled variable scan speed (drum speeds are slowed to intercept enough pulses to build up the profile). In addition they can measure beams without attenuation, as compared to camera-based systems that typically require six to nine orders of magnitude attenuation. Infrared beams at power levels of 3 kW focused to diameters of 175 µm have been directly measured with cooled profilers fitted with copper apertures.

Commercial profilers, because of their speed and accuracy, are a major improvement for frequent beam diameter measurements over the traditional practice of a manually driven translation stage carrying a razor blade (or slit) across the beam. Focused beams in particular need high instrument accuracy to resolve the small spot and provide the real time update rate to acquire a signal by overlapping the aperture with the beam. With a signal linearity range of 10^4 and a spatial resolution (if convolution error is neglected) of 0.01 microns over a 9-mm scan range (10^6 spatial resolution elements) one of these small, new profilers brings an impressive potential of 10^{10} information bits to the problem of measuring a beam diameter. Compare this to a modern CCD camera of 9-mm sensor width, 5-micron pixel spacing (2×10^3 spatial resolution elements), and 12-bit (4×10^3) linearity range, for a total of 10^7 information bits. It is understandable why in measuring beam quality M^2, profiler-based instruments surpass camera-based ones in speed and accuracy. The camera, of course, has its own advantages of giving a two-dimensional map of all the irradiance peaks in the laser spot and its ability to measure beams from low repetition rate pulsed lasers.

1.6.3 Comparing the Five Common Methods for Defining and Measuring Beam Diameters

The discussion that follows and Table 1.1 summarize the properties of the five diameter definitions.

1.6.3.1 D_{pin}, Separation of $1/e^2$ Clip Points of a Pinhole Profile

The pinhole scan reveals the structure of the irradiance variations across the beam spot with the greatest accuracy and detail, but does so working with a low light signal level and it is subject to convolution error with focused spots. To minimize convolution error, several pinholes of diameters H (10-micron and 50-micron pinholes are common) are used to keep $H < w/3$ where w here is the fundamental-mode radius or smallest feature size for a higher-order mode beam. The pinhole method requires accurate centering of the beam on the scan line of the pinhole and this makes it less adaptable to a machine measurement. This diameter definition also can give ambiguous results if the profile contains secondary

peaks of a height close to the clip level. The pinhole profile provides the basic data from which the second-moment diameter is calculated. Be sure the rule for the profile to be free of convolution error is met first!

1.6.3.2 D_{slit}, Separation of 1/e² Clip Points of a Slit Profile

The slit scan does not require centering of the beam spot and works at a medium light signal level, but does not reveal as much detail of the irradiance variations [compare Figure 1.8d and e]. This method is subject to convolution error with focused spots; the slit width S should satisfy $S/2w < 1/8$ with $2w$ as the smallest feature size of the profile. It too can give ambiguous results on profiles with secondary peaks near the clip level. This diameter definition produces a direct result (that is, without applying the conversion rules explained in Section 1.6.4.3) closest to the ISO standard second-moment diameter of the three other methods.

1.6.3.3 D_{ke}, Twice the Separation of the 15.9%
and 84.1% Clip Points of a Knife-Edge Scan

The knife-edge does not require centering of the beam spot and works at a high light signal level, but reveals almost no detail of the irradiance variations [compare Figure 1.8d and f], only the slight inflection points in the slope of the knife-edge profile show that there are any irradiance peaks at all. All modes give a simple slanted S-shaped profile. There generally is no convolution error with this method, and there are no diameter ambiguities when secondary peaks are present. Experimentally, it is the most robust diameter measurement and is least affected by beam-pointing jitter and power fluctuations, making this method fully machine readable. This diameter is the basic one measured in the most common commercial instrument[9] designed to automatically measure propagation plots and all six beam parameters.

1.6.3.4 D_{86}, Diameter of a Centered Circular Aperture
Passing 86.5% of the Total Beam Power

Unlike the other diameter measurements, the variable-aperture diameter passes light in both the *x*- and *y*-transverse planes simultaneously and cannot be used to separately measure the two principal diameters; it works best with round beams. It must also be centered in the beam for accurate results. While an iris or variable aperture can be used, more frequently sets of precision fixed apertures are used instead. A metal plate drill gauge, with some of the plate milled away on the back side of the gauge to reduce its thickness to less than the smallest aperture size to eliminate occultation error, is a convenient tool. The two diameters bracketing the 86.5% transmission point are first found, and the final result computed by interpolation. Alternatively, if there is a long propagation length available, an aperture with a transmission near 86.5% may be moved along the beam to locate the distance where that diameter produces precisely this transmission. This diameter definition is used mainly for two reasons. For high power lasers—for instance CO_2 lasers in the kilowatt range—little diagnostic analytical instrumentation is available that can absorb this power. A water-cooled copper aperture, however, can still be safely inserted in front of a power meter to give some quantification of the beam diameter. The second reason is that this diameter is readily computed from the output of a CCD camera and is available on camera instrumentation, with the computation locating the beam centroid, making physical centering of the camera unnecessary.

1.6.3.5 $D_{4\sigma}$, Four Times the Standard Deviation of the Pinhole Irradiance Profile

This diameter is computed from a pinhole irradiance profile, which for accuracy should be free of convolution error and diffractive overlay. For a beam with a rectangular cross-sectional symmetry described by a weighted sum of Hermite–Gaussian modes the calculation proceeds by finding the rectangular moments of the profile treated as a distribution function. The zeroth moment gives the total power P of the beam, the first moment the centroid, and the second moment leads to the variance σ^2 of the distribution:

$$\text{Zeroth moment or total power } P = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} I(x,y)\mathrm{d}x\mathrm{d}y \qquad (1.32)$$

$$\text{First moment or centroid } \langle x \rangle = \left(\frac{1}{P}\right)\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xI(x,y)\mathrm{d}x\mathrm{d}y \qquad (1.33)$$

$$\text{Second moment } \langle x^2 \rangle = \left(\frac{1}{P}\right)\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x^2 I(x,y)\mathrm{d}x\mathrm{d}y \qquad (1.34)$$

$$\text{Variance of the distribution } s_x^2 = \langle x^2 \rangle - \langle x \rangle^2 \qquad (1.35)$$

$$\text{Linear second-moment diameter } D_{4sx} = 4s_x \qquad (1.36)$$

This last equation comes from the requirement that the second-moment diameter reduce to the $1/e^2$ diameter when applied to a fundamental-mode beam, as explained in arriving at Equation 1.6. A precisely similar set of equations holds for the moments in the vertical plane (y, z) to define a vertical principal plane centroid and diameter [Equations 1.33 through 1.36 with x and y interchanged]:

$$\text{Linear second moment diameter } D_{4sy} = 4s_y. \qquad (1.37)$$

A similar set of moment equations defines a radial second-moment diameter, applicable to beams with cylindrical symmetry described by a weighted sum of Laguerre–Gaussian functions. Here the pinhole x-scan profile is split in half at the centroid point $\langle x \rangle$, and the half profile is taken as the radial variation of the cylindrically symmetric beam. In the transverse radial coordinate plane (r, θ), the origin is the center of the beam spot defined by the centroid $(\langle x \rangle, \langle y \rangle)$ given by the rectangular first moments, Equation 1.33.

$$\text{Zeroth – moment or total power } P = \int_{0}^{2p}\int_{0}^{\infty} I(r,q)r\mathrm{d}r\mathrm{d}q \qquad (1.38)$$

$$\text{Radial second moment } \langle r^2 \rangle = \left(\frac{1}{P}\right)\int_{0}^{2p}\int_{0}^{\infty} r^3 I(r,q)\mathrm{d}r\mathrm{d}q \qquad (1.39)$$

$$\text{Variance of the distribution } s_r^2 \equiv \langle r^2 \rangle \qquad (1.40)$$

$$\text{Radial second-moment diameter } D_{2\sqrt{2}sr} = 2\sqrt{2}s_r \qquad (1.41)$$

This last equation derives from the requirement that the linear and radial variances are related[6] by:

$$s_x^2 + s_y^2 = s_r^2. \qquad (1.42)$$

Then for a cylindrically symmetric mode $\sigma_x = \sigma_y$, yielding $2\sigma_x{}^2 = \sigma_r{}^2$ or $\sigma_x = (1/\sqrt{2})\sigma_r$. Since for a fundamental-mode beam $2w = 4\sigma_x$, from the radial mode description of that beam, there results[6] $2w = 4(1/\sqrt{2})\sigma_r = 2\sqrt{2}\,\sigma_r$, which is Equation 1.41. By mixing modes, combinations of Hermite–Gaussian modes can be made to have the same irradiance profiles as Laguerre–Gaussian modes, and vice versa. Therefore, for compactness the symbols $D_{4\sigma}$ or M_{4s}^2 will be used for either linear or radial second-moment quantities unless there is a need to specifically distinguish a quantity as a radial moment.

1.6.3.6 Sensitivity of $D_{4\sigma}$ to the Signal-to-Noise Ratio of the Profile

The experimental difficulties in evaluating these integrals with noise on the profile signal come from the weighting by a high power of the transverse coordinate in the second-moment calculation, by the square in the linear case [Equation 1.34], and by the cube in the radial case [Equation 1.39]. Take as an example a measurement of a fundamental-mode spot with a CCD camera, using 256 counts to digitize the irradiance values, and 128 counts used to digitize half the integration range of the transverse coordinate. In the linear case, one noise count (0.4% noise) at the edge of the range—at the 128th transverse count—is weighted by the factor $1 \times (128)^2 = 16{,}384$ in the integration, versus 256×1 counts for the central peak. The contribution of this single noise count is 64 times that of the pixel at the central peak in the integration. In the radial case, the one noise count at the limiting transverse pixel makes a contribution $(128)^3/256 = 8192$ times that of the pixel at the central peak. A discussion of the high sensitivity of the second-moment diameter to noise on the wings of the profile is given in Reference 12. There the second-moment and knife-edge methods are compared for five simulated modes, and the knife-edge found to be considerably more forgiving and in agreement with common expectations.

To manage this sensitivity to noise, it is essential that both some measure of the detector's background illumination and noise be subtracted from the signal, and that the integration from the beam centroid outward be truncated at the edges of the illuminated region. Both means reduce the effect of noise on the wings of the profile.

A distinction is made between subtraction of background, the detector's readout with the beam blocked, and subtraction of the baseline, the noise floor of the dark detector. Because of the high directionality of laser beams, typically the background can (and should) be reduced to insignificance by inserting an aperture near the laser (blocking concomitant light) and adding a light-shielding tube to the detector (blocking ambient light).

There are differences of opinion as to the best method for subtracting the noise floor with CCD cameras, but recommended here is what is termed "thresholding." From either a dark camera frame or preferably, from the nonilluminated corners of the signal frame, a standard deviation is computed for this measured noise, and three times this value subtracted uniformly from the signal frame before data analysis. This avoids taking the difference between one random noise frame (the background frame) from another (the noise on the signal frame), which often just adds noise.

To set the integration truncation limit, the beam radius is estimated (typically by a diameter-measurement method less sensitive to noise) and the integration is carried out over the range of from three to four (estimated) beam radii. The constancy of the computed second-moment diameter is observed over this range. Then integration limits are set just wide enough to yield a stable second-moment value. When the width setting is judged to be correct, the measurement should be repeated to check reproducibility.

Other problems with CCD cameras that can look like noise are that they are subject to drift, response nonlinearity and nonuniformity, "bleeding" of signal to adjacent pixels, and low damage threshold requiring attenuation not only to prevent signal saturation but to protect them as well. For these reasons, coupled with the need for analysis software to read them out, cameras are best purchased from dealers who have assessed these problems and will stand behind their instrument's measurement accuracy.

In one commercial instrument[9] two additional checks are provided to assess the effect of noise on the radial second-moment calculation done on a pinhole single line scan. The first check compares the second-moment diameter calculated from the right half profile, to that from the left half profile. If the beam is indeed cylindrically symmetric and the contribution from noise on the profile is negligible, the ratio of these two results should be near unity. The second check is an option in the calculation called "noise-clip ON/OFF." In the wings of the 256 count wide profile where the signal is near zero, noise counts vary the trace above and below the average dark level, and the lowest noise pixels acquire a negative sign when the linear baseline (between the means of the 20 points on either end of the scan line) is subtracted. This is desirable, these negative noise pixels help cancel positive ones, but it is straightforward for the processor in the instrument to clip these pixels to a zero value with the "noise-clip" option turned ON. The size of the resulting change in the calculated second-moment diameter provides a test of how large the contribution is from noise in the wings.

It is also recommended when measuring a second-moment diameter to vary the sources of noise on the laser beam. Check that the resonator alignment is peaked, the sources of microphonics impinging on the laser are minimized, the laser is warmed up and bolted down to the stable table, and so forth, and watch for variations in the second-moment diameter. A more complete analysis[9] of the effect of noise on diameter measurements showed that the standard deviation over the mean of ten repeated second-moment diameter measurements was five to ten times larger than that for knife-edge measurements of the same beam at (low) signal-to-noise levels from 50 down to ten. With these precautions required in interpreting $D_{4\sigma}$ results, it is fair to say that the second moment as currently implemented is not a "machine readable" diameter definition.

1.6.3.7 Reasons for $D_{4\sigma}$ Being the ISO Choice of Standard Diameter

Since there is considerable experimental difficulty in measuring second-moment diameters, why is this definition the one adopted[11] as the standard by the International Organization for Standards? The primary answer is that this definition is the one best supported by theory. The general theories of the propagation of ray bundles[4,6,19] are based on the Fourier transform relationship[6] between the irradiance distribution and angular spatial-frequency distribution. These show two essential requirements are met if the beam width is defined by the second-moment diameter [Equation 1.36]. The beam width is rigorously defined[6] for all realizable beams [excluding only those with discontinuous edges,[6] for which the integration Equation 1.34 may not converge] and the square of this width (the variance) increases as a quadratic function of the free space propagation distance away from the waist. That is, $D_{4\sigma}(z)$ increases with z according to the hyperbolic form (Equation 1.16a). All other diameter definitions gain legitimacy in propagation theory by being shown to be proportional to the second-moment diameter.

Another important feature of the second-moment diameter is that the beam quality (M^2 values) calculated using it turn out to be integers for either the pure, rectangular-symmetry Hermite–Gaussian modes, or the pure, cylindrical-symmetry Laguerre–Gaussian modes. Thus not only for the fundamental mode is $M_{4s}^2 = 1$, which happens by definition,

but for the next higher-order mode, the donut mode, $M_{4s}^2 = 2$, and so on counting up by unity each time the mode order increases. In general[6] the formulas are:

$$\text{Hermite – Gaussian modes} \quad M_{4s}^2 = (m+n+1) \tag{1.43}$$

$$\text{Laguerre – gaussian modes} \quad M_{2\sqrt{2}s}^2 = (2p+l+1) \tag{1.44}$$

where *m*, *n* are the order numbers of the Hermite polynomials, and *p*, *l* the order numbers for the generalized Laguerre polynomials associated with the modes as before (Equation 1.1). For the six modes shown in Figure 1.2, of increasing order from (a) to (f), the values are $M_{4\sigma}^2$ = 1, 2, 3, 3, 4, 4 respectively. The integers $(m+n+1)$ or $(2p+l+1)$ are termed the mode order numbers, and they determine as well the mode's optical oscillating frequency. Modes with the same frequency are termed degenerate. As the mode order number increases, the degree of degeneracy increases, there being three degenerate pure modes each for $(2p+l+1) = M^2 = 5$ or 6, four for $M^2 = 7$ or 8, five for $M^2 = 9$ or 10, and so on. The diameters of the pure modes in second-moment units are just the square root of the mode order numbers times the fundamental-mode diameter (by Equation 1.3):

$$\text{Pure Hermite – Gaussian modes} \quad \frac{D_{4s}}{2w} = \sqrt{m+n+1} \tag{1.45}$$

$$\text{Pure Laguerre – Gaussian modes} \quad \frac{D_{2\sqrt{2}s}}{2w} = \sqrt{2p+l+1} \tag{1.46}$$

Another consequence of the pure modes having integer values of beam quality is that for mixed modes, the M_{4s}^2 value is a simple power-weighted sum of the integer M_{4s}^2 values of the component modes. Finding integers like this in a physical theory is strong indication that the quantities have been defined and measured "the way nature intended."

Another reason for the ISO Committee's choice of $D_{4\sigma}$ as the diameter standard is that the committee members were aware that conversion formulae were available to permit diameters measured according to the other definitions to be put in standard form. These formulae are discussed in the next section.

The last line of Table 1.1 refers to CCD camera properties. A CCD camera together with frame-grabber electronics and appropriate software can be a universal instrument capable of providing diameter measurements according to any or all of the definitions. Affordable cameras do not provide as large a dynamic range for irradiance levels (useful range ~1000:1) compared to that for a silicon detector (~10^4) but good variable attenuators are readily available[33] to allow camera operation just below saturation to make the most of the range that exists. Spatial resolution of 5 micron per pixel may be inadequate for direct measurement of focused beams but flexibility, ease of use, and quick access to colorful 2-D irradiance maps make it an attractive choice for beam diameters large enough to fill an adequate number of pixels. Imaging optics can be used if necessary to measure smaller beams. If improvements in CCD cameras continue at their recent pace, they are likely to become superior to all the older methods of measuring beam diameters.

1.6.3.8 Diameter Definitions: Final Note

It is important to emphasize that the M^2 model can be applied using any reasonable definition of beam diameter as long as the definition is used consistently both in making measurements and interpreting calculated values. Results will then be meaningful and reliable.

In fact, there can be cases where it is important to use a "nonstandard" diameter definition. For example, there is a trend toward steeper sides and flattened tops as M^2 increases. The effect becomes pronounced for M^2 values above ten and at 50 or more, profiles can be aptly described[5] as a "top hat" shape. The diameter of such a beam becomes unambiguous and it makes sense to abandon the standard definitions ($D_{4\sigma}$, D_{86}, etc.) and just measure the diameter of the "top hat" cylinder. The good news is that for such beams, pinhole scans would show the diameter at half-maximum irradiance to be insignificantly different from that at the $1/e^2$ level. The aperture size that passes 86.5% of the total power will not provide as meaningful a result in this situation as the aperture that transmits 95% of the power. The latter would likely be little different in size from the one that passes 98%. Curve fitting to a series of D_{95} measurements will yield a set of valid parameters describing the beam but this defines a new "currency" and one must stay consistent and not mix these diameters with those arrived at by a different method or definition.

1.6.4 Conversions between Diameter Definitions

For a diameter conversion algorithm to be widely applied, it must be normalized, with the natural normalization being the diameter of the fundamental mode generated in the same resonator as the measured beam, the embedded gaussian. Using Equation 1.3, this essentially changes the problem of converting diameters into one of converting M^2 values.

The conversion rules that are now part of the ISO beam widths document[11] were first derived empirically and later found to have theoretical support. They apply to cylindrically symmetric modes generated in a resonator with a circular limiting aperture and an approximately uniform gain medium. In this case, if $M^2_{2\sqrt{2}s}$ is known, then the mixture and relative amplitudes of the modes oscillating can also be reasonably estimated.

1.6.4.1 Is M² Unique?

Determining the fractions of the pure modes in a mixture for a cylindrically symmetric beam from the beam quality alone seems unlikely at first, because the beam quality M^2 is not unique in the mathematical sense. Consider the case of a beam with $M^2 = 1.1$ in second-moment units. An experienced laser engineer might guess the likely composition is 90% fundamental mode ($M^2 = 1$) and 10% donut mode ($M^2 = 2$) to give $M^2 = (0.9) + 2(0.1) = 1.1$ for the mixed mode, and she/he would be right. For a beam of $M^2 = 5$ however, the problem is much harder. The number of possible modes above threshold makes for a large range of possible mix fractions within the $M^2 = 5$ constraint.

Our empirical results showed, however, that for the class of lasers with round beams described just, M^2 was unique at least up to values of $M^2_{4s} = 3.2$.[14] In these resonators, diffraction losses and spatial mode competition in saturating the gain determine the mixed-mode composition. As the circular limiting aperture is opened—as the Fresnel number of the resonator is increased—some modes grow and others decrease in a predictable and reproducible way, such that for each M^2 there is a unique known mode mixture. Furthermore, this knowledge has allowed us to establish mathematical rules for interconversion of beam diameters between the various measurement definitions.

1.6.4.2 Empirical Basis for the Conversion Rules

We acquired the empirical data[14] by using an argon ion laser set up to give beams with a large range of M^2 values as a function of the diameter of the circular mode-limiting

aperture. By varying this aperture diameter and the gain—the latter by adjusting the laser tube's current—values of $M^2_{2\sqrt{2}s}$ from 1 to 2.5 were covered with the green line at 514 nm; the upper limit was increased to 3.2 by changing to the higher gain of the 488 nm blue line. As the blue line was generated in the same resonator, the blue beam diameters here could be scaled by multiplying the square root of the ratio of the wavelengths, a factor of 1.027, for comparison to the green line diameters. The beam from this laser was split to feed an array of monitoring equipment. A radio frequency photodiode and rf spectrum analyzer gave how many modes and what mode orders were oscillating. Profiles were recorded with a commercial slit and pinhole profiler[8] and a commercial beam propagation analyzer[9] to obtain knife-edge diameters, M^2_{ke}, and radial second-moment diameters. A CCD camera and software computed the variable aperture diameter. In front of the camera, a lens provided a known (1.47 times) magnification to fill an adequate number of pixels, and a variable attenuator set the light level.

As the laser's internal aperture was opened and the beam diameter enlarged, the mode spot alternated from one with a peak in the center to one with a dip at the center in over one-and-a-half cycles as shown in the profiles of Figure 1.10b. Seven aperture settings were chosen spanning the range of M^2 values, two giving the highest central peaks (A and E), two at the deepest dips (C and F), and three transitional ones (AP, named the "perturbed A-mode", B, and D). The full set of diagnostic data at these settings was recorded. Knowing the number of modes oscillating and the mode orders at each setting from the rf spectrum, trial mode mixtures were assumed. The resulting theoretical profiles were adjusted[14] to match the experimental pinhole profiles. An example is Figure 1.2, where the theoretical mixed-mode profile, (g), is matched to experimental profile, (h), which is the same as Mode E in Figure 1.10b.

Once the TEM^*_{0n} modes were included[14] in the mode mix, good matches of profiles were found. These modes are like the donut mode, for which $n = 1$, but with increasingly larger holes in the center as their order ($n + 1$) increases. Because they have $p = 0$ they are "all null" (nearly zero in amplitude) in the middle. They make the most of the r^3 weighting factor in the second-moment integral to reach a given second-moment diameter $Mw = \sqrt{(2p + l + 1)}\, w$ at the smallest radius, resulting in the lowest tails[14] to their profiles of all modes of the same-order number. They thus have the lowest diffraction loss for a limiting circular aperture and always oscillate first among pure modes of the same order as the Fresnel number of the resonator is increased. It was noted in Reference 20 that in this aperture-opening process there was a gradual extinction of a mode of lower order soon after a mode of next higher order reached threshold. This is clearly a gain competition effect won by the higher-order mode. A possible physical reason of general applicability discussed in Reference 20 was that the larger spatial extent of the higher-order mode provided access to a region of gain not addressed by the competing lower-order mode.

The final mode fractions for the seven mixed modes were determined using a Mathematica function called SimpleFit made available by Wolfram Research. These fractions are plotted in Figure 1.10a as a function of the resultant beam quality M^2_{4s} for the mixed modes. The modes turn on in the order of decreasing diffraction loss as shown by McCumber[21] and then gradually extinguish, as predicted in the preceding paragraph. At each value of M^2_{4s} for this argon ion laser there is a characteristic set of oscillating modes, mode fractions, and mode profiles (Figure 1.10). Here for every M^2 value there is a unique mixture of modes. From all the data gathered, simple conversion rules given in the next section between diameter definitions were derived. Over the range measured of $M^2_{4s} = 1$ to 3.2, the error to convert knife-edge, slit, and variable aperture diameters to second-moment diameters was ±2% (one standard deviation). This is a ±4% error in converting M^2. The error was ±4% for conversion of pinhole diameters to second-moment diameters.
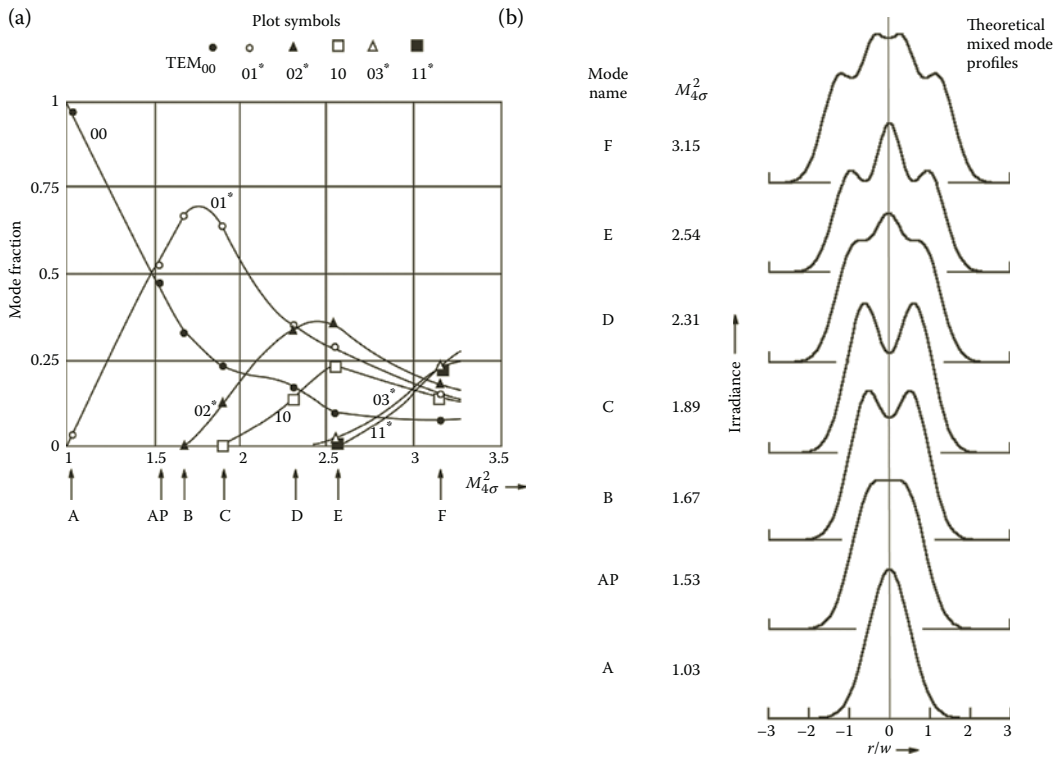
FIGURE 1.10
Observed mode fractions for a beam from a resonator with a limiting circular aperture. As the aperture diameter increases M_{4s}^2 follows, with the mode fractions changing in a characteristic fashion as higher-order modes come above threshold. (a) The mode fractions as a function of M_{4s}^2. (b) The computed pinhole profiles and their $M_{4\sigma}^2$ values for the characteristic set of mixed modes A to F measured to determine the mode fractions.

We then tested the rules on other lasers[14] within this M^2 range and found that knife-edge diameter measurements converted to second-moment diameters agreed with directly measured second-moment diameters within ±2%. The conversion error is defined as the fraction in excess of unity of the $D_{4\sigma}$ diameter obtained by the conversion rule, over that obtained directly from the variance of the irradiance profile, expressed in percent. The knife-edge diameter conversion subsequently was tested on three other gas lasers at $M_{4s}^2 = 4.2, 7.5,$ and 7.7 and found to remain valid to ±2%. However, a test[25] on a pulsed Ho:YAG laser at $M_{4s}^2 = 13.8$ gave a conversion error of −9%; this is thought to be due to the strong transient thermal lensing in this medium affecting the spatial gain saturation. This consistency in the face of an extrapolation by a factor of two indicates that these conversion rules are fairly robust, valid to the stated accuracy, and that the mixed modes on which they are based exist in this large class of lasers. Apparently, for many lasers, M^2 is unique.

1.6.4.3 Rules for Converting Diameters between Different Definitions

The empirical results showed there was a linear relationship between $M_i = \sqrt{M_i^2}$ and the square root of the second-moment beam quality $M_{4s} = \sqrt{M_{4s}^2}$, where M_i is the square root

of the beam quality obtained by method "*i*" and *i* can signify any of the other definitions. Since all the diameter definitions give the same result for the fundamental-mode beam (for which the beam quality is unity) the linear relationship can be expressed with a single proportionality constant $c_{i\sigma}$ in the form:

$$M_{4s} - 1 = c_{is}(M_i - 1) \tag{1.47}$$

for the conversion from the method "*i*" to second-moment quantities. This form ensures that the linear plot of $M_{4\sigma}$ versus M_i passes through the origin with no offset term and that only the slope constant *c* is required to define the relationship.

In the same resonator, the fundamental-mode diameter is given by the ratio of the mixed-mode diameter to *M*. This is true independent of what diameter definition is used, and thus a second relationship is:

$$\frac{D_i}{M_i} = 2w = \frac{D_{4s}}{M_{4s}}. \tag{1.48}$$

Here D_i is the diameter obtained by method "*i*." Substituting Equation 1.48 into Equation 1.47 yields:

$$D_{4s} = \left(\frac{D_i}{M_i}\right)[c_{is}(M_i - 1) + 1]. \tag{1.49}$$

The values of the conversion constants $c_{i\sigma}$ are listed in Table 1.1 to convert from the diameter definitions summarized there to the second-moment diameter, $D_{4\sigma}$.

Since each of the other diameter methods is linearly related to the second-moment diameter, they all are linearly related. The conversion constants between the other methods can be obtained from those for the second-moment conversions. Let one of the other methods be denoted by subscript "*j*." From Equation 1.47 there results:

$$(M_{4s} - 1) = c_{is}(M_i - 1) = c_{js}(M_j - 1)$$

therefore

$$(M_i - 1) = \left(\frac{c_{js}}{c_{is}}\right)(M_j - 1).$$

By definition of a conversion constant for method $i \rightarrow j$,

$$(M_i - 1) = c_{ji}(M_j - 1).$$

Hence:

$$c_{ji} = \left(\frac{c_{js}}{c_{is}}\right). \tag{1.50}$$

This gives the conversion constants between any two methods in Table 1.1, by taking the ratios of their constants for conversion to the second-moment values. Note that Equation 1.50 also implies that $c_{ji} = 1/c_{ij}$, which is also useful.

The values for the $c_{i\sigma}$ constants in Table 1.1 are an improvement over our earlier results[14] that were incorporated in the ISO beam-test document.[11] More experimental data later became available, but also it was realized once the mode fractions were determined experimentally that the conversion constants could then be calculated from theory alone. From the mixed-mode set A to F defined by the mode fractions of Figure 1.10a, each of the

theoretical diameters D_i for the different methods was calculated. By Equation 1.3, these were converted to M_i's. Then plots of $M_{4\sigma} - 1$ versus $M_i - 1$ were least-squares curve fit to determine by Equation 1.47 the values of $c_{i\sigma}$ listed in Table 1.1. The fit for the slope $c_{i\sigma}$, was for one parameter only with the intercept forced to be zero. This gives an internally consistent set of $c_{i\sigma}$'s so that Equation 1.50 is valid.

1.7 PRACTICAL ASPECTS OF BEAM QUALITY *M²* MEASUREMENT: THE FOUR-CUTS METHOD

The four-cuts method means measuring the beam diameter at four judicious axial positions, the minimum number—as explained in this section—to permit an accurate determination of M^2. To execute this method well, several subtleties should first be understood.

The simplest way to measure M would be to take the ratio of the mixed-mode beam diameter to that of the embedded gaussian as by Equation 1.3, $M = W/w$, except that the embedded gaussian is inaccessible by being enclosed inside the mixed mode. However, both beams have the same Rayleigh range. By measuring z_R and the waist diameter $2W_0$ for the accessible mixed mode, the beam quality is determined through Equation 1.20:

$$M^2 = \frac{pW_0^2}{1z_R}. \tag{1.20}$$

The general approach is to measure beam diameters $2W_i$ at multiple locations z_i along the propagation path and least-squares curve fit this data to a hyperbolic form to determine z_R and $2W_0$. But even by taking this computer-intensive approach, unreliable values will sometimes result unless a number of subtle pitfalls[25] (often ignored) are avoided on the way to good (±5%) M^2 values. The pitfalls are highlighted in *italics* as they are encountered in this discussion.

Well-designed commercial instruments[9] avoid these pitfalls, and a button push yields a good answer. For the engineer performing the measurement on his or her own, and who can start by roughly estimating the beam's waist diameter and location (using burn paper, a card inserted in the beam, or a profiler slid along the propagation axis) a minimum effort, logical, quick method exists, which circumvents the subtle difficulties. This is the method[25] of "four cuts," the subject of this section.

The first pitfall is avoided by realizing that in the M^2 model the beam divergence is no longer determined by the inverse of the waist diameter alone (as it is for a fundamental mode) but has the additional proportionality factor M^2:

$$\Theta = \frac{2M^2 1}{pW_0}. \tag{1.19}$$

The first implication of this additional degree of freedom is that the beam waist must be measured directly, not inferred from a divergence measurement. Consider the propagation plots shown in Figure 1.11a. Several beams are plotted, all with the same values of the ratio M^2/W_0 and therefore the same divergence, but with different M^2 [accomplished by having the Rayleigh range proportional to W_0, see the second form of Equation 1.19 in Section 1.4]. From measurements all far from the waist it would be impossible to distinguish between
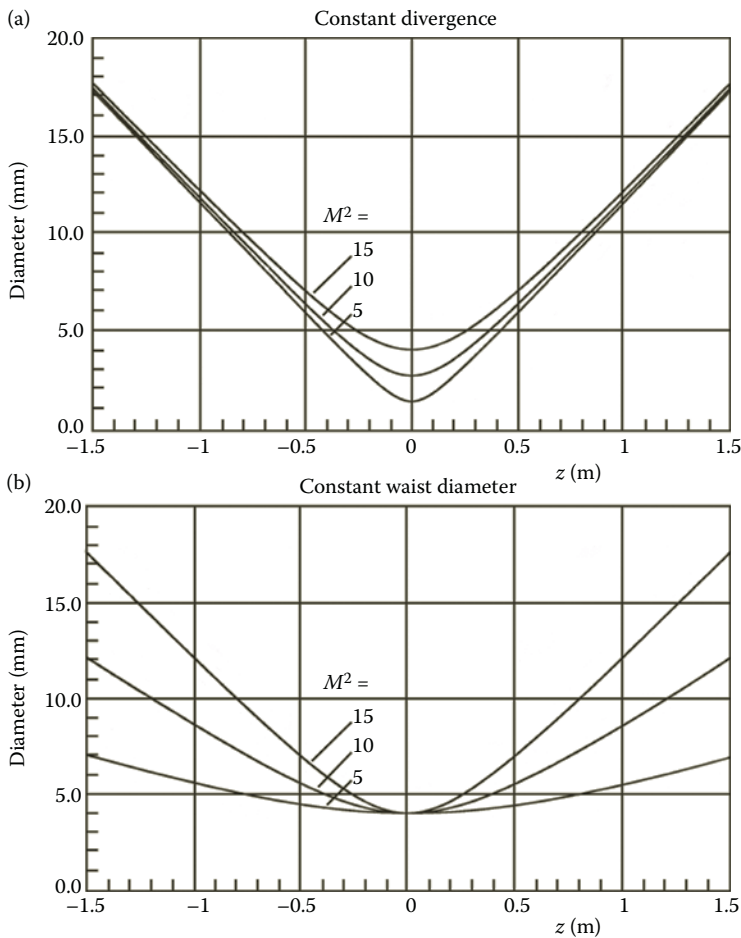
FIGURE 1.11
Beams of constant divergence (a) and constant waist diameter (b) to illustrate the consequences of $M^2 \neq 1$. The beam must be sampled in both near- and far-fields to distinguish these possibilities. The curves are drawn with values appropriate for a beam of $\lambda = 2.1$ microns. (Redrawn from Johnston, T.F., Jr. *Appl. Opt.* 1998, *37*, 4840–4850.)

these curves to determine M^2. On the other hand, in Figure 1.11b are propagation plots for several beams with the same waist diameters but different M^2 and therefore divergences. Here $\Theta \propto M^2$ and by Equation 1.18, $z_R \propto 1/M^2$. Measurements all near the waist could not distinguish these curves to determine the divergences. Both near- and far-field diameter measurements are needed to measure M^2.

Any of the diameter-measurement methods can be used to define an M^2 value, and *the next pitfall is avoided by staying in one currency,* and do not mix, for instance, the knife-edge divergence measurement with the laser manufacturer's quoted $D_{4\sigma}$ (second-moment) waist value. Consistently use the most reliable diameter-measurement method you have available, and in the end convert your results to values in the standard $D_{4\sigma}$ currency.

1.7.1 The Logic of the Four-Cuts Method

The four-cuts method starts with the error estimate for your best method for measuring diameters, and uses that to set the tolerances on all other measurements. Let diameters be determined to a fractional error g,

$$g = \left(\frac{2W_{\text{meas}}}{2W} \right) - 1 \tag{1.51}$$

where $2W_{\text{meas}}$ is the measured diameter, and $2W$ the correct diameter. It is assumed g is small, usually 1%–2%. This will yield a fractional precision h for the beam quality of $h = 3\%$–5% since M^2 varies as the product of two diameters, with a small error added for a required lens transform (discussed in Section 1.7.1.1). The term "cut" is used for a diameter measurement, after the common use of a knife-edge scan cutting across the beam to determine a diameter. Let us define the normalized or fractional propagation distance from the waist as:

$$h(z) = \frac{(z - z_0)}{z_{\text{R}}}. \tag{1.52}$$

Let the fractional error in locating the waist be η_0. For this miss in cut placement in measuring the waist diameter $2W_0$ to cause an error of less than g, Equation 1.16a gives:

$$\sqrt{1 + h_0^2} < g + 1 \quad \text{or} \quad h_0 < \sqrt{2g} \tag{1.53}$$

for $g \ll 1$. If $g = 0.01$, then $\eta_0 < \sqrt{(0.02)} \cong 1/7$. *The tolerable error in locating z_0 is one-seventh of a Rayleigh range for a 1% precision in diameter measurements.*

To locate the waist to this precision, beam cuts must be taken far enough away from the waist to detect the growth in beam diameter with distance. At the waist location the diameter change with propagation is nil; to precisely locate a waist requires observations far from it where the diameter variation can be reliably detected. On both sides away from the waist, cuts must be made at distances of a sizeable fraction of the Rayleigh range.

To find the optimum cut distances, look at the fractional change Q in beam diameter versus normalized propagation distance:

$$Q \equiv \left(\frac{1}{W} \right) \frac{dW}{dh} = \frac{h}{(1 + h^2)}. \tag{1.54}$$

Figure 1.12 is a plot of this function, Equation 1.54, in which it is easy to see that the maximum fractional change of Q occurs at $\eta = \pm 1$. By making cuts within −2 to −0.5 and +0.5 to +2.0 Rayleigh ranges from the waist corresponding to η within these numerical values, 80% of the maximum fractional change is available. This will significantly enhance the reliability of the position determination over that made using diameters from less than 0.5 z_{R} away from the waist. *An accessible span of at least a Rayleigh range centered on the waist is needed for diameter measurements.*

Note that Figure 1.12 highlights the physical significance of the propagation locations one Rayleigh range to either side of the waist. The wavefront curvature is largest in absolute magnitude there, resulting in the fractional change in diameter Q with propagation coordinate z reaching extremes of ±0.5 there as well.

1.7.1.1 Requirement of an Auxiliary Lens to Make an Accessible Waist

Most lasers have their beam waists located internally where they are inaccessible. Therefore, an accessible auxiliary waist related to the inaccessible one is achieved by inserting a lens
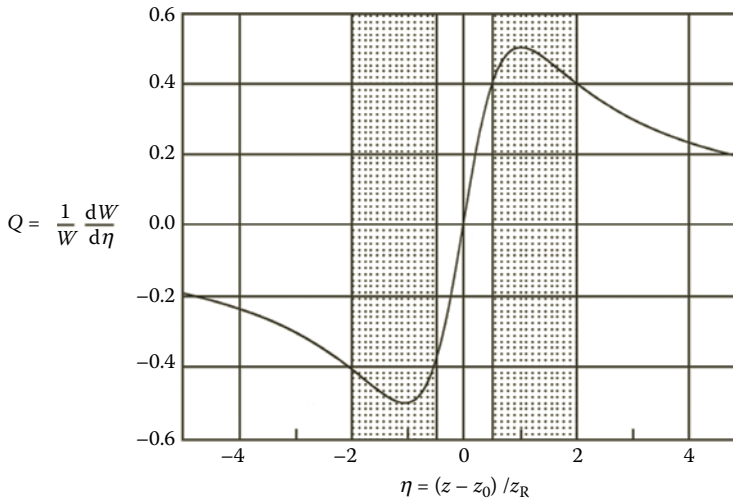
FIGURE 1.12
The fractional change Q in beam diameter as a function of the normalized propagation distance from the waist. Cuts made to locate the waist in the shaded regions benefit from a fractional change of 80% or more of the maximum change. This requires a minimum of one Rayleigh range of access to the beam around the waist location. (Redrawn from Johnston, T.F., Jr. *Appl. Opt.* 1998, *37*, 4840–4850.)

or concave mirror into the beam, and making the M^2 measurement on the new beam. Then the constants found are transformed back through the lens to determine the constants for the original beam. *This requirement to insert a lens, and then transform through the lens back to the original beam constants, is an often-ignored pitfall in making accurate beam measurements.*

The temptation is to use what is available, and just measure the beam on the output side of the output coupler. Usually this means the data is all on the diverging side of the waist. The problem is that nothing in this data constrains the waist location very well. In the curve fit, small errors in the measured diameters will send the waist location skittering back and forth to the detriment of the extrapolation to find the waist diameter. Inserting a lens and making a beam that is accessible on both sides of its waist is a significantly more reliable procedure.

There are three constants $(z_{02}, 2W_{02}, M^2)$ needed to fix the 2-space beam shown in Figure 1.6 for one of the principal propagation planes, so, in principle, only three cuts should suffice, but then one of them would have to be within the range $|\eta_0| < 1/7$. The location of this narrow range $z_{02} \pm z_{R2}/7$ is at this point unknown. Therefore four cuts are used, the first an estimated Rayleigh range z_{R2} to one side of the estimated waist location z_{02}, the second and third at about 0.9 and 1.1 times this estimated Rayleigh range to the other side (see Figure 1.13). These cut locations and the diameters determined there are labeled by their cut numbers $i = 1, 2, 3$. Between z_2 and z_3 there is a diameter that matches $2W_1$ and the location z_{match} of this is determined by interpolation:

$$z_{match} = z_2 + \frac{(z_3 - z_2)(W_1 - W_2)}{(W_3 - W_2)} \tag{1.55}$$

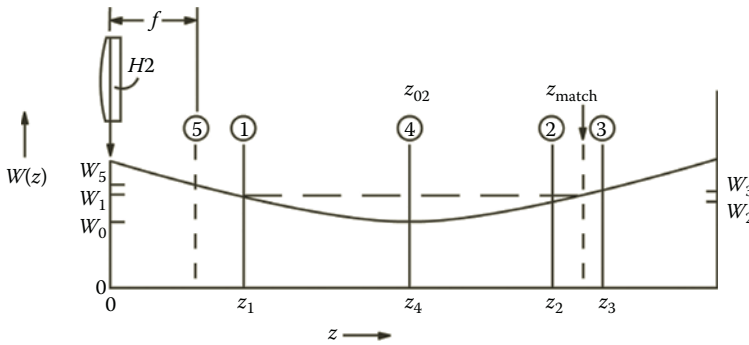$$z_4 = z_{02} = \frac{(z_1 + z_{match})}{2} \tag{1.56}$$

FIGURE 1.13

The four-cuts method. Shown is the beam propagation plot in 2-space, behind the inserted auxiliary lens; the circled numbers indicate the order of the cuts made to locate the waist. The propagation distance z_{match} of the diameter matching that at the first cut at z_1 determines the waist location z_{02} as halfway between these equal diameters. (Redrawn from Johnston, T.F., Jr. *Appl. Opt.* 1998, *37*, 4840–4850.)

The waist is located exactly halfway between z_1 and z_{match}, and the fourth cut is made there at z_4 to directly measure the waist diameter $2W_{02} = 2W_4$ of the 2-space beam and complete the minimum data to determine $M.^2$

1.7.1.2 Accuracy of the Location Found for the Waist

If the locating cuts (1, 2, and 3 of Figure 1.13) are within the ranges specified from $|Q| > 0.4$ and the diameters are measured to the fractional error g, then the error in the normalized waist location η_0 is no worse than $g/Q = 2.5\%$. This is much less (since g is small) than the tolerance $\sqrt{2g} = 14.1\% = 1/7$ determined from inequality Equation 1.53. The measured waist diameter is then correct to the fractional error g.

The fractional error in measurement of diameters g when divided by the fractional change in diameter with normalized propagation Q, yields the fractional error in normalized waist location $\eta_0 = g/Q$. The plot of Figure 1.12 is thus actually a quantitative version of the statement "to precisely locate a null requires observations far from the null" when locating the waist. Diameter measurements inside the range $z_{02} \pm z_R/2$ quickly lose any ability to contribute precision in locating the waist as here Q drops to zero.

There is much value in locating the waist as accurately as the diameter-measurement tolerance will allow in that it reduces the number of unknown constants to be determined by curve fitting from three to two. The number of terms in the curve fit drops by a factor of four, and the remaining terms are made more accurate. Some of these terms depend on the distance from the waist to the ith-cut location, $z_i - z_{02}$, either squared or raised to the fourth power. It is often useful to take a fifth cut at $z_5 = f$ as shown by the vertical dashed line in Figure 1.13. This cross checks the input beam divergence by Equation 1.29 and balances the number of points on either side of the auxiliary waist at z_{02} to improve the curve fit.

1.7.2 Graphical Analysis of the Data

The data, which consists of a table of four- or five-cut locations and their beam diameters for each of the two independent principal propagation planes, is next plotted. A sample plot for the $\lambda = 2.1$ micron Ho:YAG laser beam analyzed in Reference 25 is shown in Figure 1.14. There it was found that with as few data points as required in the four-cuts method, and
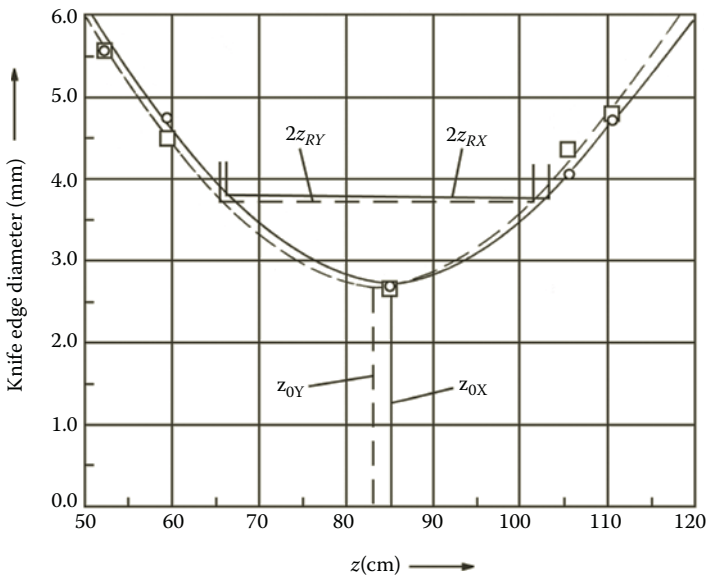
FIGURE 1.14

An example of graphical analysis of propagation data for the auxiliary beam in 2-space. The chords give the Rayleigh ranges for the *x*- and *y*-planes. They are drawn at ordinates on the plot √2 larger than the waist diameters located at z_{0x} and $z_{0y.}$ (Redrawn from Johnston, T.F., Jr. *Appl. Opt.* 1998, *37*, 4840–4850.)

with the initial waist location and Rayleigh range estimates close to the final values (within ~10%), a simple and quick graphical analysis is as accurate as a curve fit.

Generally, with more points as in commercial instrumentation, a weighted least-squares curve fit of the data to a hyperbolic form is required,[25] discussed in Section 1.7.3. The curve fit also generates a sum of residuals for a statistical measure of the goodness of fit.

In the graphical analysis after the points are plotted, smooth curves of approximately hyperbolic form are laid in symmetrically about the known waist locations for each principal propagation plane, here in Figure 1.14 with a French curve. Next, horizontal chords are marked off at heights √2 times the waist diameters $2W_4$ to intersect the smooth curves. The distance between these intersection points on each curve are twice the Rayleigh ranges $2z_{Rx}$, $2z_{Ry}$ respectively, and these lengths are measured off the plot for use in Equation 1.20 with $2W_{0x} = 2W_{4x}$ (and $2W_{0y} = 2W_{4y}$) to determine M_x^2 (and M_y^2) for the auxiliary 2-space beam. For the data of Figure 1.14 the results were $z_{Rx} = 17.6$ cm and $z_{Ry} = 17.8$ cm, resulting in knife-edge beam qualities $M_x^2 = 15.4$ and $M_y^2 = 14.9$.

These results are termed the initial graphical solution and can be improved to give the corrected graphical solution by using the fact that a better estimate of the waist diameter is available than just the closest measured point. By the propagation law, Equation 1.16, if the miss distance of the closest point (Cut 4) is η_0 then the best estimate of the corrected waist diameter is:

$$2W_{02} = \frac{2W_4}{\sqrt{(1 + h_0^2)}}. \tag{1.57}$$

The corrected solution uses the Rayleigh range and waist values from the initial graphical solution in Equation 1.57 to obtain a corrected waist diameter, and plots a chord at a height of $\sqrt{2}$ times this diameter to determine a corrected length $2z_R$ and M^2 from Equation 1.20. In the example of Figure 1.14, the chords shown are the corrected chords; only the y-axis data changed slightly to $z_{Ry} = 17.3$ cm and $M_y^2 = 15.2$. After curve fitting the same data, the fractional rms error (goodness of fit) for the five diameter points were the same at <1.9%.

This good accuracy is a consequence of the four-cuts strategy. The waist diameter is directly measured and if the initial estimate for the Rayleigh range is close, the other cuts give data points near the intersection points of the chords fixing the $2z_R$ values on the plot. The graphical analysis then amounts to an analog interpolation to find the best positions for the intersection points.

There are two last steps. The first is to transform the 2-space data back to 1-space to get the constants for the original beam, using Equations 1.24 through 1.28. This adds a small fractional error to the end result due to the uncertainties in z_{02} and z_{R2}, which contribute a slight uncertainty to the transformation constant Γ of Equation 1.24 (in the example of Reference 25, a 2% error in Γ, 1% additional error in transformed diameters).

The second step is to convert these knife-edge measurements of Figure 1.14 to standard second-moment units as done in Table 3 of Reference 25. The beam of Reference 25 is the one that did not work well with the conversion rules of Section 1.1.5. Instead the conversion of $M_{ke}^2 = 15.4$ to $M_{4s}^2 = 13.8$ was done by comparing measurements at cut 5, the focal plane of the auxiliary lens, of the knife-edge diameter to the second-moment diameter calculated from a pinhole scan. This gave the ratio $D_{ke}/D_{4\sigma} = 1.055$ or a factor of $1/(1.055)^2 = 0.897$ for the M^2 conversion.

1.7.3 Discussion of Curve-Fit Analysis of the Data

A complete numerical example of a full weighted least-squares curve fit to analyze the four-cuts data, or a larger data set, is given in Reference 25 and need not be repeated. There are some subtle pitfalls to avoid in using curve fits on beam propagation data and these are briefly discussed.

A least-squares curve fit is the only general way to account for all the data properly. A common mistake *is to use the wrong function for the curve fit*, which necessitates a discussion of what is the correct one. The fit should be to a hyperbolic form, Equation 1.16, but that is not all. It also should be a weighted curve fit, with the weight of the ith squared residual in the least-squares sum being the inverse square power of the measured diameter $2W_i$.

There are three reasons for this choice of weighting. The first is that in general in a weighted curve fit, the weights[36] should be the inverse squares of the uncertainties in the original measurements. For many lasers, the fractional error in the measured diameter is observed to increase with the diameter; this is probably due to the longer time it takes to scan a larger diameter. The spectrum of both amplitude noise and pointing jitter on a beam tends to increase towards lower frequencies and longer measurement times give this noise a greater influence.

The second reason arises from an empirical study[25] of different weightings one of us did during the development of a commercial M^2 measuring instrument.[9] Amplitude noise was impressed on the beam of a fundamental-mode ion laser with a known $M_{4s}^2 = 1.03$, by rapid manual dithering of the tube current while the instrument's data gathering run[9] was underway. (Note, the ModeMaster[9] collects 260 knife-edge cuts in each of two orthogonal planes in a 30-s "focus" run, generating the beam propagation plots.) The same data was then fitted to a hyperbola five times, with five different weighting factors.

The weights were the measured diameter raised to the nth power, $(2W_i)^n$, with $n = -1$, -0.5, 0, $+0.5$, or $+1$. The weight with $n = 0$ is unity or equal weight for all data points. Data runs were repeated many times with increasing noise amplitude, and the resulting M^2 values for all five weighting schemes were compared each time. The equal or negative power weightings gave stable M^2 values within 3% of the correct value up to 5% peak-to-peak amplitude noise. The positive power weightings $n = +0.5$ gave 4%–5% and $n = +1$ gave 12%–19% errors in M^2 respectively at this noise level. With larger noise amplitudes, the positive power weightings gave errors that grew rapidly and nonlinearly.

A common curve-fitting technique is to use a polynomial fit for the square of the beam diameter versus propagation distance. This may be convenient but it could give an unsatisfactory result. This technique takes advantage of the wide availability of polynomial curve-fit software, and the fact that the square of Equation 1.16 gives a quadratic for $W(z)^2$ as a function of z. However, look at what this does. Let $2W_i$ be the measured ith diameter, and $2W_i'$ be the exact diameter with the small deviation between them $2\delta_i = 2W_i - 2W_i'$. In the W^2 polynomial curve fit, the ith term is

$$(W_i)^2 = (W_i' + d_i)^2 = (W_i')^2 + 2W_i' d_i$$

making the residual

$$(W_i)^2 - (W_i')^2 = 2W_i' d_i.$$

The residual from the exact polynomial curve is weighted in the fit by $2W_i'$, a positive power (+1) of W_i', and so will give unstable results if there is more than a few percent amplitude noise on the beam. At the time of completion of the 1995 ISO document on beam-test procedures, this difficulty with a polynomial curve fit was unrecognized, and a polynomial fit was (incorrectly) recommended there; the 2004 version correctly recommends fitting to a hyperbolic form.[11]

The third reason for an inverse-power weighting is that mathematically the least fractional error results for a ratio quantity like $M^2 = \Theta/\theta_n$ in Equation 1.22 if the fractional errors from the denominator and numerator roughly balance. The residuals from the more numerous cuts far from the waist—the points giving the measurement of divergence Θ, or numerator—would swamp with equal weighting the fewer (or single) cut at the waist—the point(s) giving the divergence of the normalizing gaussian, or denominator. An inverse square weighting approximately halves the influence of the three or four far points, compared to the unity weighting at the waist in the four-cuts method, giving the desired rough balance.

1.7.4 Commercial Instruments and Software Packages

There are three main commercial instruments for measuring beam quality and a host of less well developed others. The first is the original[9,35] system designed as a beam propagation analyzer and believed at this time to be the most fully developed, the ModeMaster™ from Coherent, Inc. The cylindrical scan head (10 cm diameter by 31 cm length) mounts through angle and translation alignment stages to a heavy stable-table post. The basic diameter measurements are achieved with two orthogonal knife-edge cuts. Both principal propagation planes are measured nearly simultaneously on a drum spinning at 10 Hz behind an auxiliary lens. Measurements are restricted to continuous-wave laser beams or high repetition rate (>100 kHz) pulsed lasers. The lens moves to carry the auxiliary beam through the plane of the knife-edges to assemble 260 cuts in each principal propagation plane making up a

pair of propagation plots for the auxiliary beam in a 30-s "focus" pass. A curve fit, with an inverse-diameter weighting, to a hyperbola is done and the fitted parameters are transformed through the lens by the on-board processor to present a data report for the original beam.* The M^2 measurement accuracy is specified at 5% and the waist diameter accuracy at 2%, with a minimum of 100 sample points taken across the profile. The drum also carries two pinholes, each of different diameters, giving pinhole profiles that are processed to give direct second-moment diameters. The instrument also measures beam-pointing stability. Electronic alignment aids are included. In the original instrument operation was controlled through a dedicated electronics console; the current version is driven by a laptop PC.

The second is the ModeScan™ family of instruments from Photon, Inc. Originally intended as an upgrade of a user's existing 10 Hz rotating drum profiler to a beam propagation analyzer, the simplest version is a modular package consisting of a 0.5 m rail to manually translate the profiler in the beam behind a fixed input lens, with software for the user's PC computer that prompts the user for input of position data. When the data fields are filled, the software calculates the M^2 of this auxiliary beam (the same as the input beam's) and transforms the data through the lens for the other input beam constants. Later versions automate the stage drive and data acquisition, expand the software features, and include a new profiler with a selection of five rotation speeds to measure pulsed beams. The latest instrument, the ModeScan™ 1780 model, is a new design in a 26 cm × 18 cm × 8 cm housing gimbal mounted to a 1/2" diameter stable-table post. It incorporates beam splitters behind a fixed input lens to pick off ten sample beams, and direct them to a CCD array camera. The diameters at these ten different throw distances are measured simultaneously and this data fit to a hyperbola to generate the beam constants. Placing ten spots at once on the camera sensor reduces the number of illuminated pixels per spot (a minimum of 15 pixels per diameter measurement is recommended), but enough are lit that the M^2 measurement accuracy of this and the rest of these systems is listed as 4%–5%. This latest model is the only commercial instrument capable of determining beam constants from a single pulse, and therefore of showing pulse-to-pulse beam variations.

The third instrument is the CCD camera-based M^2-200 Beam Propagation Analyzer from Ophir Spiricon, Inc., operating with pulsed or cw laser beams. The original instrument with a 500-mm focal length input lens occupied a 28 cm × 82 cm footprint on the stable table; a new model, the M^2-200s, uses a 300-mm lens and is a half-size version fitting into 26 cm × 44 cm. Here a stepper motor and translation stage on a rail moves an optical delay line behind the fixed input lens to effectively scan the detector surface through the auxiliary beam. The PC computer attached to the system automatically adjusts filter wheel attenuation, subtracts background, sets spot truncation, and calculates the second-moment diameters[37] directly from the CCD profiles. A curve fit to a hyperbola is done[37] with the results transformed back through the input lens to present the constants of the original beam. The M^2 measurement accuracy is given as 5%. This product, with its long focal length input lenses, is positioned to measure the large beams of industrial lasers for process monitoring and control.

1.8 TYPES OF BEAM ASYMMETRY

In the previous sections, the means for the spatial characterization of laser beams were established. This section looks at commonly found beam shapes and others that are

---

* An example of a data report is shown later in Figure 1.17 of Section 1.9.
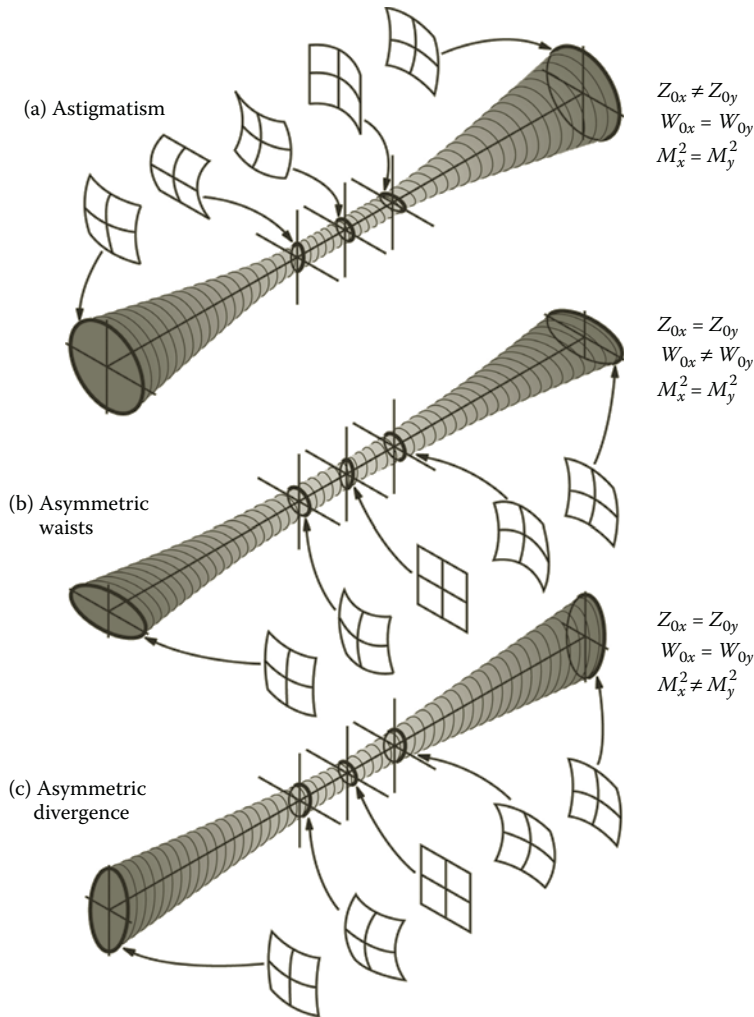
FIGURE 1.15
Depiction of the three-dimensional appearance of the three basic types of asymmetry for a mixed-mode beam: (a) astigmatism, (b) asymmetric waist diameters, and (c) asymmetric divergence. The window insets show the wavefront curvatures along the beam path. (Redrawn from Johnston, T.F., Jr. *Appl. Opt.* 1998, 37, 4840–4850.)

possible. The three common types of beam asymmetry are depicted in Figure 1.15. These are the pure forms but mixtures of all three are common in real beams.

1.8.1 Common Types of Beam Asymmetry

The first is *simple astigmatism* (Figure 1.15a), where the waist locations for the two orthogonal principal propagation planes do not coincide, $z_{0x} \neq z_{0y}$, but $W_{0x} = W_{0y}$, and $M_x^2 = M_y^2$. Because here the waist diameters and beam qualities are the same for the principal propagation planes, so are the divergences, $\Theta_x \propto M_x^2/W_{0x} = M_y^2/W_{0y} \propto \Theta_y$ [see Equation 1.19]. This makes the beams round in the converging and diverging far-fields. At the two waist planes the beam cross sections are elliptical, one oriented in the vertical and the other the

horizontal plane, with the minor diameters equal. Midway between the waists, the beam becomes round like the "circle of least confusion" point in the treatment of astigmatism[28] in geometrical optics. The simple astigmatic beam is characterized by three round cross sections, at the distant ends and midpoint, with orthogonally oriented elliptical cross sections in between.

The window frame insets of Figure 1.15 show the wavefront curvatures, which are spherical in the far-field, cylindrical at the waist planes with one cylindrical axis horizontal, the other vertical, and saddle-shaped at the midpoint between the waists. The wavefront curvatures determine the nature of the focus when a lens is inserted.

Simple astigmatic beams can be generated in resonators with three spherical mirrors, with one used off-axis to give an internal focus,[38] unless there is astigmatic compensation built in as with a Brewster plate of the correct thickness[38] added to the focusing arm. Many diode lasers are astigmatic but with the other two types of asymmetry as well because the channeling effects in the plane parallel to the junction differ from those in the plane perpendicular to it, giving two different effective source points for the parallel and perpendicular wavefronts. Beams formed using angle-matched second harmonic generation can be astigmatic due to walk-off in the phase matching plane of the beam in the birefringent doubling crystal. The diode lasers in laser pointers frequently have a large astigmatism, as large as the Rayleigh range for the high-divergence axis.

The next is *asymmetric waists* (Figure 1.15b), where the waist diameters are unequal. Because of the different waist diameters but with equal beam qualities, in the far-fields where divergence dominates, the cross sections are elliptical with the long axes of the ellipses (shown as horizontal) perpendicular to the long axis of the waist ellipse (here vertical). In between there are round cross sections at planes symmetrically placed around the waist location—the same geometry as in Figure 1.15a, with the ellipses and circles interchanged. The wavefronts are plane at the waist, and ellipsoidal everywhere else, with curvatures at the round cross sections in the ratio of the square of the waist diameters.

Lasers having an out-of-round gain medium are likely to produce beams with asymmetric waists. A solid-state laser pumped from the end by an elliptical beam from a diode laser is an example. Mode selection is by the combined effects of gain aperturing and absorption in the unpumped regions. The resonant beam shape will mimic the geometry of the pumped region. Beam walk-off from angle-matched nonlinear processes can also produce asymmetric waists.

The third type is *asymmetric divergence* (Figure 1.15) where the beam qualities differ in the principal propagation planes to give proportionally different divergence angles, $\Theta_x \propto M_x^2 \neq M_y^2 \propto \Theta_y$, but $W_{0x} = W_{0y}$, and $z_{0x} = z_{0y}$. The simplest description of this beam is that the mode in one principal propagation plane is of higher order than in the other. In the far-field, cross sections are elliptical as in case (b), but the beam is round only at the waist plane. The wavefronts are plane at the waist and ellipsoidal everywhere else and the Rayleigh ranges are different in the two principal propagation planes.

A CW dye laser using a high-viscosity dye jet provides an example of pure asymmetric divergence.[39] The pump-beam spot was round, but the heat it deposited in the dye stream was cooled differentially by the flow. In the flow direction the temperature gradient was smoothed by the forced convection but in the other direction a more severe thermal gradient existed, causing an aberration that resulted in $M_{4sy}^2 = 1.51$ for that plane compared to $M_{4sx}^2 = 1.06$ for the plane parallel to the flow with negligible aberration. Because of the round pump beam, waist asymmetry was only $2W_{0y}/2W_{0x} = 1.06$.

1.8.2 The Equivalent Cylindrical Beam Concept

Beams with combinations of these asymmetries can be depicted with superposed (x, z) plane and (y, z) plane propagation plots as shown in Figure 1.16a. More generally there is a propagation plot $W(\alpha, z)$ for each azimuth angle α around the propagation axis z. The angle α is measured from the x-axis and $W(\alpha, z)$ lies in the plane containing α and the z-axis. The three-dimensional beam envelope shown in Figure 1.15a, b, or c is called the beam caustic
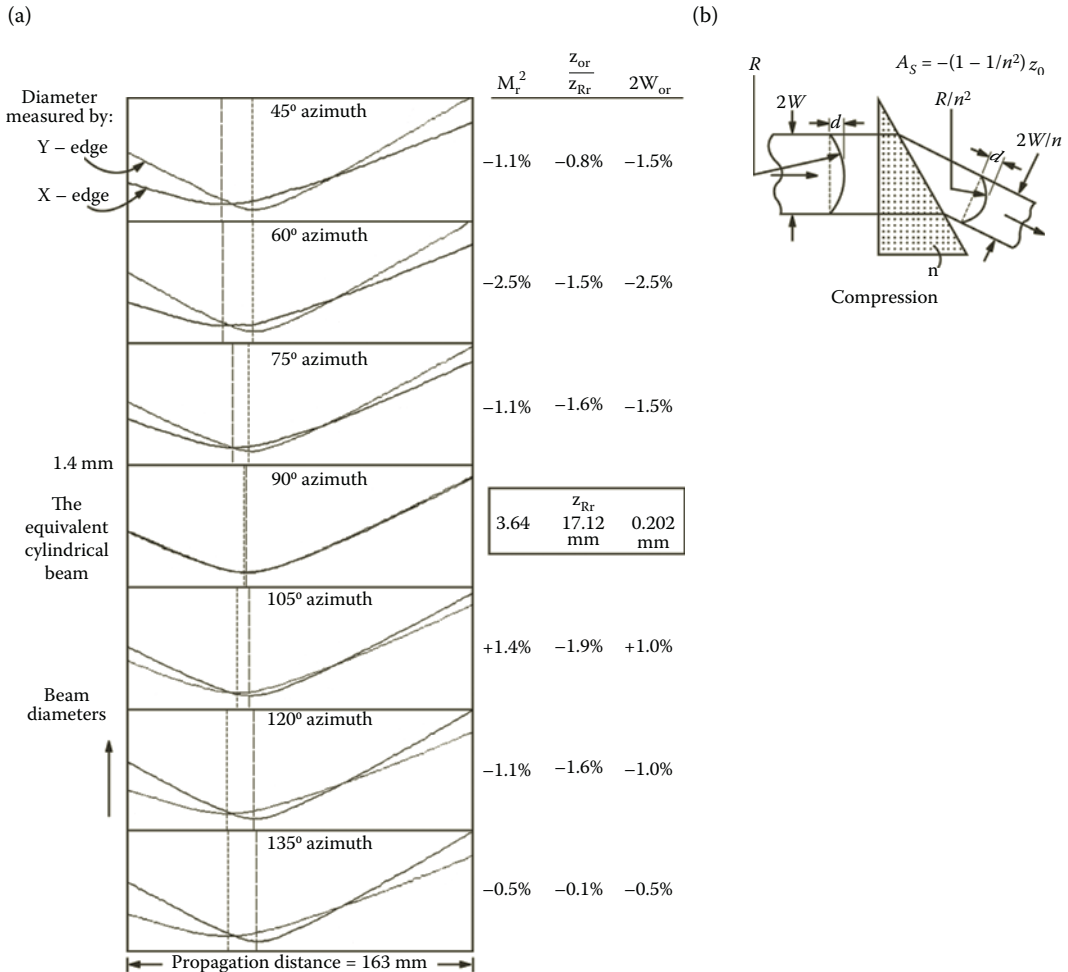

FIGURE 1.16
(a) Experimental propagation plots with beam diameters measured by orthogonal knife-edges for a beam with both astigmatism and waist asymmetry. The percentage variation of the constants of the equivalent cylindrical beam, computed from the plots for each instrument azimuth, is listed in the right-hand columns. The small variations demonstrate the constants are independent of the azimuth of the two orthogonal cutting planes intersecting the beam caustic surface. The constants of the equivalent cylindrical beam, in the box, correspond to the cuts at an instrument azimuth of 90°. (b) Diagram showing how a half-Brewster prism introduces both astigmatism A_s and waist asymmetry $W_{0y}/W_{0x} = n$ to the beam.

surface, and is swept out by $W(\alpha, z)$ as the azimuth angle α rotates through a full circle from 0 to 2π.

For beams with combinations of moderate asymmetries, it is convenient to define an *equivalent cylindrical* beam. This is a beam with cylindrical symmetry—with a round spot for all z—and the real-beam asymmetries are treated as deviations from this round beam. The constants defining this equivalent cylindrical beam are the best average of the beam constants for the two independent principal propagation planes. Many problems can be treated with just this simpler equivalent beam. In particular, it has been predicted theoretically (A.E. Siegman, personal communication, 1990) and demonstrated experimentally[40] that the centered circular aperture computed to give 86% transmission for the equivalent cylindrical beam, has this same transmission for the out-of-round real beam. The minimum aperture sizes for the real beam after propagation in free space can be computed using just the three equivalent cylindrical beam constants. Because the equivalent cylindrical beam is round for all z like the radial modes discussed in Section 1.4, the subscript r is used to denote its constants, and the beam is sometimes called the *equivalent radial mode*.

The equivalent cylindrical beam is best understood by considering the plots of Figure 1.16a. These were measured with the ModeMaster beam propagation analyzer.[9] The profiler built into this instrument uses two knife-edge masks at right angles to each other. They are mounted on a rotating drum at 45° to the scan direction of the drum. This arrangement is equivalent to a vertical and a horizontal knife-edge, each scanned at $1/\sqrt{2}$ times the actual scan speed of the drum, when the analyzer head's azimuth angle is set to 45° to align one knife-edge with the horizontal. Each run to measure beam diameters versus propagation distance produces two propagation plots for the diameters at right angles to the two edges. Normally the analyzer head's azimuth angle is adjusted to record the propagation plots in the two principal planes of the beam. For Figure 1.16a, the analyzer head's azimuth angle was incremented in 15° steps through 90° and new sets of propagation plots recorded for each increment, generating the seven plots shown.

The asymmetric beam of Figure 1.16a was formed by inserting a Brewster-angle half prism[41] in the cylindrically symmetric beam Mode E of Figure 1.10b and Figures 1.2g and h. The prism was oriented as in Figure 1.16b to produce a compression of the beam diameter in one dimension in the (x, z) plane. The prism thus introduces astigmatism and waist asymmetry to the beam. From Figure 1.16b the incoming wavefront of radius of curvature R has a sagitta of the arc, $d = W^2/2R$, which remains unchanged upon the transit of the prism while the beam diameter is compressed. For the Brewster-angle prism, it can be shown[41] that the exiting beam diameter is smaller by the factor $1/n$, where n is the index of refraction of the prism material, here silica with $n = 1.46$. The radius of curvature exiting the prism is thus R/n^2. The M^2 of the beam is unchanged in traversing the prism. From these three conditions both the reduced waist diameter in the x-direction and the astigmatic distance introduced in the exiting beam can be determined [using Equations 1.16b and 1.17b and a little algebra] to be $2W_0/n$ and $A_s = (z_{0y} - z_{0x}) = -(1 - 1/n^2)z_0$, where z_0 is the propagation distance from the input waist location to the prism.

The propagation plots of Figure 1.16a are for the directly measured internal beam, behind the lens of the beam propagation analyzer. These were used because the beam diameter and propagation distance scales of the internal plots remain the same as the instrument azimuth is varied and this facilitates comparison of the plots. Notice in the top plot [45° instrument azimuth], because the internal propagation plots are shown, the axis with the n-times larger divergence and $1/n$-times smaller waist is the y-axis, interchanged with the compressed x-axis of the external beam in the beam-lens transform of Section 1.5.

As the instrument azimuth angle moves around from the initial 45° value (which measures the principal propagation planes for this beam) to 90°, the plots from the two orthogonal edges coalesce into a single "average" curve, then separate with continuing azimuth increments. The plots at 135° are identical to the 45° plots with the *x*-edge and *y*-edge curves interchanged. The dashed and dotted vertical lines on each plot locate the waists for the *x*-edge and *y*-edge curves, respectively. The beam constants for the symmetric, 90°-azimuth plots are those for the equivalent cylindrical beam.

To visualize this process of cutting the beam caustic surface with two orthogonal planes, then rotating the azimuth of the cutting planes, look at Figure 1.15c. The initially vertical (*y*-edge) plane is cutting the caustic in its highest divergence plane, and moves towards a lower divergence $W(\alpha, z)$ plot as the azimuth is incremented. The initially horizontal (*x*-edge) plane is cutting the caustic in its lowest divergence plane, and moves toward a higher divergence $W(\alpha, z)$ plot as the azimuth is incremented. When the cutting planes reach 45° azimuth to the principal planes of the beam, the orthogonal propagation plots match as they would for a round beam with no asymmetries.

Siegman[42] gives the following expressions for the beam constants of the equivalent cylindrical beam in terms of the six constants of the real beam:

$$z_{0r} = \left(\frac{M_x^4 W_{0y}^2}{M_x^4 W_{0y}^2 + M_y^4 W_{0x}^2} \right) z_{0x} + \left(\frac{M_y^4 W_{0x}^2}{M_x^4 W_{0y}^2 + M_y^4 W_{0x}^2} \right) z_{0y} \tag{1.58}$$

$$2W_{0r}^2 = W_{0x}^2 + W_{0y}^2 + \left(\frac{1}{p^2} \right) \left(\frac{M_x^4 M_{0y}^4}{M_x^4 M_{0y}^2 + M_y^4 M_{0x}^2} \right) \mathcal{I}^2 (z_{0x} - z_{0y})^2 \tag{1.59}$$

and

$$M_r^4 = \frac{W_{0r}^2}{4} \left(\frac{M_x^4}{W_{0x}^2} + \frac{M_y^4}{W_{0y}^2} \right). \tag{1.60}$$

The columns of numbers in Figure 1.16a demonstrate that the beam constants of the equivalent cylindrical beam are the same when computed from plots for any azimuth, a necessary condition for the equivalent cylindrical beam concept to be useful. The equivalent cylindrical beam quality, waist location, and waist diameter were computed for each azimuth increment from the plots shown, and normalized to the constants measured for the 90° azimuth shown in the box. The percentage errors for these measurements are given in the three columns; the magnitudes of all errors are no larger than 2.5% and are within the instrument measurement tolerances.

From Equations 1.58 and 1.59 for an astigmatic beam the equivalent cylindrical waist lies between the two astigmatic waists, and the square of the cylindrical waist diameter exceeds the sum of the squares of the two astigmatic waist diameters. For a beam with no astigmatism ($z_{0x} = z_{0y}$) the equivalent cylindrical constants become:

$$2W_{0r}^2 = W_{0x}^2 + W_{0y}^2 \tag{1.61}$$

$$M_r^4 = \left(\frac{W_{0x}^2 + W_{0y}^2}{4W_{0x}^2} \right) M_x^4 + \left(\frac{W_{0x}^2 + W_{0y}^2}{4W_{0y}^2} \right) M_y^4. \tag{1.62}$$

For a beam with no astigmatism and no waist asymmetry the equivalent cylindrical beam quality is:

$$M_r^4 = \frac{(M_x^4 + M_y^4)}{2}.$$ (1.63)

A beam of this type with different values of M_x^2 and M_y^2 will have a round spot at the waist plane, but not in the far-field as illustrated in Figure 1.15c.

1.8.3 Other Beam Asymmetries: Twisted Beams, General Astigmatism

The shape of a beam caustic surface is determined by the straight-line paths of rays where they emerge at the margin of the particular beam. Such shapes are all examples of ruled surfaces and those depicted in Figure 1.15 are hyperboloids. In principle, any paraxial ensemble of light rays (i.e., a beam) will be enclosed by a ruled surface. Another example is a taut ribbon, and these surfaces can be twisted. Imagine the shapes of Figure 1.15 as taut, flexible membranes. Start with a shape similar to Figure 1.15b except with *all* of the cross sections being horizontally elongated ellipses (a beam with *both* asymmetric waists and divergence). Mentally rotate the far-field ellipses to vertical, the distant one by +90° and the foreground one by −90° azimuth, while keeping the waist ellipse horizontal. In propagating from $z = -\infty$ to $+\infty$ the elliptical cross sections of this beam twist through 180° of azimuth. Such a twisted beam can be physically realized and is said to have *general astigmatism*.[15,16] Here all spots can be ellipses,[15] a waist location is defined by a cross section having a uniform phasefront,[16] and the Rayleigh range is defined as the distance of propagation away from the waist that increments[16] the Gouy phase by $\pi/4$.

Such beams are produced by nonorthogonal[5] optical systems, for example, two astigmatic elements in cascade with azimuth angles that differ by something other than 0° or 90°. Rays in the (x, z) and the (y, z) planes are coupled and cannot be independently analyzed. The general theory for spatial characterization of such beams uses ray matrices weighted by the Wigner density function[4,17] averaged over a four-dimensional geometrical optics "phase space." Rays are described by 4×1 column vectors; each vector gives the position x, y and slope u (= θ_x), v (= θ_y) of the ray at the location z along the propagation axis. There are 16 possible second-order moments of these variables; they propagate in free space with a quadratic expansion law.[6,26] The square of the second-moment diameter D_{4s}^2 is such a second-order moment and this is theoretical support this diameter definition enjoys. The beam matrix P, the 4×4 array of these 16 second-order moments, then fully characterizes the beam with general astigmatism.

The 16 possible second moments can be listed as

$$\langle x^2 \rangle; \langle xy \rangle; \langle xu \rangle; \langle xv \rangle; \quad \langle y^2 \rangle; '\langle yx \rangle'; \langle yu \rangle; \langle yv \rangle;$$

$$\langle u^2 \rangle; '\langle ux \rangle'; '\langle uy \rangle'; \langle uv \rangle; \quad \langle v^2 \rangle; '\langle vx \rangle'; '\langle vy \rangle'; '\langle vu \rangle'.$$

However, by symmetry $\langle xy \rangle = \langle yx \rangle$, and so on; so that only ten of these are independent and in the list those in single quotes are redundant. The moments containing only spatial variables $\langle x^2 \rangle$, $\langle xy \rangle$, $\langle y^2 \rangle$ can be evaluated as the variances of irradiance pinhole profiles in the proper direction; the $\langle xy \rangle$ profile is at 45° to the x- or y-axes. The moments containing the angular variables cannot be evaluated directly, but are found by inserting optics (usually a cylindrical lens) and measuring downstream irradiance moments at appropriate propagation distances.

From these moments, beam constants are calculated. The first six are the familiar set $2W_{0x}$, $2W_{0y}$, z_{0x}, z_{0y}, M_{0x}^2, M_{0y}^2. The other four address the rate of twist of the phasefront and spot pattern with propagation distance, the generalized radii of curvature of the wavefronts and the orbital angular momentum[43] carried per photon by the beam. Beam classes are defined[44] by values of invariants calculated from the ten second-order moments. A simple association of the resultant shapes for the beam caustic envelopes with each class is *not* immediately available from these definitions.

Twisted-phase beams have been generated by inserting an appropriate computer designed diffractive optical element into an ordinary beam.[45,46] The first report of a beam from an "ordinary" laser (not one deliberately perturbed to produce a twisted phase) that required all ten matrix elements for its characterization, appeared in 2001.[18] Nonorthogonal beams can arise in nonorthogonal resonators,[5] such as in twisted-ring resonators. Until instruments are developed[17] to measure all elements of the beam matrix P and then are applied to characterize beams from many of lasers, the fraction of laser beams with general astigmatism will not be clear. The techniques discussed in previous sections are the methods that can be used together with various auxiliary optics to measure these second-order moments.

1.9 APPLICATIONS OF THE M^2 MODEL TO LASER BEAM SCANNERS

This section applies previous concepts and results to determine appropriate specifications for a laser used in an industrial scanning system, by working backward from the beam properties needed at the work surface. This example shows how parts of the M^2 model interact in the design of a system and how the model can be applied to solve simpler individual problems.

1.9.1 A Stereolithography Scanner

The example analyzes an actual stereolithography scanning system, shown in Figure 1.17. A multimode ultraviolet beam (of 325 nm wavelength) writes on a liquid photopolymer surface under computer control, selectively hardening tiny volume elements of plastic to build up a three-dimensional part. After a 1/4-mm thick slice (cross section) of the part is completed, a jack supporting the growing part inside the vat of liquid lowers the part and brings it back up to 1/4 mm below the surface for the next slice to be written. Parts of great complexity can be formed overnight directly from their CAD-file specifications. This process is called stereolithography and has created what is termed the "rapid-prototyping" industry.

Beam characteristics for the laser are shown in the data report of Figure 1.17. Much of the scanning system design elements used in this analysis are available in the literature.[47–50] The beam from the laser is expanded in an adjustable telescope that also focuses the spot on the liquid surface at the optimum beam spot size[47] for the solidification process, $2W_{02} = 0.25$ mm ± 10%, measured with a slit profiler.

Notice first that the system geometry defines a maximum M^2 for the laser beam in this application. The rapidly moving y-scan mirror benefits from a low moment of inertia and has a small diameter A. This is the minimum diameter needed to just pass the expanded beam incident on the mirror, making the beam diameter at the mirror $2W_A$ be smaller than A only by some safety factor γ or $2W_A = A/\gamma$. From this mirror, the beam is focused on the
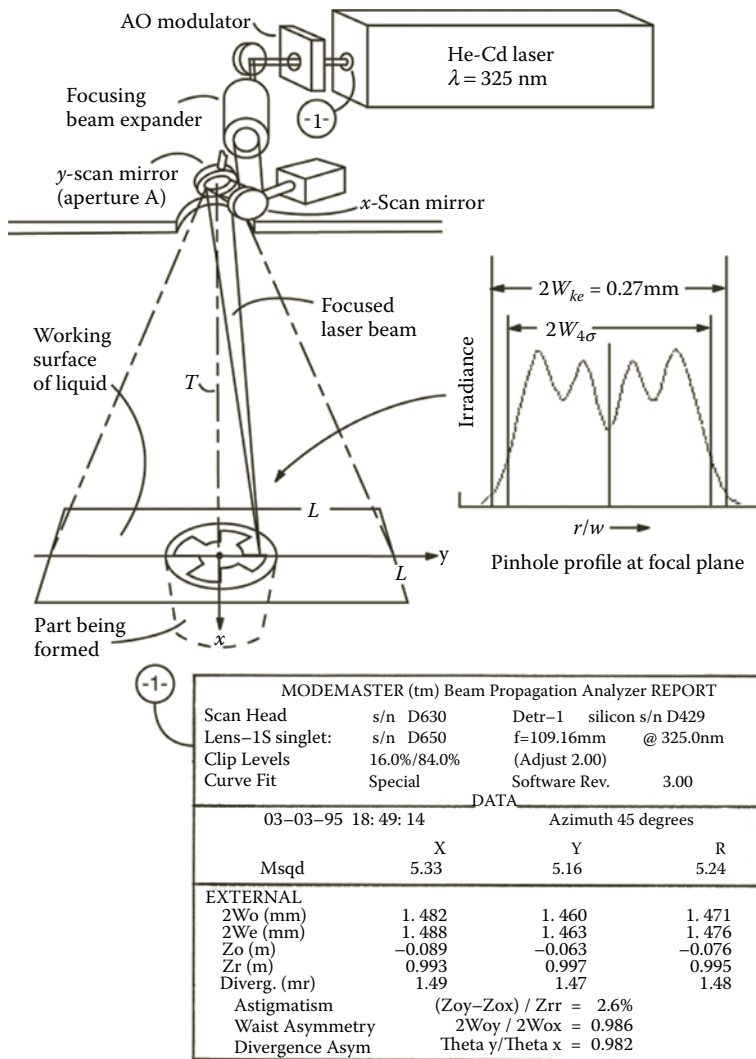
FIGURE 1.17

A stereolithography scanning system based on a helium–cadmium ultraviolet multimode laser. The pinhole focal plane profile (upper inset) shows the irradiance profile at the surface of the liquid photopolymer. The printout from the commercial beam propagation analyzer (lower inset) applies to the beam at the laser output, location (–1-). Laser data courtesy of CVI Melles Griot, Inc.

liquid surface below at the throw distance T shown in Figure 1.17. The maximum convergence angle of the beam focused on the vat surface is therefore $\Theta_2|_{max} = A/\gamma T$ (a larger angle would overfill the y-scan mirror). The focused beam waist diameter, given in the previous paragraph is $2W_{02}$; a diffraction-limited beam of that waist diameter—a normalizing gaussian—has a divergence angle of $\theta_n = 2\lambda/\pi W_{02}$. This defines a maximum M^2 for this application by Equation 1.22 of $M^2|_{max} = \Theta_2|_{max}/\theta_n = \pi W_{02}A/2\lambda\gamma T$.

This may be evaluated in two different ways. From scaling a photograph of the system,[49] an estimate for T can be made as between 0.6 m and 0.7 m, or a reasonable value is

$T = 0.65$ m. The y-scan mirror diameter A is likely that of a small standard substrate, such as $A \sim 7.75$ mm, and a likely safety factor is about $\gamma \sim 1.5$, yielding $2W_A \sim 5.2$ mm and $M_{\text{slit}}^2 \sim 4.8$. This rough estimate is refined in Section 1.9.5. This beam quality is given in slit units because this is the currency for the focal diameter at the vat; the assumption being made that the value of $2W_A$ used is also in slit units for this estimate.

Alternatively, the beam diameter A/γ can be determined working back from the vat to the y-scan mirror since it is known that the laser of Figure 1.17 is designed for this application and that the measured data (given in Figure 1.17) are within the nominal beam specifications. Those measurements are in knife-edge currency[9] (see Section 1.9.4). Once the knife-edge waist diameter at the vat is found, so is $\theta_n = 2\lambda/\pi W_{02}$ and $2W_A = T\Theta = TM^2\theta_n$, all in knife-edge units. A diameter conversion is thus required to bring the diameter at the waist into knife-edge units for a consistent currency.

1.9.2 Conversion to a Consistent Knife-Edge Currency

By Equation 1.48, for any diameter definition i, the ratio D_i/M_i equals the embedded gaussian diameter $2w$ and therefore the conversion from slit to knife-edge diameters at the vat is just $D_{\text{ke}} = D_{\text{slit}}(M_{\text{ke}}/M_{\text{slit}})$. The square root of the beam quality M_{ke} is known from the report (Figure 1.17), $M_{\text{ke}} = \sqrt{(5.24)} = 2.289$. Here the R or "round beam" column value was used, the equivalent cylindrical beam constants as discussed in Section 1.9.4. To determine M_{slit}, use is made of the expression just above Equation 1.50 relating any M_i to any M_j for different diameter definitions i and j. This conversion formula requires knowledge of the M^2 of the starting currency j; M^2 is known here only for knife-edge units, so $j = $ knife-edge. The desired ending currency is in slit units, $i = $ slit. Then Equation 1.50 gives the required conversion constant, in terms of the conversion constants to second-moment diameters from Table 1.1, as:

$$c_{\text{ke}\to\text{slit}} = c_{ji} = \frac{c_{js}}{c_{is}} = \frac{c_{\text{ke}\to s}}{c_{\text{slit}\to s}} = \frac{(0.813)}{(0.950)} = 0.856.$$

This gives $(M_{\text{slit}} - 1) = 0.856(M_{\text{ke}} - 1) = 1.103$, thus $M_{\text{slit}} = 2.103$ and $M_{\text{slit}}^2 = 4.423$.

Then Equation 1.48 yields the focal diameter at the vat in knife-edge units, $2W_{02\,\text{ke}} = 0.272$ mm, a knife edge to slit diameter ratio of 1.088 for this beam. The "normalizing gaussian" divergence angle above is then evaluated as $\theta_n = 1.521$ mr, the maximum convergence angle is larger than θ_n by $M^2 = 5.24$, making the beam diameter at the y-scan mirror be $2W_A = TM^2\theta_n = 5.180$ mm, all in knife-edge units.

For comparison, using the knife-edge to second-moment conversion constant from Table 1.1 and Equation 1.47 gives the second-moment beam quality and beam diameter at the vat of $M_{4\sigma}^2 = 4.19$ and $2W_{02}|_{4\sigma} = 0.243$ mm. The irradiance profile in Figure 1.17 shows the relative size of the second-moment diameter to the knife-edge diameter. It is evident that the former would require a larger safety factor γ than the latter if used in estimating a safe minimum mirror aperture.

For the remainder of this section, diameters are all in knife-edge units and for simplicity the subscripts indicating this are suppressed.

1.9.3 Why Use a Multimode Laser?

What is the advantage of a multimode laser in this application? First, the critical optic, the scan mirror of diameter A, required for the larger multimode beam diameter is

of reasonable size, so it is possible to use such a larger beam here. The significant advantage is seen from the product data sheet for this laser (CVI-Melles Griot Model 74 Helium–Cadmium laser): with single isotope cadmium used in the laser (the X models on the data sheet) the multimode power is 55 mW, the fundamental-mode power is 13 mW, *a ratio of 4.2 times*. With natural isotopic mix cadmium, the numbers are 40 mW and 8 mW, *a ratio of 5 times*. Thus the laser's output power is roughly proportional to its M^2, making the multimode laser considerably smaller and less expensive than a fundamental-mode laser would be at the power level required for this application.

1.9.4 How to Read the Laser Test Report

Notice that the beam quality number used in Section 1.9.2 was from the "*R*" column (for radial or round mode) of the REPORT shown in Figure 1.17. These are the beam constants for the equivalent cylindrical beam discussed in Section 1.8.2, the best theoretical average[6,40,42] of the X and Y column constants for the two principal propagation planes on the report. Since there is less than 4% difference between M_x^2 and M_y^2, it is appropriate to use the average values in the *R* column and treat the beam as round for this exercise. The fact that the report is all in knife-edge units is signified by the "clip-levels" line reading 16%/84% (adjust: times 2.00) as explained in Reference 9. The EXTERNAL label means these constants are for the original beam external to the instrument, after the lens transform has been done from the constants measured for the INTERNAL auxiliary beam inside. Next, listed for the two principal propagation planes in the X and Y columns, and the equivalent cylindrical beam in the *R* column, are: the external beam waist diameter $2W_0$; the beam diameter $2W_e$ at the instrument's reference surface (the beam entrance plane at the front bezel, designated plane B); the waist location z_0 measured from plane B (the zero point of the beam propagation axis) with negative values being back towards the laser;[9] the Rayleigh range z_R; and the beam divergence. Lastly, the significant beam asymmetry ratios are listed, with the astigmatism normalized to the equivalent cylindrical beam's Rayleigh range.

The whole report is readily converted into a different currency with a diameter a factor of τ larger, if desired, by multiplying the M^2 values by τ^2, the diameters and the divergence by τ, and leaving the z_0, z_R, values and the asymmetry ratios unchanged.

1.9.5 Replacing the Focusing Beam Expander with an Equivalent Lens

The beam expander of Figure 1.17, when left at a fixed focus setting, can be replaced with an equivalent thin lens placed at the *y*-scan mirror location, with the laser moved back a distance z_{01} from the lens, as shown in Figure 1.18b. The propagation over z_{01} expands the beam to match the spot size at this mirror. To find z_{01}, Equation 1.16a is used to find the propagation distance for the required beam expansion of $\rho = 2W_A/2W_{01} = (5.180 \text{ mm})/(1.471 \text{ mm}) = 3.521$. Here $2W_{01}$ is the laser's waist diameter in 1-space, on the input side of the equivalent lens, from the report of Figure 1.17. From Equation 1.16a, $\rho = \sqrt{[1 + (z_{01}/z_{R1})^2]}$, yielding $z_{01}/z_{R1} = \sqrt{[\rho^2 - 1]}$, and with the 1-space Rayleigh range $z_{R1} = 0.995$ m taken from the report, there results $z_{01} = 3.359$ m, also shown in Figure 1.18b.

The equivalent lens focal length $f_{equiv} \equiv f$ is next properly chosen to bring this beam to a focus at the vat. Since the waist diameters on either side of the equivalent lens are known, by Equation 1.26 the required transformation constant Γ is also known. This then gives by
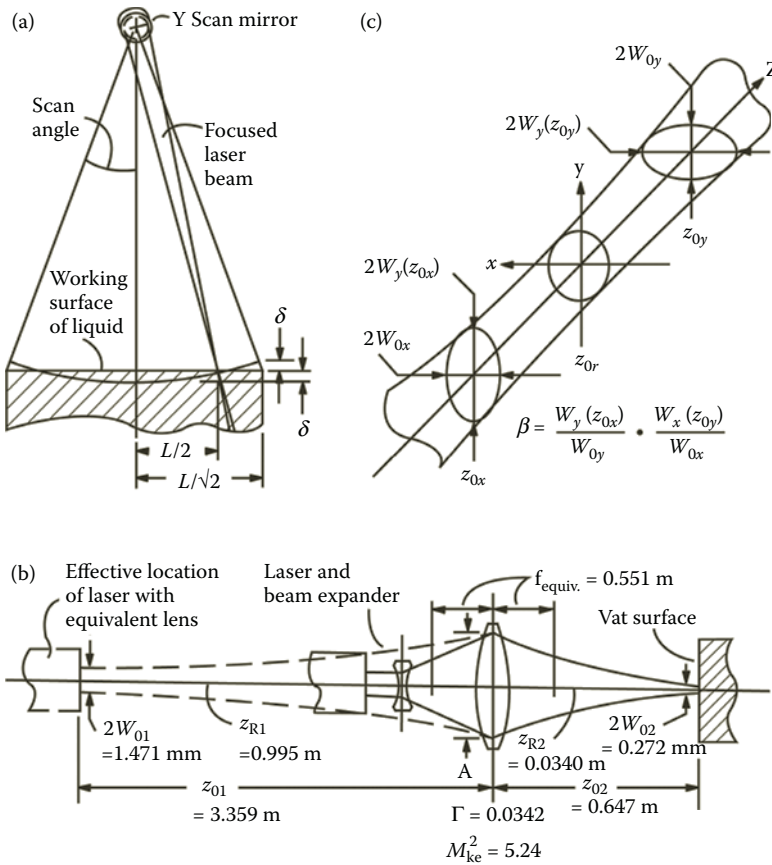
FIGURE 1.18

Analysis of the stereolithography system. (a) Optimum focus to minimize spot-size change over the working surface. For clarity the scan angle shown is larger than the actual scan angle of ±15°. (b) Replacement of the focusing beam expander with an equivalent thin lens of focal length f_{equiv}. Parameters for the equivalent lens transform of the unperturbed beam are shown. (c) Definition of an out-of-roundness parameter $\beta \equiv$ (quadratic ratio of astigmatic diameters) for the focal region of an astigmatic beam. The quantities shown are all for the vat-side focal region or 2-space.

Equation 1.24 a quadratic equation solvable for f:

$$\Gamma = \left(\frac{2W_{02}}{2W_{01}} \right)^2 = \frac{f^2}{[(z_{01} - f)^2 + z_{R1}^2]^2} \tag{1.64}$$

yielding

$$f = z_{01} \left[\frac{\Gamma}{\Gamma - 1} \right] \left\{ 1 \pm \sqrt{1 - \left(\frac{\Gamma - 1}{\Gamma} \right) \left[1 + \frac{z_{R1}^2}{z_{01}^2} \right]} \right\}. \tag{1.65}$$

Inserting $2W_{02} = 0.272$ mm and $2W_{01} = 1.471$ mm in Equation 1.64 gives $\Gamma = 0.03422$, and this in Equation 1.65 produces $f = f_{equiv} = 0.5511$ m. In what follows, a precise value of z_{02} that corresponds to T in Figure 1.17 is needed and the only value at hand is the previous estimate of $T = 0.65$ m for the y-scan mirror to vat distance. A precise value is needed

consistent with the quantities used in the lens transform z_{01} and f_{equiv} above. This is given by Equation 1.28, now that Γ, z_{01}, and f_{equiv} are known, as $z_{02} = 0.6472$ m. This also shows that the original estimate of T was reasonable. The nominal values for the quantities involved in the equivalent lens transform are shown in Figure 1.18b. The effects of perturbing the nominal values are studied in Section 1.9.7.

1.9.6 Depth of Field and Spot-Size Variation at the Scanned Surface

With the equivalent lens transform defined, questions relating the input beam to the scanned beam can be answered. First, what is the amount of defocus over the scanned field? From Equation 1.27, the Rayleigh range in 2-space at the vat is $z_{R2} = \Gamma z_{R1} = 3.404$ cm. The longest radial scan distance is to the corner of the square vat of side length $L = 250$ mm, Figure 1.17, a distance of $\sqrt{2}L/2 = L/\sqrt{2}$. The variation in distance from the y-scan mirror to the corner of the square vat's working surface is $\sqrt{[T^2 + (L/\sqrt{2})^2]} - T = 2.371$ cm, or 0.696 times the vat side Rayleigh range. By Equation 1.16a, from the center of the vat to the far corner, the spot size of the beam will grow by a factor of $\sqrt{[1 + (0.696)^2]} = 1.219$. The simplest way to reduce this range is to focus the beam at the middle of a side edge of the vat, at a radial distance of $L/2$ [see Figures 1.17 and 18a]. This splits the defocus amount 2δ equally among the corners and the middle, to 11% maximum change in spot diameter on the liquid surface over the scanned field. Equivalently, the liquid level could be raised by δ. However, for simplicity the focal distance will be left at T here for the remainder of this analysis.

1.9.7 Laser Specifications to Limit Spot Out-of-Roundness on the Scanned Surface

Next, the inverse lens transform, from the vat side back to the laser side, is applied to transfer scanning beam specifications into laser-beam specifications. From Equation 1.31, for the transform equations going from 2-space to 1-space, use the inverse transformation constant $\Gamma_{21} = 1/\Gamma_{12}$.

Since the lens transformation constant depends on both the input waist location and the Rayleigh range, in general beams without astigmatism but with some other asymmetry when transformed become astigmatic as the results in the remainder of this section show. The plan, starting from the nominal, round, equivalent cylindrical beam of Figure 1.17 transformed to a round beam at the vat, is to perturb the beam at the vat to have a ~10% out-of-round spot. This beam is then transformed back to the laser to see which 1-space variables change and by how much, to account for the perturbation on the scanned side of the lens. The 10% out-of-roundness of the scanned spot is deemed acceptable because that amount of growth in spot diameter over the field was found acceptable above.

The perturbations are made as equal changes of opposite sign in the two independent propagation planes. For example, 10% out-of-roundness due to waist asymmetry is accomplished with a +5% change in W_{02y} and a −5% change in W_{02x}. The resulting changes in the 1-space beam constants are not completely symmetrical, due to the nonlinearity of the beam-lens transform. The effect of perturbing a constant in only one principal propagation plane is given directly in the tables by the percentage changes, the column in parentheses, for 1-space shown for that plane. Because the propagation planes are independent, so are the percentage changes in each plane.

1.9.7.1 Case A: 10% Waist Asymmetry

Assume $2W_{02x}$ is reduced 5% (to 0.259 mm), and $2W_{02y}$ is raised 5% (to 0.286 mm) to give a *waist asymmetry* different from unity by 10%. To calculate the effect on the input beam, first

the new Rayleigh ranges for the beam at the vat are found as $z_{R2x} = 3.088$ cm (reduced 10%) and $z_{R2y} = 3.753$ cm (increased 10%). For each of these a new $1/\Gamma$ for the inverse transform is computed from Equation 1.24, followed by the remaining constants through Equations 1.26 through 1.28. The results for the 1-space beam constants, and their percentage change shown in parentheses are summarized in Table 1.2. The initial value of $1/\Gamma$ is 29.2259. In the table A_s/z_{Rr} stands for the normalized astigmatism $A_s/z_{Rr} = (z_{01y} - z_{02x})/z_{R1r}$, where z_{R1r} is the Rayleigh range in 1-space of the equivalent cylindrical beam.

The +10% pure waist asymmetry at the vat (i.e., accompanied by no astigmatism or divergence asymmetry) for the most part transforms through the lens to a corresponding +8% waist asymmetry at the laser. The same is true for the divergence asymmetry. The different waist diameters at the vat, generate different Rayleigh ranges there and in the lens transform produce a –12% normalized astigmatism at the laser. Specify the laser to have less than these asymmetries to keep the scanned beam out-of-roundness below 10%.

1.9.7.2 Case B: 10% Divergence Asymmetry

Here it is assumed M_x^2 is reduced 5% and M_y^2 is increased by 5% to give a +10% change in the 2-space *divergence asymmetry* without changing the waist asymmetry $W_{02y} / W_{02x} = 1$. By Equation 1.18 or 1.19 the Rayleigh ranges on the vat side of the lens change inversely with their M^2 to make them $z_{R2x} = 3.088$ cm, $z_{R2y} = 3.753$ cm. Applying Equations 1.24 through 1.28 to each principal plane produces Table 1.3, the results for the 1-space beam constants and their percentage change.

The divergence asymmetry of the beam at the vat carries through the lens to the laser, and implies some astigmatism is necessary at the laser (but half as much as Case A) to get pure divergence asymmetry at the vat.

TABLE 1.2

Laser Constants Corresponding to *10% Waist Asymmetry* at the Scanned Surface

Quantity	x	y	Ratios, y/x	Ratio was:
$1/\Gamma$	29.816 (+2.0 %)	28.540 (–2.4%)	0.957 (–4.3%)	1
$2W_{01}$ (mm)	1.415 (–3.8%)	1.526 (+3.8%)	1.079 (+7.9%)	1
z_{01} (m)	3.416 (+1.7%)	3.294 (–2.0%)	$A_s/z_{Rr} = -12.3\%$	0
z_{R1} (m)	0.921 (–7.5%)	1.071 (+7.7%)	1.163 (+16.4%)	1
Θ_1 (mr)	1.537 (+3.8%)	1.425 (–3.7%)	0.927 (–7.3%)	1

TABLE 1.3

Laser Constants Corresponding to *10% Divergence Asymmetry* at the Scanned Surface

Quantity	x	y	Ratios, y/x	Ratio was:
$1/\Gamma$	28.896 (–1.1%)	29.532 (+1.0%)	1.022 (+2.2%)	1
$2W_{01}$ (mm)	1.463 (–0.6%)	1.478 (+0.5%)	1.011 (+1.1%)	1
z_{01} (m)	3.328 (–0.9%)	3.389 (+0.9%)	$A_s/z_{Rr} = +6.1\%$	0
z_{R1} (m)	1.033 (+3.8%)	0.958 (–3.8%)	0.927 (–7.3%)	1
Θ_1 (mr)	1.416 (–4.3%)	1.544 (+4.3%)	1.091 (+9.1%)	1

1.9.7.3 Case C: 12% Out-of-Roundness across the Scanned Surface Due to Astigmatism

A little discussion is required to define an out-of-roundness parameter for the focal region of an astigmatic beam in general before applying the concept to the focus at the vat. It has already been shown (Section 1.9.6) that the path length to the liquid surface changes over the scanned field by 2.37 cm. This path change causes a spot-size variation of 21.9% if the spot is focused at the center of the vat, and 11% if focused at the middle of a vat edge to reduce the variation across the scanned surface. On top of this, there is an out-of-round change in the spot, if the beam is astigmatic. The fastest change of shape with z of the elliptical spots in a beam with pure astigmatism, Figure 1.16a, takes place between the two astigmatic waists in the focal region, which is the working region for the beam in the stereolithography system. Suppose the astigmatic distance $z_{02y} - z_{02x}$ is matched to the path length change of 2.37 cm but the edge focus is used to split this distance (see Figure 1.18a). This makes the largest path between an astigmatic waist and the liquid working surface anywhere in the field be 1.19 cm. Then from Equation 1.16a and Figure 1.18c

$$\frac{W_{2x}(z_{02r})}{W_{02x}} = \sqrt{1 + \left(\frac{1.19}{3.40}\right)^2} = 1.059$$

where z_{02r} is the equivalent cylindrical beam waist location halfway between the x and y waist locations. The spot at the vat only goes to 5.9% out-of-round, but the orientation of the out-of-round ellipse is along the y-axis in the corners where the liquid is below z_{02r} and along the x-axis in the middle of the field where the liquid is above z_{02r}. This can have an unpleasant effect on the part, because the textures of the x- and y-formed surfaces differ. Therefore, an adequate out-of-round parameter for the focal region of an astigmatic beam can be defined (Figure 1.18c) as $\beta \equiv$ (quadratic ratio of astigmatic diameters), where the product of the x-direction and y-direction out-of-round diameter ratios is

$$b = \left[\frac{W_{2y}(z_{02x})}{W_{02y}}\right]\left[\frac{W_{2x}(z_{02y})}{W_{02x}}\right]. \tag{1.66}$$

The ratios are evaluated at the two astigmatic waist locations as indicated in Equation 1.66. From the above, $\beta = (1.059)^2 = 1.12$ for an astigmatic distance equal to the scanned depth of field; this is taken here as "12% out-of-roundness due to astigmatism" for the final example.

The calculations proceed in this example with $z_{02x} = (0.6472 - 0.0119)$ m = 0.6353 m and $z_{02y} = (0.6472 + 0.0119)$ m = 0.6591 m, with the other 2-space beam parameters left at their unperturbed values of Figure 1.18b. Table 1.4 gives the results for the 1-space beam.

TABLE 1.4

Laser Constants Corresponding to a *12% Out-of-Roundness* due to Astigmatism ($\beta = 1.121$) across the Scanned Surface

Quantity	x	y	Ratios, y/x	Ratio was:
$1/\Gamma$	36.794(+25.9%)	23.708 (−18.9%)	0.644 (−35.6%)	1
$2W_{01}$ (mm)	1.651 (+12.2%)	1.345 (−9.9%)	0.803 (−19.7%)	1
z_{01} (m)	3.651 (+8.7%)	3.110 (−7.4%)	$A_s/z_{Rr} = -52.4\%$	0
z_{R1} (m)	1.253 (+25.9%)	0.807 (−18.9%)	0.644 (−35.6%)	1
Θ_1 (mr)	1.318 (−11.0%)	1.641 (+10.9%)	1.216 (+24.6%)	1

This type of asymmetry, transformed back to the laser side of the equivalent lens, is devastating to the 1-space beam constants. More correctly, it would take devastating input beam characteristics to produce this large a "quadratic-ratio-of-astigmatic-diameters" parameter. There are large percentage changes in 1-space waist asymmetry, astigmatism, and divergence asymmetry. Actual lasers with asymmetries this large would be rejected by the laser manufacturer, and the scanner manufacturer would not have to deal with them. Lasers with sufficient beam asymmetry to give $\beta = 1.12$ at the scanned surface would not make it into the field.

In conclusion, the strictest specifications found for the laser to meet upon incoming testing were from Case A, yielding 10% waist asymmetry at the vat surface. To stay below this out-of-roundness at the vat, the analysis gave bounds at the laser of less than 12% normalized astigmatism and less than 8% waist and divergence asymmetry. These values were easily met by the laser tested and reported in Figure 1.17. In an actual situation of setting laser specifications, several more examples should be run, including cases starting on the laser side and calculating the asymmetries that result in the scanning beam. Readers journeying this far into this applications section should now have sufficient analytical tools provided by the M^2 model to complete those calculations themselves.

1.10 CONCLUSION: OVERVIEW OF THE M^2 MODEL

The M^2 model provides description of real beams by generalizing the equations for the ideal fundamental-mode beam. Any real beam, whether made up of modes from a stable laser resonator or not, is larger in diameter by the factor M—for all propagation distances z—than the embedded gaussian beam implicit within it. Thus the change in equations takes the form of replacing the $1/e^2$ radius w of the embedded gaussian beam by W/M, where W is the radius of the real beam. This replacement generalizes both the beam propagation and beam-lens transform equations.

The real beam, with waist diameter $2W_0$ being larger than the embedded gaussian by the factor M for all z, diverges at an M times larger rate. All diffraction-limited beams have a gaussian irradiance profile, and one of waist diameter $2W_0$, being M times larger than the embedded gaussian diverges at a rate $1/M$ as fast as the embedded gaussian. Hence the real beam divergence is M^2 times larger than the diffraction limit. This identifies M^2 as a beam invariant unchanging in free space propagation or transmission through lenses, and as a measure of the beam quality. An M^2 of unity is the highest quality, a diffraction-limited beam, and real beams with larger values have increasing degrees of higher-order mode content and wavefront aberration (and hence are also called mixed-mode beams).

To apply this analytical description of a mixed-mode beam, its M^2 must first be measured, and here the simplicity of the ideas becomes more complex. The measurement requires finding the scale length for expansion of the beam diameter with propagation, z_R, the Rayleigh range. Several diameters at well-chosen z locations on both sides of the waist are determined and this data is fit to the correct hyperbolic form. The fit gives three beam constants—the beam quality, the waist diameter, and the waist location—for each independent and orthogonal principal propagation plane.

The first additional complexity is that different definitions give different numerical values for the diameters for the mixed mode and the higher-order modes it contains. Beam diameters are still measured from beam irradiance profiles, but different profiling masks

(pinholes, slits, knife-edges, or centered circular apertures) all give different shaped profiles for higher-order modes and hence different diameters. Care is required to keep track of which measurement "currency" is in use and to not mix different currencies. A standard diameter definition—the second-moment diameter, four times the standard deviation of a pinhole irradiance distribution of the beam—has been adopted by the ISO. However, this diameter is computation intensive and difficult to measure reliably because of sensitivity to noise on the wings of the profile. Therefore, conversion rules have been developed applicable to cylindrically symmetric mixed-mode beams permitting measurements done in one diameter currency to be changed into another. The basis of these rules is our observation that higher-order modes turn on and off in a characteristic sequence as the diameter of the circular limiting aperture in the resonator is opened. This associates with the increasing second-moment M^2 a unique set of mode fractions, allowing accurate conversion rules to be derived.

The second additional complexity is that the diameter measurements and curve fits done to determine M^2 may give unreliable answers unless several pitfalls in the process are avoided. Chief among these is that the mixed-mode waist must be accurately located and its diameter physically measured and not inferred or assumed. Since the waists of most lasers are buried inside the resonator, this requires the use of an auxiliary lens to form an auxiliary beam with an accessible waist. The constants determined for this auxiliary beam then are transformed back to those for the original beam by means of the beam-lens transform equations. Commercial instruments that do this automatically are available.

Beams with pure forms of the classic asymmetries have been illustrated, those with only astigmatism, or waist asymmetry, or divergence asymmetry. Beams with combinations of asymmetries may be represented by pairs of propagation plots, one for each principal propagation plane. Beam asymmetries can also be interpreted as deviations from a theoretical "best weighted average" round beam, the equivalent cylindrical beam. There are also beams not directly covered in the M^2 model whose principal propagation planes twist in space like a twisted ribbon—beams with "general astigmatism."

Lastly, the M^2 model was demonstrated by analyzing an actual laser-beam scanning system used in stereolithography. Asymmetries causing out-of-round spots on the scanned surface were analytically projected back through the delivery system to determine the size of the corresponding asymmetries at the laser source.

There are many applications of the M^2 model. It quantifies the mode specifications for commercial lasers, with the means to test to these specifications. Its use permits design of multimode lasers and their beam delivery systems. The beam transformations occurring in nonlinear optics can be analyzed. The divergence of a high M^2 laser beam can be matched into the acceptance angle a high numerical aperture fiber, to take advantage of the lower cost per unit of output power of a multimode laser. These are just a few of many other applications, all with the backup of commercial instrumentation to make the beam measurement process easy and efficient. This chapter has provided the analytical tools to make these applications realities.

discussions: Gerald F. Marshall, the original editor of this book, who always has another intriguing question; and G. Nemes of Astigmat, who taught us about beams with general astigmatism. In this revised edition, Jeff Guttman of Photon, Inc. updated us on recent developments in cameras and profilers. Lastly, David Bacher and John O'Shaughnessy of CVI Melles Griot, Inc., and especially Gerald F. Marshall contributed very helpful and constructive reviews of the manuscript.

GLOSSARY

Astigmatism, general: The property of beams having elliptical cross sections for all z, with the principal axes of the ellipses rotating with propagation along the beam axis (nonorthogonal beams; "twisted" beams).

Astigmatism, normalized: The difference in waist locations for the two independent principal propagation planes divided by the Rayleigh range of the equivalent cylindrical beam, $A_s/z_{Rr} = (z_{0y} - z_{0x})/z_{Rr}$, usually expressed in percent.

Astigmatism, simple: Having different waist locations in the two principal propagation planes, $z_{0x} \neq z_{0y}$.

Asymmetric divergence: Having different divergence angles $\Theta_x \neq \Theta_y$ in the two principal propagation planes.

Asymmetric waists: Having different waist diameters in the two principal propagation planes, $2W_{0x} \neq 2W_{0y}$.

Beam caustic surface: The envelope of the beam swept out by rotating the curve of the beam radius $W(z)$ versus propagation distance about the propagation axis z. When a plane containing the z-axis and at an angle α to the x-axis cuts the caustic surface the intersection gives the propagation plot for azimuth α. See the discussion of Figure 1.16a.

Beam, equivalent cylindrical: A cylindrically symmetric beam constructed mathematically in the M^2 model from the beam constants measured in the two principal propagation planes of an asymmetric beam, see the explanation of Figure 1.16a. The propagation plot for the equivalent cylindrical beam is obtained from the beams of Figure 1.15 by slicing their caustic surfaces along the z-axis at a 45° inclination to the x- or y-axes. This is the best cylindrically symmetric average beam for a beam with asymmetry. The subscript r is used to denote the constants for this beam, for round or radial symmetry.

Beam, gaussian: A uniphase beam with spherical wavefronts whose transverse irradiance profiles everywhere have the form of a gaussian function. Such an idealized beam is diffraction limited with $M^2 = 1$, a condition that can only be approached by real beams.

Beam, idealized: The abstract mathematical description of a beam (which can have $M^2 = 1$).

Beam propagation analyzer: An instrument that measures beam diameters as a function of propagation distance, displays the $2W(z)$ versus z propagation plot, and curve fits this data to a hyperbola to determine beam quality M^2, waist location z_0, and waist diameter $2W_0$.

Beam propagation constant M^2: So called because replacing the fundamental-mode radius $w(z)$ in its propagation equation by $w(z) = W(z)/M$ predicts the propagation of the mixed mode, of radius $W(z)$.

Beam quality: The quantity M^2, so called because a real beam has M^2 times the divergence of a diffraction-limited beam of the same waist diameter; see also "normalizing gaussian."

Beam, real: An actual beam; all have at least slight imperfections and thus an M^2 greater than one.

Clip width: Distance (on the mask translation axis) between the points on an irradiance profile at a specified fraction (such as 13.5%) of the height of the highest peak.

Conversions, beam diameter: Empirical rules derived for beams of cylindrical symmetry, to convert diameters measured by one method to those measured by another, such as slit diameters to knife-edge diameters.

Convolution error: Contribution to the measured diameter from the finite size of the scanning aperture; minimizing this is an important consideration for pinhole and slit measurements.

Cut: A beam diameter measurement, from the cutting action of a profiler's scanning aperture.

Diameter, $1/e^2$: Beam diameter defined by the aperture translation distance between clip points on an irradiance profile at a height of $13.5\% = 1/e^2$ relative to the highest peak at 100%.

Diffractive overlay: Interference from high angle rays overlapping the beam, diffracted from the limiting aperture in the resonator. This can distort profiles taken close to the laser output coupler (within a Rayleigh range).

Eigenfunctions: A set of functions f_n associated with a linear operator Q satisfying $Qf_n = c_n f_n$, where the c_n are scalar constants (the eigenvalues). Because of this self-replicating property these functions occur in many physical problems, for example, the laser mode functions that also describe the harmonic oscillator and hydrogen wave functions in quantum mechanics.

Embedded gaussian: The fundamental mode of the resonator that generates a mixed-mode beam. The mixed-mode beam diameter is M times larger than the embedded gaussian beam diameter at all propagation distances z.

Far-field: The propagation region(s) of a beam many Rayleigh ranges away from the waist locations. In the far-field, the transverse extent of the beam grows linearly with increasing distance from the waist.

Four-cuts method: The simplest method for determining M^2, requiring only four well-chosen diameter measurements both straddling the waist location and at the waist location.

Fresnel number: The square of the radius of the limiting aperture in a resonator, divided by the mirror separation and the wavelength. As the aperture is opened and this number increases, modes of higher order oscillate and join the mix of modes.

Gaussian: A mathematical function of the form $\exp(-x^2)$; see also "beam, gaussian."

Hermite–Gaussian function: An eigenfunction of the wave equation including diffraction, which describes beams of rectangular symmetry, of the form of a gaussian function times a pair of Hermite polynomials of orders (m, n).

Invariant, beam: A quantity that is unchanged by propagation in free space or transmission through ordinary, nonaberrating, optical elements (lenses, Brewster windows, etc.).

Irradiance: The power per unit cross-sectional area of the beam.

Laguerre–Gaussian function: An eigenfunction of the wave equation including diffraction, that describes beams of cylindrical symmetry of the form of a gaussian function times a generalized Laguerre polynomial of order (p, l).

M²: The ratio of the waist diameter-divergence angle product of the mixed-mode beam, to that for the embedded gaussian of that beam. A beam invariant, this is also called the "times-diffraction-limit" number, the beam quality, and the beam propagation factor.

Mode: The characteristic frequencies and transverse irradiance patterns of beams formed in laser oscillators, described by Hermite–Gaussian and Laguerre–Gaussian functions, denoted by the symbols $TEM_{m,n}$, $TEM_{p,l}$ with m, n or p, l the order numbers of the function's polynomials.

Mode, degenerate: Two modes with the same optical frequency, and therefore, order numbers.

Mode, donut: A starred mode, TEM_{01}^*, with the second-lowest diffraction loss through a circular limiting aperture, and an irradiance profile with a hole (null) in the center (see Figure 1.1).

Mode, fundamental: The TEM_{00} mode, with a gaussian irradiance distribution, a single-spot peaked profile, the lowest mode order and smallest beam diameter from a given resonator, and with $M^2 = 1$ in the limit of perfection. Thus this mode has the lowest diffraction loss through a centered circular limiting aperture.

Mode, higher order: Any mode of order number greater than that of the fundamental mode.

Mode, longitudinal: A mode of frequency $q(c/2L)$, where c is the speed of light and q is a large integer equal to the number of beam wavelengths that fit in the round trip path $2L$ of the resonator. The $(q + 1)^{th}$ longitudinal mode has a frequency $(c/2L)$ higher than the qth; each longitudinal mode is associated with a given transverse mode.

Mode, lowest order: The fundamental mode, of order number one.

Mode, mixed: An incoherent superposition of pure modes, all from the same resonator, with a diameter $2W$ that is M times larger for all z than $2w$, the fundamental-mode diameter from the set. Also called a real beam as only idealized beams have $M^2 = 1$ (indicating zero higher-order mode content).

Mode order number: For Hermite–Gaussian modes, $(m + n + 1)$; for Laguerre–Gaussian modes, $(2p + l +1)$; the order numbers determine the mode frequencies and phase shifts, and give the mode's beam quality M_{4s}^2 measured in second-moment units.

Mode or spot pattern: The two-dimensional pattern of the irradiance distribution as would be viewed on a flat surface inserted normally in the beam.

Mode, pure: Any transverse mode that is *not* a mixture of modes of different orders.

Mode, starred: A circularly symmetric mode that is a composite of two degenerate modes combined in space and phase quadrature, that is, superposed with a copy of itself after a 90° rotation (see Figure 1.1).

Mode, transverse: A mode, designated by the symbols $TEM_{m,n}$, $TEM_{p,l}$, whose transverse irradiance distribution is described by the Hermite–Gaussian or Laguerre–Gaussian functions of m, n or p, l order numbers.

Near-field: The beam propagation region(s) within a Rayleigh range from the waist location.

Noise-clip option: A test of the sensitivity to noise of the second-moment diameter computed from a pinhole profile, consisting of discarding any profile data with negative values after subtraction of the background to see the change this makes in the computed diameter.

Normalizing Gaussian: A diffraction-limited, idealized gaussian beam of the same waist diameter as a mixed-mode real beam, whose divergence is used as the denominator in a ratio with the real beam's divergence to compute the real beam's M^2.

Paraxial: Meaning close to the beam axis, this refers to a ray (or bundle of rays) propagating at an angle small enough with respect to the central axis that this angle and its tangent are essentially equal.

Power-in-the-bucket: Alternate term for D_{86}, the variable-aperture beam diameter definition.

Principal diameters (of an elliptical spot): The diameters along the major and minor axes of the ellipse.

Principal propagation planes, independent: The two perpendicular planes containing the major and minor axes of an elliptical beam spot (*x*- and *y*-axes) and the propagation axis (*z*). In the M^2 model the three propagation constants for each of these two planes are independent.

Profile: The record of transmitted power versus translation distance of a small aperture or other mask scanned across the beam.

Profile, knife-edge: A profile taken with a knife edge mask, yielding a tilted S-shaped curve.

Profile, pinhole: A profile taken with a pinhole aperture and capable of showing all the irradiance highs and lows but requiring careful centering of the beam to the scanned track of the pinhole. Signal-to-noise ratio and convolution error are inversely dependent on the pinhole diameter making the hole diameter an important consideration.

Profiler: An instrument for measuring beam diameters, that scans a mask (pinhole, slit, or knife-edge) through the beam, displays the resulting profile, and (usually) reports the beam diameter on a digital readout as the scan distance—or clip width—between preset clip points on the profile.

Profile, slit: A profile taken with a slit aperture, showing something of the irradiance highs and lows and *not* requiring centering of the beam to the scanned track. Signal-to-noise ratio and convolution error are counterdependent on the slit width, making it an important consideration.

Propagation constants: The set of parameters: waist diameter $2W_0$, waist location z_0, and beam quality M^2, in each of the two principal propagation planes that define how the transverse extent of a beam changes as it propagates.

Propagation plot: The plot of beam diameter versus propagation distance, $2W(z)$ versus z. For the beams covered in the M^2 model, the form of this plot is a hyperbola.

Rayleigh range: In the M^2 model, the propagation distance z_R from the waist location to where the wavefront reaches maximum curvature, also the distance from the waist to where the beam diameter has increased by $\sqrt{2}$, and the scale length for beam expansion with propagation, $z_R = \pi W_0^2 / M^2 \lambda$.

Resonator: The aligned set of mirrors providing light feedback in a closed path through the gain medium in a laser. Since the wavefront curvatures and surface curvatures must match at the mirrors, the resonator determines the mode properties of the beam.

Scan: Movement of a mask or aperture transversely across a beam while recording the transmitted power; see "cut."

Second-moment diameter: $D_{4\sigma}$, equal to four times the standard deviation, σ, of the transverse irradiance distribution obtained from a pinhole profile.

Second-moment, linear: The integral over the transverse plane of the square of the linear coordinate times the irradiance distribution, for example, $\langle x^2 \rangle$, used in calculating the variance of the distribution $\sigma^2 = \langle x^2 \rangle - \langle x \rangle^2$.

Second-moment, radial: The integral over the transverse plane of the irradiance distribution times the square of the radial coordinate measured outwardly from the

centroid of the spot, for example $\langle r^2 \rangle$, used in calculating the variance of the distribution $\sigma_r^2 = \langle r^2 \rangle$. In the integration the distribution is weighted by r^3 since the area element is $dA = r\,dr\,d\theta$.

Spot: The two-dimensional irradiance distribution or cross section of a beam as seen on a flat surface normal to the beam axis.

Stigmatic: Describes a beam that maintains a round cross section as it propagates, or more formally, a beam that maintains a rotationally symmetric irradiance distribution in free space. (The opposite term is *astigmatic*, where cross sections are elliptical at some propagation distances z.)

TDL, times-diffraction-limit number: The number of times the divergence of a real beam is larger than that of a diffraction-limited beam (called the normalizing gaussian) of the same waist diameter; TDL $= \Theta/\theta_n = M^2$. Also the factor by which a real-beam waist diameter is larger than that for a gaussian beam ($M^2 = 1$) converging at the same numerical aperture (NA).

TEM_{mn}: (For Transverse ElectroMagnetic wave). A symbol used to designate a transverse mode of rectangular symmetry described by a Hermite–Gaussian function with polynomial orders m, n.

TEM_{pl}: (For Transverse ElectroMagnetic wave). A symbol used to designate a transverse mode of cylindrical symmetry described by a Laguerre–Gaussian function with polynomial orders p, l.

Thresholding: A method for noise reduction in the readout of a CCD camera frame by measuring the noise level in nonilluminated portions of the frame (such as the corners) from which a standard deviation σ is calculated, then subtracting a uniform noise floor level (typically of 3σ amplitude) from the entire frame before processing the signal.

Variable-aperture diameter: D_{86}, (or D_{xx}) the diameter of a centered circular aperture passing 86.5% (or xx%) of the total power in the beam.

Waist, beam: The location on the beam propagation plot where the beam diameter is at a minimum; also used for the value of this minimum diameter.

Waist diameters: $2W_{0x}$, $2W_{0y}$, the minimum diameters in each principal propagation plane.

Waist locations: z_{0x}, z_{0y}, the points along the propagation axis where the minimum (waist) diameter(s) of the beam in each of the independent principal propagation planes are located.

Wave equation: Propagation of paraxial rays including the effect of diffraction are described by either the Fresnel–Kirchhoff diffraction integral equation of Boyd and Gordon[2] or the simple scalar wave equation used by Kogelnik and Li;[1] both have the Hermite–Gaussian and Laguerre–Gaussian functions as eigenfunction solutions.

REFERENCES

1. Kogelnik, H.; Li, T. Laser beams and resonators. *Applied Optics* 1966, *5*, 1550–1567.
2. Boyd, G.D.; Gordon, J.P. Confocal multimode resonator for millimeter through optical wavelength masers. *Bell System Technical Journal* 1961, *40*, 489–508.
3. Marshall, L. Applications à la mode. *Laser Focus* 1971, *7*(4), 26–29.

4. Bastiaans, M.J. Wigner distribution function and its application to first-order optics. *Journal of the Optical Society of America* 1979, *69,* 1710–1716.

5. Siegman, A.E. *Lasers;* University Science Books: Sausalito, CA, 1986; ISBN 0-935702-11-3.

6. Siegman, A.E. New developments in laser resonators. *Proceedings of SPIE* 1990, *1224,* 2–14.

7. Sasnett, M.W. Propagation of multimode laser beams—The M² factor. In *The Physics and Technology of Laser Resonators;* Hall, D.R., Jackson, P.E., Eds.; Adam Hilger: New York, 1989; Chapter 9, ISBN 0–85274-117–0.

8. Johnston, T.F., Jr.; Fleischer, J.M. Calibration standard for laser beam profilers: method for absolute accuracy measurement with a Fresnel diffraction test pattern. *Applied Optics* 1996, *35,* 1719–1734.

9. The Coherent, Inc., ModeMaster™. The manual for the PC version of this instrument containing much useful information is available upon request or on their website from Coherent Laser Measurement and Control, 27650 SW 95th Avenue, Wilsonville, OR 97070.

10. Johnston, T.F., Jr. M² concept characterizes beam quality. *Laser Focus* 1990, *26*(5), 173–183.

11. Test methods for laser beam widths, divergence angles, and beam propagation ratios, ISO/FDIS 11146:2004 in three parts: -1, Stigmatic and simple astigmatic beams; -2, General astigmatic beams; -3 Intrinsic and geometrical laser beam classification, propagation and details of test methods; available from Deutsches Institut fur Normung, Pforzheim, Germany.

12. Lawrence, G.N. Proposed international standard for laser-beam quality falls short. *Laser Focus World* 1994, *30*(7), 109–114.

13. Sasnett, M. et al. Toward an ISO beam geometry standard. *Laser Focus World* 1994, *30*(9), 53.

14. Johnston, T.F., Jr.; Sasnett, M.W.; Austin, L.W. Measurement of "standard" beam diameters. In *Laser Beam Characterization;* Mejias, P.M., Weber, H., Martinez-Herrero, R., Gonzales-Urena, A., Eds.; SEDO: Madrid, 1993; 111–121.

15. Arnaud, J. A.; Kogelnik, H. Gaussian light beams with general astigmatism. *Applied Optics* 1969, *8,* 1687–1693.

16. Mansuripur, M. Gaussian beam optics. *Optics and Photonics News* 2001, *12*(1), 44–47.

17. Nemes, G.; Siegman, A.E. Measurement of all ten second-order moments of an astigmatic beam by the use of rotating simple astigmatic (anamorphic) optics. *Journal of the Optical Society of America* 1994, *11,* 2257–2264.

18. Serna, J.; Encinas-Sanz, F.; Nemes, G. Complete spatial characterization of a pulsed doughnut-type beam by use of spherical optics and a cylinder lens. *Journal of the Optical Society of America* 2001, *18,* 1726–1733.

19. Silfvast, W.T. *Laser Fundamentals;* Cambridge University Press: New York, 1996; Chapter 10, ISBN 0-521-55617-1.

20. Rigrod, W.W. Isolation of axi-symmetric optical resonator modes. *Applied Physics Letters* 1963, *2,* 51–53.

21. McCumber, D.E. Eigenmodes of a symmetric cylindrical confocal laser resonator and their perturbation by output-coupling apertures. *Bell System Technical Journal* 1965, *44,* 333–363.

22. Koechner, W. *Solid–State Laser Engineering,* 5th Ed.; Springer-Verlag: New York, 1999; Figure 5.10.

23. Wolfram, S. *The Mathematica Book,* 3rd Ed.; Cambridge University Press: Cambridge, UK, 1996; ISBN 0-521-58889-8.

24. Feng, S.; Winful, H.G. Physical origin of the Gouy phase shift. *Optics Letters* 2001, *26,* 485–489.

25. Johnston, T.F., Jr. Beam propagation (M²) measurement made as easy as it gets: The four-cuts method. *Applied Optics* 1998, *37,* 4840–4850.

26. Belanger, P.A. Beam propagation and the ABCD ray matrices. *Optics Letters* 1991, *16,* 196–198.

27. Serna, J.; Nemes, G. Decoupling of coherent Gaussian beams with general astigmatism. *Optics Letters* 1993, *18,* 1774–1776.

28. Hecht, E. *Optics,* 2nd Ed.; Addison–Wesley Publishing Co.: Menlo Park, CA, 1987; ISBN 0-201-11609-X.

29. Kogelnik, H. Imaging of optical modes—Resonators with internal lenses. *Bell System Technical Journal* 1965, *44,* 455–494.

30. Self, S.A. Focusing of spherical Gaussian beams. *Applied Optics* 1983, *22*, 658–661.
31. Herman, R.M.; Wiggins, T.A. Focusing and magnification in Gaussian beams. *Applied Optics* 1986, *25*, 2473–2474.
32. O'Shea, D.C. *Elements of Modern Optical Design;* John Wiley & Sons: New York, 1985; ISBN 0-471-07796-8, 235–237.
33. Wright, D.L.; Fleischer, J.M. Measuring Laser Beam Parameters Using Non-Distorting Attenuation and Multiple Simultaneous Samples. US Patent No. 5,329,350, 1994.
34. McCally, R.L. Measurement of Gaussian beam parameters. *Optics Letters* 1984, *23*, 2227.
35. Sasnett, M.W.; Johnston, T.F., Jr. Apparatus for Measuring the Mode Quality of a Laser Beam. US Patent No. 5,100,231, March 31, 1992.
36. Taylor, J.R. *An Introduction to Error Analysis;* University Science: Mill Valley, CA, 1982; ISBN 0–935702-10–5.
37. Green, L. Automated measurement tool enhances beam consistency. *Laser Focus World* 2001, *37*(3), 165–166.
38. Kogelnik, H.; Ippen, E.P.; Dienes, A.; Shank, C.V. Astigmatically compensated cavities for CW dye lasers. *IEEE Journal of Quantum Electronics* 1972, *3*, 373–379.
39. Johnston, T.F., Jr.; Sasnett, M.W. The effect of pump laser mode quality on the mode quality of the CW dye laser. SPIE Proceedings 1992, *1834*, Optcon Conference, Boston, 1992, Paper #29.
40. Johnston, T.F., Jr.; Sasnett, M.W. Modeling multimode CW laser beams with the beam quality meter. OPTCON, Boston, MA, 5 November 1990, Paper OSM 2.4.
41. Firester, A.H.; Gayeski, T.E.; Heller, M.E. Efficient generation of laser beams with an elliptic cross section. *Applied Optics* 1972, *11*, 1648–1649.
42. Siegman, A.E. Laser beam propagation and beam quality formulas using spatial-frequency and intensity-moment analysis, distributed to the ISO Committee on test methods for laser beam parameters, August 1990, 32.
43. Simpson, N.B.; Dholakia, K.; Allen, L.; Padgett, M.J. Mechanical equivalence of spin and orbital angular momentum of light: and optical spanner. *Optics Letters* 1997, *22*, 52–54.
44. Nemes, G.; Serna, J. Laser beam characterization with use of second order moments: An overview. In *DPSS Lasers: Applications and Issues, OSA TOPS;* Dowley, M. W., Ed.; 1998; 17, 200–207.
45. Piestun, R. Multidimensional synthesis of light fields. *Optics and Photonics News* 2001, *12*(11), 28–32.
46. Kivsharand, Y.S.; Ostrovskaya, E.A. Optical vortices. *Optics and Photonics News* 2002, *13*(4), 24–28.
47. Partanen, J.P.; Jacobs, P.F. Lasers for stereolithography. In *OSA TOPS on Lasers and Optics for Manufacturing;* Tam, A.C., Ed.; Optical Society of America: Washington, DC, 1997; Vol. 9, 9–13.
48. Partanen, J. Lasers for solid imaging. *Optics and Photonics News* 2002, *13*(5), 44–48.
49. Ibbs, K.; Iverson, N.J. Rapid prototyping: New lasers make better parts, faster. *Photonics Spectra* 1997, *31*(6), 4 pages.
50. SLA 250/30 Product Data Sheet from 3D Systems, 26081 Avenue Hall: Valencia, CA 91355.

2

Optical Systems for Laser Scanners

Stephen F. Sagan

NeoOptics
Lexington, Massachusetts, USA

CONTENTS

2.1 INTRODUCTION

This chapter builds on the original work of Robert E. Hopkins and David Stephenson on optical systems for laser scanners[1] to provide yet another perspective. The goal of this chapter is to provide the background knowledge that will help develop an insight and intuition for optical designs in general and scanning systems in particular. Combined with a familiarity with optical design tools, these insights will help lead to optical designs with higher performance and fewer components. Design issues and considerations for holographic scanning systems are discussed in detail.

The interactions between optical requirements and constraints imposed on optical systems for laser scanners are explored. The optical components that many applications of laser scanning depend on to direct and focus the laser beam are discussed, including lenses, mirrors, and prisms.

The optical invariant, first-order issues, and third-order lens design theory as they relate to scanning systems are presented as a foundation to the layout and design of the optical systems. Representative optical systems, with their characteristics, are listed along with drawings showing the lenses and ray trajectories. Some of the optical systems used for scanning that require special methods for testing and quality control are reviewed.

2.2 LASER SCANNER CONFIGURATIONS

Optical system configurations for laser scanners can vary in complexity from a simple collimated laser source and scanner to one including beam conditioning optical components, modulators, cylinders, anamorphic optical relays, laser beam expanders, multiple scanners, and anamorphic optical components for projecting the scanned beam.

The scanned laser beam can be converging, diverging, or collimated. Figure 2.1 illustrates the three basic scanning configurations: *objective*, *post-objective*, and *pre-objective* scanning.[2]

2.2.1 Objective Scanning

The objective scanning configuration, where the objective, laser source, image plane, or a combination of these is moved, is the least common method of optical scanning. Objective
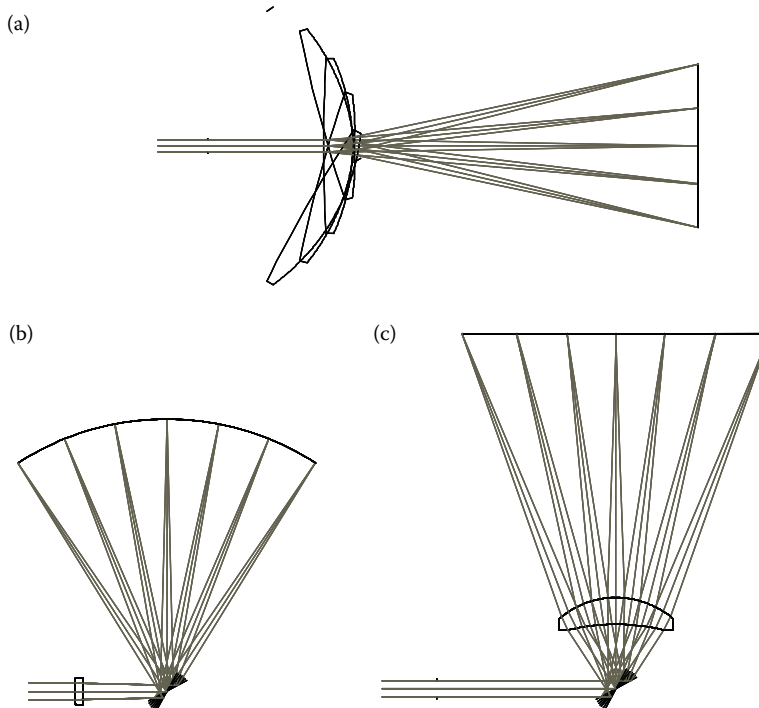
FIGURE 2.1
Three basic scanning configurations. (a) objective scanning; (b) postobjective scanning; (c) preobjective scanning.

scanning is accomplished by rotating about a remote axis as illustrated in Figure 2.1 (or translating in a linear fashion) a focusing objective across the collimated beam. The moving objective can be a reflective mirror, refractive lens, or diffractive element (such as a holographic disc). The fundamentals of holographic scanning will be described beginning in Section 2.11.

2.2.2 Post-objective Scanning

The post-objective scanning configuration requires one of the simplest optical systems because it works on-axis. The rotation axis of the scanner can be orthogonal to the optical axis as with a galvanometer (as illustrated in Figure 2.1) or coaxial, as in the case of a monogon scanner.

For many low-resolution applications (barcode scanners, for instance), simple lenses are sufficient to expand or begin focusing the beam prior to being scanned. As system resolution requirements increase, larger numerical apertures and better optical correction will require additional lens elements and element complexity (such as doublets for spherical and color correction). The disadvantage of post-objective scanning is that the focal plane is curved, requiring an internal drum surface.

From an optical viewpoint, internal drum scanning offers high resolution over large formats with relatively simple optics. The laser beam, lens elements, and monogon scanner can be mounted coaxially into a carriage, with the scanner rotated about the optical axis of the incident laser beam. In such a system, the scanned spot would trace a complete circle on the inside of a cylinder. Translating either the carriage (scanning optical subsystem) or the drum will generate a complete two-dimensional raster. This type of system is ideal for inspecting the inside surface of a tube or writing documents inserted on the inside of a drum.

2.2.3 Pre-objective Scanning

In a pre-objective scanning configuration the beam is first scanned into an angular field and then usually imaged onto a flat surface. The entrance pupil of the scan lens is located at or near the scanning element. The clearance from the scanner to the scan lens is dependent on the entrance pupil diameter, the input beam geometry, and the angle of the scanned field. The complexity of the scan lens is dependent on the optical correction required over a finite scanned field, that is, spot size, scan linearity, astigmatism, and depth of focus (DOF).

Pre-objective scanners are the most commonly seen systems; these systems often require multielement flat-field lenses. The special conditions described in the next few sections must be considered during the design of these lenses.

2.3 OPTICAL DESIGN AND OPTIMIZATION: OVERVIEW

Computers and software packages available to the optical designer for the layout, design/ optimization, and analysis of optical systems (including developments in global optimization and synthesis algorithms) can be very powerful tools. Despite these advances, the most important tools available to the optical designer are simply a calculator, pen and

paper, and a keen understanding of the first- and third-order fundamentals. These fundamentals provide key tools for back-of-the-envelope assessment of the issues and limitations in the preliminary phases of an optical design.

A successful design begins with an appropriate starting point including: (1) a list of the system specifications to scope the design problem, assess its feasibility, and guide the design process (see example list given in Table 2.1); (2) a first-order layout of the system configuration—the position of optical component groups, the aperture stop, and intermediate pupils and images; and (3) the selection of candidate design forms for the design of the optical component groups. Parameters that are entirely dependent on other specifications (in other words redundant) can be listed as reference parameters to provide further clarity.

The important fundamentals in the design of an optical system are as follows:

1. First-order parameters, particularly the optical invariant
2. First-order diffraction theory
3. Third-order aberrations

and then the rest.

Understanding the fundamentals can often mean the difference between achieving a simple "relaxed" design (with fewer optical components and a reduced sensitivity to fabrication and alignment errors), and a complicated "stressed" design, which meets the nominal performance goals but is difficult to assemble to meet as-built specifications.

A relaxed design will have low net third-order aberrations with reduced and distributed individual surface contributions to minimize induced higher order aberrations that can affect the performance of the as-built system. Lens elements in a relaxed design will generally bend with the marginal or chief ray, based on the intermediate speed and field of view demands of the system. Figure 2.2 shows a microscope objective where most air/glass surfaces are bent to minimize marginal ray angles of incidence and therefore minimize

TABLE 2.1

Example List of Optical Specification for a Scanning System

Parameter	Specification or Goal
1. Image format (line length)	216 mm
2. Wavelength	770–795 nm
3. Nominal $1/e^2$ spot size	26 µm diameter, $\pm$ 10% (~1000 DPI)
4. Spot size variation	<4% (over image field)
5. RMS wavefront error	<1/30 wave (or Strehl for optimization)
6. Scan linearity (F–Θ distortion)	<1% (<0.2% over $\pm$ 25°)
7. Scanned field angle	$\pm$30°
8. Effective focal length	206 mm (reference)
9. F-number	F/26 (reference)
10. Depth of focus	>1 mm (reference)
11. Overall length	335 mm
12. Scanner clearance	25 mm
13. Image clearance	270 mm
14. Optical throughput	>50% (including source truncation)

DPI, dots per inch.

(a)   (b)

Microscope objective, plan-achromat   Fisheye lens, U.S. Pat. 4,412,726
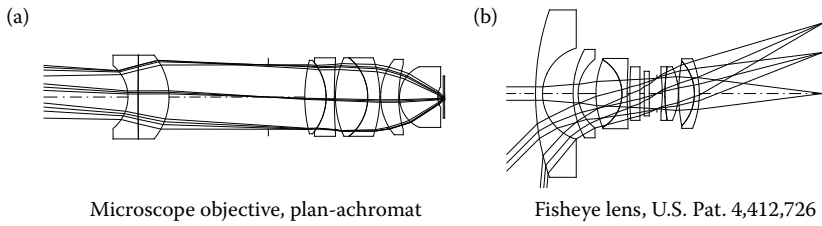
FIGURE 2.2
Example lens designs configured for (a) aperture (microscope objective at left) and (b) primarily field (wide-field fisheye objective at right).

individual surface aperture-dependent aberrations and a wide field of view fisheye objective where most surfaces are bent to minimize chief ray angles of incidence, minimizing individual surface field dependent aberrations. The ability to recognize which design forms work better over the field and which work better over the aperture will help in developing relaxed design forms.

Spherical surfaces are naturally easier to fabricate and test. However, aspheric surfaces (as a design variable) can be used to gain insight into what is holding back a design, or help find a new design form. Their moderate use can save weight and space, or they can often be replaced later in the design process with additional spherical elements. Aspheric surfaces can also be over used, with surfaces competing for correction during the optimization process, leading to overly complex and tolerance-sensitive design solutions.

Design variables such as surface curvatures, the airspace or glass thickness between surfaces, and glass types, and optimization constraints appropriate to the design should be used. Too many variables and/or constraints, particularly conflicting ones, will limit the optimization convergence and performance of the design. Glass type and element thickness are often weak design variables. When glass variables are important, parameters such as cost, availability, production schedule, weight, and transmission, in addition to baseline performance must be considered in their final selection. Changing the glass map boundaries during the optimization process (allowing a wider range early in the design) can lend insight into possible alternative solutions. Vendor glass maps and catalogs are useful reference tools during the selection of glass types.

Anamorphic optical systems using combinations of cylindrical, toroidal, and anamorphic surfaces (with different radii in X and Y directions) can add more degrees of freedom and lens complexity, but are substantially more difficult to fabricate and test.

2.4 OPTICAL INVARIANTS

The optical invariant is defined at any arbitrary plane in a medium with refractive index n as a function of the paraxial marginal ray height and angle (y_m and nu_m) and paraxial chief ray height and angle (y_c and nu_c), as illustrated in Figure 2.3 and given by the relationship

$$I = (y_m nu_c - y_c nu_m). \tag{2.1}$$
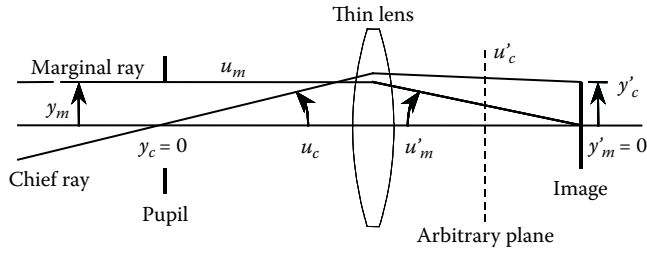
FIGURE 2.3
Paraxial marginal and chief rays for a simple lens.

The optical invariant, as the name implies, is a constant throughout the optical system, provided it is not modified by discontinuities in the optical system such as diffusers, gratings, or other discontinuities such as vignetting apertures. The optical invariant is typically calculated at the object, aperture stop, or final image of the system, conveniently defined by the product of the object (or image) height times marginal ray angle or pupil height times chief ray angle. At the aperture stop or a pupil plane the chief ray height y_c is equal to zero, and the optical invariant reduces to

$$I = y_m n u_c \qquad (2.2)$$

where the chief ray angle term (nu_c) is the paraxial half-field or scan angle. At the object or an image plane the marginal ray height y_m is equal to zero, and the optical invariant reduces to

$$I = -y_m n u_m \qquad (2.3)$$

where the marginal ray angle term (nu_m) is the paraxial equivalent of the sine of the cone half angle in air of the light focused on the image plane, known as the *numerical aperture* (NA). These reduced invariant equations are very useful when dealing with the optical properties at intermediate images or pupil conjugates within the system.

The *f*-number of a lens, defined as the lens focal length F divided by the design entrance aperture diameter (D_L), is also used to describe the image cone angle, with the relationship between NA and *f*-number (*F*/#) for infinite conjugates given by

$$F/\# = \frac{F}{D_L} = \frac{1}{2NA}. \qquad (2.4a)$$

This relationship is clear for a collimated object, but at finite conjugates the lens *f*-number no longer describes the operating *f*-number, which is simply defined by the relative aperture

$$F/\# = \frac{1}{2NA}. \qquad (2.4b)$$

Most scan lenses operate in collimated space and it is convenient to use the *F*/# to describe the image-side cone angle. It is this relative aperture definition that will be used throughout this section.

2.4.1 The Diffraction Limit

Most scanning systems are required to perform at or very near the diffraction limit. The fundamental limit of performance for an imaging system of focal length F, illuminated by a uniform plane wave of wavelength λ and truncated by an aperture of diameter D is defined by the Airy disk first diffraction ring diameter

$$d = \frac{2.44\,\lambda}{D/f} = 2.44\,\lambda(F/\#).$$

(2.5)

This diffraction limit is an optical invariant that determines the resolution in both the spatial and angular domain, and can be thought of as a *spot-invariant* (or spot-divergence product) that can be rewritten as

$$d(2NA) = 2.44\,\lambda.$$

(2.6)

The fundamental diffraction limit for an ideal Gaussian beam with no truncation is defined by the *waist-invariant* (or waist-divergence product)

$$w_0\theta_{1/2} = \frac{\lambda}{\pi}$$

(2.7)

where w_0 is the radius of the beam waist and $\theta_{1/2}$ is the half divergence angle in the far-field (where z is much greater than the Raleigh range $w_0{}^2/\lambda$) at the $1/e^2$ level for an ideal Gaussian beam of wavelength λ. Defined in terms of the $1/e^2$ waist diameter and full divergence angle, the waist-invariant becomes

$$d_0\theta = \frac{4\lambda}{\pi} = 1.27\,\lambda.$$

(2.8)

2.4.2 Real Gaussian Beams

Laser scanning systems typically use a near-Gaussian input beam. The degree to which the beam is Gaussian (TEM00) depends on the type of laser and the quality of the beam. Siegman[3] has shown that real laser beams (irregular or multimode) can be described analytically by simply knowing the near-field beamwidth radius W_0 and far-field half divergence angle $\Theta_{1/2}$, defined for each as the standard deviations measured in two orthogonal planes coincident with the axis of propagation. The product of these parameters defines the real beam waist-invariant and is proportional to the Gaussian beam diffraction limit given by

$$W_0\Theta_{1/2} = \left(\frac{\lambda}{\pi}\right) M^2$$

(2.9)

where the factor M^2 defines the "times-diffraction-limit." When comparing beams of equal waist or divergence, the real beam divergence or waist, respectively, will be greater than the diffraction limit by a factor M^2, as illustrated in Figure 2.4. The waist-invariant of a real beam will always be greater than the Gaussian diffraction limit.

Engineers developing scanning systems often use the concept of spot diameter. The specifications will call for a spot diameter measured at a specified intensity level, typically the $1/e^2$ and the 50% intensity levels. The maximum allowable growth of spot size across the length of the scan line is also included in the specification. Measurements by
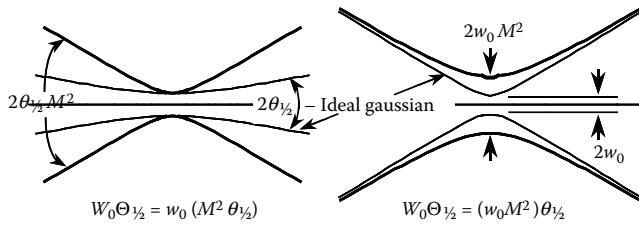
FIGURE 2.4
Relationship between ideal and real Gaussian beams.

commercially available instruments that measure spot profile with a scanning slit will differ from the calculated point-spread function of the point image because the spot profile is determined by integrating the irradiance as the slit passes over the point image. This line-spread function measurement of the Airy disc does not have zeros in the irradiance distribution and is a more appropriate measure of integrated exposure when the spot is constantly moving during the exposure.

2.4.3 Truncation Ratio

Laser scanning systems typically use a near-Gaussian input beam with some truncation. *Truncation* means that a hard aperture restricts the diameter extent of the Gaussian beam, usually located in the input collimator. The truncation ratio (W) is the ratio of the diameter of the Gaussian beam D_B (usually defined at the $1/e^2$ irradiance level), to the diameter of the truncating aperture D_L, defined as

$$W = \frac{D_B}{D_L}. \tag{2.10}$$

Figure 2.5 shows how the image of a diffraction-limited beam is affected by different truncation ratios.

It is important to remember that a scan lens does not have a fixed aperture stop and that the lens diameters are actually much larger than the design aperture to pass the oblique ray bundles of the scanned beam. The prescan collimated beam, often called the feed beam, usually determines the aperture. The diameter of the beam should be no larger than the diameter of the largest beam for which the lens can provide the required image quality, which is usually diffraction-limited. This is called the *design aperture* and is the value to use for D_L, when calculating truncation ratio W, but does not refer to the actual physical diameter of the scan lens.

Extending the definition of the diffraction limit to include the effect of the truncation ratio leads to the definition

$$d_x = \frac{k_x 1}{2NA} = k_x 1(F/\#). \tag{2.11}$$

The value of k_x depends on the truncation ratio W and the level of irradiance in the image spot used to measure the diameter of the image. Figure 2.6 shows how the value of k_x and consequently the diameter of the image of a point source is affected by different amounts of the truncation ratio W. Figure 2.6 also shows two criteria for the image diameter d, one
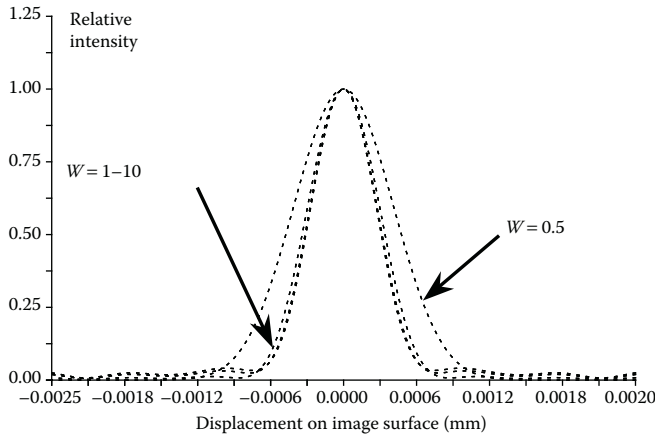
FIGURE 2.5
Point spread for a perfect wavefront and various truncations.
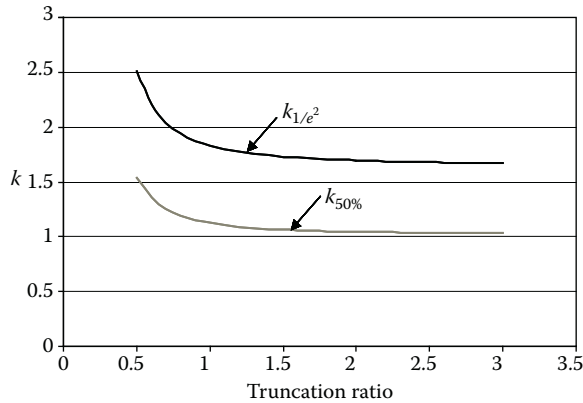
FIGURE 2.6
Effect of truncation ratio on relative spot diameter.

for the $1/e^2$ irradiance level and another for the 50% irradiance level. Equations for these two cases may be found in Reference 4.

$$k_{\text{FWHM}} = \frac{1.021 + 0.7125}{(W - 0.2161)^{2.179}} - \frac{0.6445}{(W - 0.2161)^{2.221}} \tag{2.12}$$

$$k_{1/e^2} = \frac{1.6449 + 0.6460}{(W - 0.2816)^{1.821}} - \frac{0.5320}{(W - 0.2816)^{1.891}} \tag{2.13}$$

A truncation ratio of 1 generally provides a reasonable trade-off between spot diameter and conservation of total energy (86.5%). With $W = 1$, the following equations can be used to estimate a spot diameter:

$$d_{1/e^2} = 1.891(F/\#) \tag{2.14}$$

and

$$d_{50} = 1.13 \lambda (F/\#).$$

(2.15)

There are several points to consider when deciding what truncation ratio to use. It would appear that the Gaussian beam spot with a 1.83λ diameter dependence is smaller than the uniform beam Airy disc with a 2.44λ diameter dependence. However, the Gaussian beam diameter formula refers to the 13.5% irradiance level in the image, while the Airy disc formula relates to the diameter of the first zero in irradiance. Figure 2.5 illustrates this with the irradiance distributions of the near-uniform illumination of the $W = 10$ curve approaching that of an Airy disc pattern that is narrower than the truncated Gaussian illumination beam of the $W = 1$. On the other hand, the Airy disc image has more energy out in the wings of the image than does the Gaussian beam.

It is clear that the heavier truncation ratios ($W \gg 1$, e.g. where a given fixed aperture is overfilled by a Gaussian beam to a nearly uniform illumination) yield smaller spot sizes, but they also suffer the flare or side lobes from the diffraction rings formed by the truncated beam. For this reason, many designers believe that lower values of W (in the 0.5–1.0 range) are a better compromise, providing more light with less danger of image flare.

2.5 PERFORMANCE ISSUES

This section describes the terminology and unique image requirements of laser scan-lens design that are not typical factors in the design for most photographic objectives.

2.5.1 Image Irradiance

There are subtle differences to be considered in the calculation of image irradiance produced by a scan lens, compared to that of a normal camera lens. In galvanometer and polygon laser beam scanners, while the design aperture stop of scan lenses should be located on or near the deflecting mirror surface, these turning mirrors do not alter the circular diameter of the incoming beam as the deflection changes. This is different from a camera lens, which has a fixed aperture stop perpendicular to the lens optical axis. The oblique beam in a camera lens is foreshortened by the cosine of the angle of obliquity on the aperture stop. Designing the scan lens with a slightly larger entrance pupil (by the inverse cosine of the field angle) will provide a good first-order solution.

Most lens design programs do not automatically take this aperture effect into account, so at some point in the design process it will be necessary to use the proper tilts in the design program to maintain the beam diameter at each field angle to be optimized. This can be done in the multiconfiguration (or zoomed) setup available in most of the commercial design programs. The design program then optimizes several versions of the design simultaneously. Section 2.8.4 discusses in greater detail how multiconfiguration design procedures can be used in scan-lens design.

2.5.2 Image Quality

Addressability is an important term widely used in laser scanning. It refers to the least resolvable separation between two independent addressable points on a scan line. When

the concept of spot diameter is used to describe optical performance, it is difficult to know how close the two spots can be to recognize them as separate points. Electrical engineers tend to think in terms of Fourier analysis, suggesting the concept of the modulation transfer function (MTF).[21]

The MTF specification can offer advantages in describing the optical performance of laser scanners. Figure 2.7 shows MTF plots for diffraction-limited images formed with truncation ratios W of 10 (near Airy disc), 2, 1, and 0.5. It is clear that the lower values of truncation have higher MTF for the low frequencies. The best value for W is close to 1. At this truncation ratio the MTF is highest, up to 43% of the design aperture theoretical cutoff frequency. This suggests that the principle of design to follow is to use a value of W close to 1 and design to as small an $F/\#$ as possible, consistent with the performance and cost considerations.

The foregoing rule is based on a perfect image. In attempting to increase the MTF at frequencies below 43% of the cutoff frequency, problems with aberration eventually occur in the large design apertures required. Fortunately, the small values of W mean that the intensity of the rays near the edge of the aperture is reduced, so the acceptable tolerance on the wavefront aberrations can be relaxed. It is not as easy to give a rule-of-thumb tolerance on the wavefront errors because it depends on the type of aberration. The higher-order aberrations near the edges of the pupil will have less effect than will lower-order aberrations, such as out-of-focus or astigmatism errors.

Specifying the performance of the optical system of a laser scanner and the appropriate measure, be it point-spread or line-spread function-based spot size or MTF, can be a point of confusion. In terms of writing an image with independent image points, the point-spread function is a convenient concept. The lens point-spread function would be evaluated at several points along the scan, and the written spot sizes determined by the exposure profile defined by the point spread; that is, as the exposure level increases or decreases, the observable spot size determined by the irradiance level in the point-spread function also increases or decreases. The effect of exposure on spot size depends on the type of image being written—analog gray scale or digital half tone—and the response of the medium being written on. Often the choice of intensity level used to define the design spot size is $1/e^2$ because it draws more attention to energy pushed into side lobes
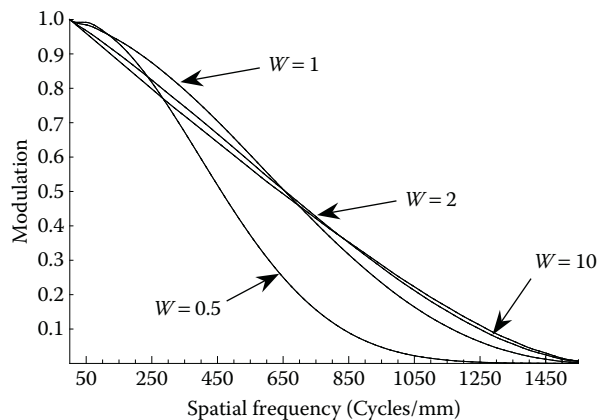
FIGURE 2.7
Modulation transfer function (MTF) curves for a perfect image with truncation ratio W.

of the point spread that can deteriorate the performance of a real lens with aberrations. Sometimes even side lobes above 5% are of consequence to the written image performance. The specification of full-width half maximum (FWHM) most often relates to the final written product.

A real lens with aberrations will spread the energy of an imaged spot beyond the Airy disc. This redistribution of energy reduces the irradiance in the center of the point-spread function. A measure of that redistribution is the Strehl ratio, the ratio of the spot peak intensity relative to the diffraction limit. The Strehl ratio is a convenient measure of the lens image quality during the design process (along with RMS wavefront error) and it is useful in calibrating normalized point-spread function calculations when evaluating a lens over several field angles.

The problem with the above concept is that the laser beam is constantly moving as it writes, smearing the imaged spot along the scan line. As the spot moves, the beam irradiance level is also being modulated to write the information required. This as-used writing process would point to the MTF as an appropriate measure of performance. However, the use of MTF assumes that the recording medium records irradiance level linearly over the complete range of exposures. This may or may not be true, depending on the recording medium. It is important for the lens designer to discuss these differences with the system designer to ensure that all parties understand the issues and trade-offs. It may be easier to correlate a specified spot size variation through focus and across field with the Strehl ratio rather than the MTF.

It usually pays to be conservative and over design by at least 10% on initial ventures into laser scanning system development. The time to be most critical of a new design is in the first prototype and in the testing of the first complete system.

2.5.3 Resolution and Number of Pixels

The total number of pixels along a scan line is a measure of the optical achievement, given by

$$
\begin{aligned}
n &= \frac{L}{d} \\
&= \frac{2\theta F}{k\lambda F/D_L} \\
&= \frac{2\theta D_L}{k\lambda}
\end{aligned}
\tag{2.16}
$$

where n = number of pixels, L = length of scan, d = spot diameter, D_L = diameter of lens design aperture, θ = scan half angle (radians), and F = scan lens focal length. The criterion for spot diameter will largely depend on the media sensitivity and its response to the $1/e^2$ or the 50% irradiance level.

2.5.4 Depth of Focus Considerations

Another important consideration in laser scanning systems is the DOF. The classical DOF for a perfectly spherical wavefront is given by

$$
\text{DOF} = \pm 2\lambda \left(F/\# \right)^2.
\tag{2.17}
$$

This widely used criterion is based on a one-quarter wave departure from a perfect spherical wavefront. A similar criterion defined for a Gaussian beam as the optimum balance between beam size and DOF is given by the Raleigh range

$$Z_R = \frac{p w_0^2}{l} \tag{2.18}$$

where Z_R is the distance along the beam axis on either side of the beam waist at which the wavefront has a minimum radius of curvature of

$$R_{min} = 2Z_R \tag{2.19a}$$

and the transverse $1/e^2$ beam radius is

$$w_R = \sqrt{2w_0}. \tag{2.19b}$$

Each of these generalized criteria (Equations 2.17 and 2.18] serve their particular purpose, but many system specifications state that the spot size diameter must be constant within 10% (or even less) across the entire scan line. Additionally, manufacturers of scanning systems often impose a lower limit to the tolerable DOF. There is no simple formula to relate DOF to this requirement, but spot size or MTF calculations made for several focal plane positions can provide the pertinent data. Figure 2.8 shows the MTF curves of an *F*/5 parabola under the following conditions:

1. The perfect image with uniform 632.8 nm wavelength irradiance across the entire design aperture (W = 1000).

2. The perfect image with $W = 0.85$.

3. The same image as A, but with a focal shift of 0.063 mm. This corresponds to a wavefront error of half of a wave at the maximum design aperture.

4. The same image and truncation as B, but with a focal shift of 0.063 mm.
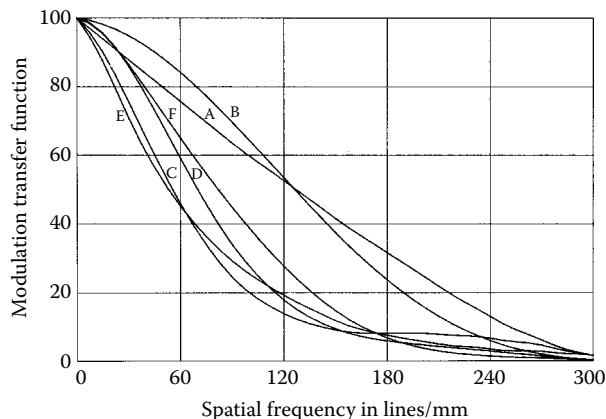
FIGURE 2.8
Effect of focus shift, spherical aberration, and truncation ratio on modulation transfer function (MTF). (From Hopkins, R.E.; Stephenson, D. Optical systems for laser scanners. In *Optical Scanning*; Marshall, G.F., Ed.; Marcel Dekker: New York, 1991; 27–81.)

5. The image from a parabola with aspheric deformation added to introduce a half-wave of fourth-order wavefront error at the edge of the design aperture; $W = 1000$, no focus shift.

6. The same as E with $W = 0.85$, no focus shift.

The truncation value of $W = 0.85$ was used for this example instead of 1.0 in order to help reduce the influence of aberrated rays near the edge of the design aperture. These curves show that the DOF is slightly improved by truncating at this value. They also show that one half-wave of spherical aberration does not have as serious an effect on the DOF as does an equivalent amount of focus error. Therefore it is most important to reduce the Petzval curvature and astigmatism in a scan lens, because these aberrations cause focal shift errors.

2.5.5 The F–Θ Condition

In order to maintain uniform exposure on the material being scanned, the constant power image spot must move at a constant velocity. As the scanner rotates through an angle $\theta/2$, the reflected beam is deflected through an angle θ, where the angle θ is measured from the optical axis of the scan lens. Because polygon scanners rotate at a constant velocity, the reflected beam will rotate at a constant angular velocity. The scanning spot will move along the scan line at a constant velocity if the displacement of the spot is linearly proportional to the angle θ. The displacement H of the spot from the optical axis should follow the equation

$$H = F\theta \tag{2.20}$$

where the constant F is the approximate focal length of the scan lens. Figure 2.9 is the distortion in an F–θ lens relative to that of a normal lens corrected for linear distortion (F-tan θ) plotted over scan angle. The curve's departure from the straight line represents the distortion required of an F–θ lens for a constant scan velocity. As the field angle increases, a classical distortion-free lens image points too far out on the scan line, causing the spot to move too fast near the end of the scan line. Fortunately, typical scan lenses begin with negative (barrel) third-order distortion—the image height curve laying below the F–tan θ curve. The distortion can be designed to match the F–θ image height at the edge of the field
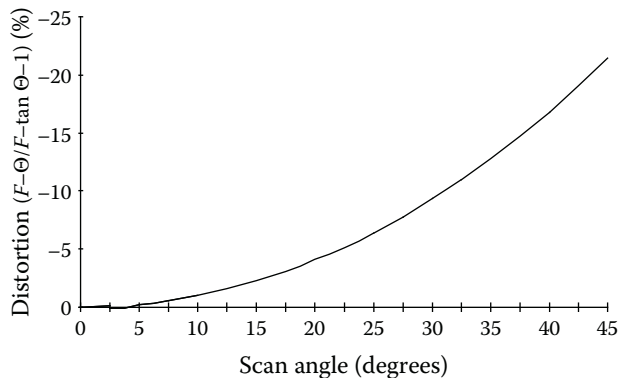
FIGURE 2.9
Error between F–tan θ and F–θ distortion correction.
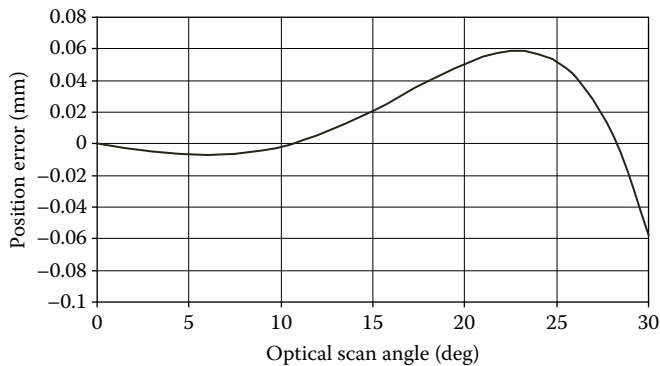
FIGURE 2.10
$F–\theta$ linearity error minimized with calibrated focal length, third-order and fifth-order distortion.

or balanced over the image. For a given distortion profile, the plus and minus departures from the ideal $F–\theta$ height over the scanned image can be balanced for minimum plus and minus departures by scaling the value of F used in defining the data rate. Scaling the data rate effectively scales the pixel spatial frequency written at the image plane. The focal length that minimizes the departures from linearity is called the *calibrated focal length*.

When the field angle is as large as $\pi/6$ radians (30°), the residual departures from linearity may still be too large for many applications. Balancing negative third-order distortion against positive fifth-order distortion can further reduce the departures. Lens designers will recognize this technique as similar to the method of reducing zonal spherical aberration by using strongly collective and dispersive surfaces, properly spaced. In this case the zonal spherical aberration of the chief ray must be reduced. Figure 2.10 shows an example of this correction.

This high-order correction should not be carried too far, since the velocity of the spot begins to change rapidly near the end of the scan and may result in unacceptable changes in exposure or pixel placement. It may also begin to distort the spot profile, turning a circular spot into an elliptical one. This *local distortion* results in a change in the resolution or spatial frequencies near the end of the scan line. A standard observer can resolve frequencies of 10 line pairs/mm (254 lines/in), but is even more sensitive to variations of frequency in a repetitive pattern. Variations of frequency as small as 10% may be detected by critical viewing. The linearity specification is often expressed as a *percent error* (the spot position error divided by the required image height). For example, the specification often reads that the $F–\theta$ error must be less than 0.1%. This means that the deviations must be smaller and smaller near the center of the scan line. It is not reasonable to specify such a small error for points near the center of the scan. The proper specification should state rate of change of the scan velocity and the allowable deviation from the ideal of Figure 2.9. More detail on this subject may be found in Reference 4.

2.6 FIRST- AND THIRD-ORDER CONSIDERATIONS

The optical system in a scanner should have a well-considered first-order layout. This means that the focal lengths and positioning of the lenses should be determined before

any aberration correction is attempted. Most of the optical systems to be discussed in this section will first be described as groups of thin lenses. The convention used for thin lenses is described in most elementary books on optics.[6, 7]

The graphical method shown in Figure 2.11 is useful for a discussion of determining individual and total system focal lengths. The diagram shows an axial ray that is parallel to the optical axis. This represents a collimated beam entering the lens. The negative lens "a" refracts the axial beam upward to the positive lens "b." The positive lens "b" then refracts the ray to the axis at the focal plane at F_{2ab}, which is the writing plane for the laser beam.

The second focal point of the negative lens is at F_{2a}. This point is located by extending the refracted axial ray backward from the negative lens until it meets the optical axis. The second focal point of the positive lens (F_{2b}) may be determined by drawing a construction line through the center of the positive lens parallel to the axial ray as it passes between the positive and negative lenses. Because the two lines are parallel, they must come to focus in the focal plane of the positive lens. The focal lengths of the two lenses are now determined. The front and back focal lengths of each lens are equal because the lenses are in air. The focal points F_{1a}, F_{2a} and F_{1b}, F_{2b} are now located. The diagram also shows the construction for finding the second principal point P_2. The distance P_2 to F_{2ab} is the focal length F of the negative–positive lens combination.

The chief ray is next traced through the two-element system. This is done using the concept that two rays that are parallel on one side of a lens must diverge or converge to the second focal plane of the lens. The chief ray enters the lens system after it passes through the entrance pupil (or aperture stop) of the system. For scan lenses, the entrance pupil is usually located at the scanning element. Note that the entrance pupil is located in front of the lens, which is in contrast to a photographic lens where the entrance pupil is usually virtual (located on the image side of the front lens) and the aperture stop is usually located between the lens elements. This is the primary reason why a photographic lens should not be used as a scan lens. It is also one of the reasons why scan lenses are limited in the field angles they can cover.

The completed diagram labels the lens focal lengths. The system focal length is 80.79, $F_a = -55.42$, and $F_b = 48.63$. The *Petzval curvature* is given by the sum of the power of each lens element divided by its index of refraction as

$$P = \sum_i \frac{\Phi_i}{n_i} = \sum_i \frac{1}{F_i n_i}. \tag{2.21}$$
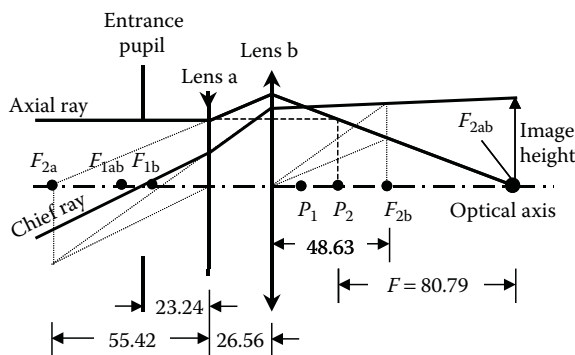
FIGURE 2.11
Graphical solutions to a system of thin lenses.

The Petzval radius ($1/P$) is 3.3 times the focal length and it is curved towards the lens. This is not flat enough for an $F/20$ system when the lens has to cover a long scan line. Equation 2.22 described in Section 2.6.4 provides a formula for estimating the required Petzval radius for a given system. When the Petzval radius is too short, the field has to be flattened by introducing positive astigmatism, which will cause an elliptically shaped writing spot. The Petzval radius is a fundamental consideration in laser scan lenses and becomes a major factor that must be reckoned with in systems requiring small spot sizes. Small spots require a large NA or a small $F/\#$. Observations to be made from this layout include:

1. The distance from the entrance pupil to the lens is 23.24% or 29% of the focal length of the scan lens.
2. If the entrance pupil is moved out toward F_{1ab}, the chief ray will emerge parallel to the optical axis and the system will be telecentric. This condition has several advantages, but lens "b" must be larger than the scan length and the large amount of refraction in lens "b" will introduce negative distortion, making it difficult to also meet the $F-\theta$ condition.
3. Reducing the power of the negative lens or decreasing the spacing between the lenses will allow for a longer distance between the lens and the entrance pupil, but this will introduce more inward-curving Petzval curvature.

This brief discussion illustrates some of the considerations involved in establishing an initial layout of lenses for a scanner. One must decide, on the basis of the required spot diameter and the length of scan, what the Petzval radius has to be in order to achieve a uniform spot size across the scan length. When field flattening is required, it is necessary to introduce more negative power in the system. The most effective way to do this is to insert a negative lens at the first, second, or both focal points of a positive-focal-length scan lens. In these positions they do not detract from the focal length of the positive lens, so the Petzval curvature can be made to be near zero when the negative lens has approximately the same power as the positive lens. The negative lens at the second focal point, however, must have a diameter equal to the scan length, and it will introduce positive distortion if it is displaced from the focal plane. This distortion will make it difficult to meet the $F-\theta$ condition. A negative lens located at the first focal point of the lens is impractical, since there would be no distance between the lens and the position of the scanning element.

The next best thing to do is to place a single negative lens between the positive lens and the image plane. When the negative lens and the positive lens have equal but opposite focal lengths and the spaces between the lenses are half the focal length of the original single lens, then the focal lengths of two lenses are $+0.707F$ and $-0.707F$. The system with the positive lens in front is a telephoto lens, and the one with a negative lens in front is an inverted telephoto. The telephoto lens has a long working distance from the first focal point to the lens, while the inverted telephoto has a long distance from the rear lens to the image plane. The question now is, "Which is the better form to use for a scan lens?"

It is well known that a telephoto lens has positive distortion, while the inverted telephoto lens has negative distortion. Scan lenses that have to be designed to follow the $F-\theta$ condition must have negative distortion. This suggests that the preferred solution is with the negative lens first, even though it makes a much longer system from the last lens to

the focal plane and the entrance pupil distance is considerably shorter. Most of the scan lenses in use are a derivative of this form of inverted telephoto lens, employing a negative element on the scanner side of the lens.

Often the clearance required for the scanning element causes aberration correction problems. A telecentric design provides more clearance. Strict telecentricity may introduce too much negative distortion because the positive lens has to bend the chief ray through a large angle. When there is a tight tolerance on the $F-\theta$ condition it is better to move the scanner (aperture stop) closer to the first lens. Experience has shown that it is difficult to achieve an overall length of the system (from the scan element to the image plane) of less than 1.6 times the lens focal length. The characteristics of several scan lenses are described in Reference 4; few have a smaller ratio. In cases where the distance from scanning element to the first lens surface has to be longer than the focal length, it is advantageous to use the telephoto configuration. However, it will be difficult to make the lens meet the $F-\theta$ condition. Systems like this have been used for galvanometer scanning. It is particularly useful for XY scanning systems where more space is needed between the aperture stop and the lens.

The lenses used in the above example are extreme lenses to illustrate the two cases. In most designs the Petzval radius is not set to infinity. A Petzval radius of 10 to 50 times the focal length is usually all that is needed. The two lenses are also usually made of different glass types in order to follow the Petzval rule: to increase the Petzval radius, the negative lenses should work at low aperture and have a low index of refraction and the positive lenses should work at high aperture and have a high index of refraction. It has been pointed out[8] that if the incoming beam is slightly diverging, instead of collimated, it increases the radius of the Petzval surface. The diverging beam in effect adds positive field curvature. The idea has occasionally been used in systems, but the focus of the collimator lens has to be set at the correct divergence—not as convenient to set as strictly collimated.

Some lenses that are required to image small spot diameters (2–4 µm) use negative lenses on both sides of the positive lens to correct the Petzval curvature. Examples are shown in Section 2.12.9.

2.6.1 Correction of First-Order Chromatic Aberrations

The correction of axial and lateral chromatic aberrations illustrated in Figure 2.12a and b, respectively, is usually a challenge with scan lenses because the aperture stop is remote from the scan lens. Some system specifications call for simultaneous scanning of two or more wavelengths. These lenses have to be color-corrected at multiple wavelengths for no change in focus or focal length—that is, designed to be achromatic. Axial (or longitudinal) color, a marginal ray aberration, is a variation of focus with wavelength and is directly proportional to the relative aperture and is independent of field. Lateral (or transverse) color, a chief ray aberration, is a variation of lens magnification or scale with wavelength and is directly proportional to the field.

The simplest way to correct axial and lateral color is to make each element into an achromatic cemented doublet. To make a positive lens achromatic it is necessary to have a positive and negative lens with glass of different dispersions. The positive lens should have low-dispersion glass and the negative lens should have high-dispersion glass. The negative focal length lens reduces the positive power, so the positive lens power must be approximately double what it would be if not achromatized.
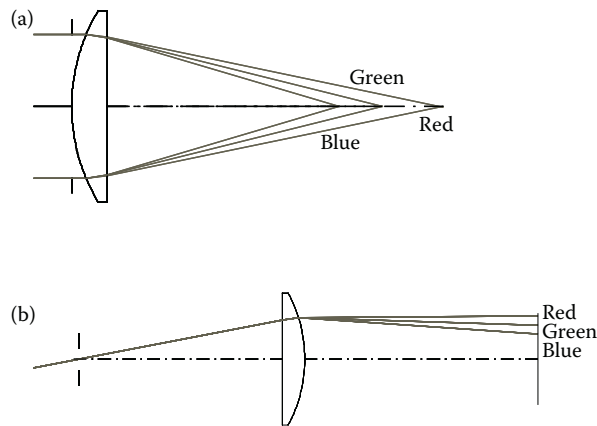
FIGURE 2.12
(a) Axial color aberration (focus change with wavelength); (b) Lateral color aberration (magnification change with wavelength) with a remote pupil.

This procedure halves the radii so the thickness must be increased in order to maintain the lens diameter. In scan lenses, the lens diameters are determined by the height of the chief ray, so the lenses are much larger in diameter than indicated by the axial beam. As thickness is increased to reach the diameters needed, the angles of incidence on the cemented surfaces increase, resulting in higher-order chromatic aberrations. When the angles of incidence in an achromatic doublet become too large, the doublet has to be split up and made into two achromatic doublets. It is safe to say that asking for simultaneous chromatic correction can more than double the number of lens elements.

Materials used for the lenses, mirrors, and mounting can be affected by environmental parameters such as temperature and pressure. For broadband systems or systems where wavelength can vary over time and/or temperature, the chromatic variation in the third-order aberrations is often the most challenging aberrations to correct. While achromats corrected for primary color use glasses with dissimilar chromatic dispersion, achromats also corrected for secondary color in addition use glasses with similar partial dispersion (i.e., glasses with similar rate of change in dispersion with wavelength). Where glasses with similar dispersion are impractical or not available, an additional element to form a triplet is used to synthesize the glass relationships needed to correct the higher-order chromatic aberrations.

Some specifications ask for good correction for a small band of wavelengths where small differences due to color can be corrected by refocus or by moving the elements. These systems do not need full color correction, and they can be designed to meet other more demanding requirements. The highest performance scan lenses are usually used with strictly monochromatic laser beams.

2.6.2 Properties of Third-Order Aberrations

The ultimate performance of any unconstrained optical design is almost always limited by a specific aberration that is an intrinsic characteristic of the design form. Familiarity with the aberrations and lens forms is still an important ingredient in a successful design optimization. Understanding of the aberrations helps designers to recognize lenses that are incapable of further optimization, and gives guidance in what direction to push a lens

TABLE 2.2

Aperture (*F/#*) and Field (*θ*) Dependence of
Third- and Fifth-Order Transverse Aberrations

Transverse Aberration	Third-order	Fifth-order
Spherical	$(F/\#)^{-3}\,\theta^0$	$(F/\#)^{-5}\,\theta^0$
Coma	$(F/\#)^{-2}\,\theta^1$	$(F/\#)^{-4}\,\theta^1$
Astigmatism	$(F/\#)^{-1}\,\theta^2$	$(F/\#)^{-1}\,\theta^4$
Field curvature	$(F/\#)^{-1}\,\theta^2$	$(F/\#)^{-1}\,\theta^4$
Distortion	$(F/\#)^0\,\theta^3$	$(F/\#)^0\,\theta^5$

Source: Thompson, K.P. *Methods for Optical Design and Analysis—Seminar Notes*; Optical Research Associates: California, 1993.

that has strayed from the optimal configuration. Table 2.2 summarizes the dependence of third- and selected fifth-order aberrations on aperture (F/#) and field (*θ*).

An understanding of the source of aberrations and their elimination comes from third-order theory. A detailed description of the theory is beyond the scope of this chapter, but can be found in References 6, 9–11. The following discussion will touch on these aberrations with the intent to provide a familiarity and some rules of thumb as guidelines for the design of scan lenses.

Third-order theory describes the lowest-order monochromatic aberrations in an optical system. Any real system will usually have some balance of third-order and higher-order aberrations, but the basic third-order surface-by-surface contributions are important to understand. These aberrations are illustrated in Figure 2.13 and briefly described below.

2.6.2.1 Spherical Aberration

This aberration is a result of a lens with different focal lengths for different zones of the aperture, a consequence of greater deviations of the sine of the angle and paraxial angle. It is an aperture-dependent aberration (varying with the cube of the aperture diameter) that causes a rotationally symmetrical blurred image of a point object on the optical axis. In rotationally symmetric optical systems it is the only aberration that occurs on the optical axis, but, if present, it will also appear at every object point in the field—in addition to other field aberrations.

2.6.2.2 Coma

This aberration is the first asymmetrical aberration that appears for points close to the optical axis. It is a result of different magnifications for different zones of the aperture. Coma gets its name from the shape of the image of a point source—the image blur is in the form of a comet. The coma aberration blur varies linearly with the field angle and with the square of the aperture diameter.

2.6.2.3 Astigmatism

When this aberration is present, the meridional fan of rays (the rays shown in a cross-sectional view of the lens) focuses at the *tangential focus* as a line perpendicular to
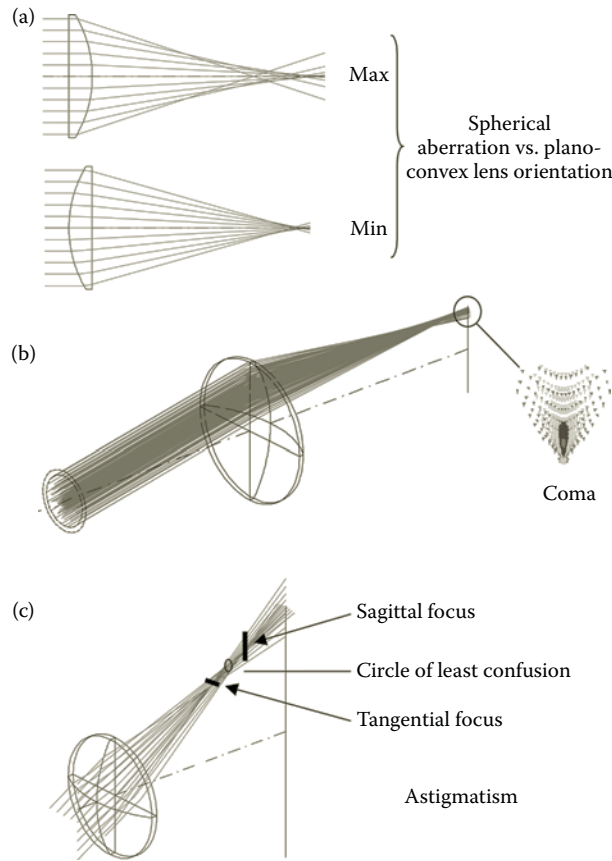
FIGURE 2.13
Aberrations: (a) spherical, (b) coma, and (c) astigmatism.

the meridional plane. The sagittal rays (rays in a plane perpendicular to the meridional plane) come to a different line focus perpendicular to the tangential line image. This focus position is called the *sagittal focus*. Midway between the two focal positions, the image is a circular blur with a diameter proportional to the NA of the lens and the distance between the focal lines. The third-order theory shows that the tangential focus position is three times as far from the Petzval surface as the sagittal focus. This is what makes the Petzval field curvature so important. If there is Petzval curvature, the image plane cannot be flat without some astigmatism. The astigmatism and the Petzval field sags both increase proportional to the square of the field. They increase faster than coma and become the most troublesome aberrations as the field (length of scan) is increased.

2.6.2.4 Distortion

Distortion is a measure of the displacement of the real chief ray from its corresponding paraxial reference point (image height $Y = F * \tan \theta$) and is independent of *f*-number. Distortion does not result in a blurred image and does not cause a reduction in any

measure of image quality (such as MTF). In an aberration-free design, the center of the energy concentration is on the chief ray. The third-order displacement of the chief ray from the paraxial image height varies with the cube of the image height. The percent distortion varies as the square of the image height.

Earlier it was noted that the distortion has to be negative in order to meet the $F-\theta$ condition. Third-order distortion refers to the displacement of the chief ray. If the image has any order of coma, it is not rotationally symmetric. The position of the chief ray may not represent the best concentration of energy in the image; there may be a displacement. Here the specification for linearity of scan becomes difficult. If there is a lack of symmetry in the image, then how does one define the error? If MTF is used as a criterion, this error is a phase shift in the tangential MTF. If an encircled energy criterion is used, then what level of energy should be used? When a design curve of the departure from the $F-\theta$ condition is provided, it usually refers to the distortion of the chief ray. The designer must therefore attempt to reduce the coma to a level that is consistent with the specification of the $F-\theta$ condition, or use an appropriate centroid criterion.

2.6.3 Third-Order Rules of Thumb

Collective surfaces[7] almost always introduce negative spherical aberration. A collective surface bends a ray above the optical axis in a clockwise direction as shown in Figure 2.14. There is a region where a collective surface introduces positive spherical aberration. This occurs when the axial ray is converging to a position between the center of curvature of the surface and its aplanatic point. When a converging ray is directed at the aplanatic point the angles of the incident and refracted rays, with respect to the optical axis, satisfy the sine condition $U/U2 = \sin U/\sin U2$, and no spherical or coma aberrations are introduced. Unfortunately this condition is usually not accessible in a scan lens. Surfaces with positive spherical aberration are important because they are the only sources of positive astigmatism.

Dispersive surfaces always introduce positive spherical aberration. A dispersive surface bends rays above the axis in a counterclockwise direction. In order to correct spherical aberration it is necessary to have dispersive surfaces that can cancel out the under correction from the collective surfaces.

Coma can be either positive or negative, depending on the angle of incidence of the chief ray. This makes it appear that the coma should be relatively easy to correct, but in the case of scan lenses it is difficult to correct the coma to zero. The primary reason is that the aperture stop of the lens is located in front of the lens. This makes it more difficult to find surfaces that balance the positive and negative coma contributions.
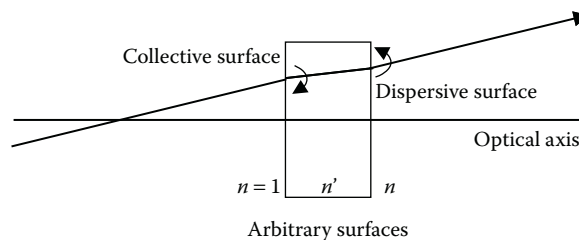
FIGURE 2.14
Simplest example of collective and dispersive surfaces.

As the field increases, astigmatism dominates the correction problem. The astigmatism introduced by a surface always has the same sign as the spherical aberration. When the lenses are all on one side of the aperture stop, this makes it difficult to control astigmatism and coma. A lens with a positive focal length usually has an inward-curving field so the astigmatism has to be positive. This is the reason that a designer must have surfaces that introduce positive spherical aberration. Because distortion is an aberration of the chief ray, surfaces that are collective to the chief ray will add negative distortion and dispersive surfaces will add positive distortion.

2.6.4 Importance of the Petzval Radius

Even though the Petzval curvature is a first-order aberration, it is closely related to the third-order because of the 3:1 relation with the tangential and sagittal astigmatism. It is not possible to eliminate Petzval curvature by merely setting up the lens powers so that the Petzval sum is zero. By doing this, the lens curves become so strong that higher-order aberrations are introduced, causing further correction problems. For this reason it is important to set up the initial design configuration with a reasonable Petzval field radius, and the designer should continually note the ratio of the Petzval radius to the focal length.

An estimate of the desired Petzval radius for a flatbed scan lens can be derived based on third-order astigmatism and DOF. Eliminating third-order astigmatism, the tangential and sagittal fields will coincide with the Petzval surface. The maximum departure of this Petzval surface from a flat image plane defined over a total scan line length L is given by the sag[10,12]

$$dz = -\frac{L^2}{8*} \quad \text{[Petzval curvature]}.$$

Setting ™z equal to the total DOF from Equation 2.17 yields the relationship

$$4l(F/\#)^2 = -\frac{L^2}{8*} \quad \text{[Petzval curvature]}.$$

The Petzval radius relative to the lens focal length F is then given by

$$\frac{\text{[Petzval radius]}}{F} = \frac{-L^2}{[32l(F/\#)^2 F]}. \tag{2.22}$$

Section 2.9 describing some typical scan lenses lists this ratio in Table 2.6 as a guideline for each application. The ratio is only an approximation. Lenses operating at large field angles or small $F/\#$ values will have high-order aberrations not accounted for in the equation. Depending on the type of correction, the final designed ratio may be higher or lower than given by the above equation. Furthermore, negative lenses working at low aperture and positive lenses working at large aperture reduce the Petzval field curvature. The negative lenses should have a low index of refraction, and the positive lenses should have a high index of refraction, atypical for an achromatized optical system.

In most monochromatic scan lenses, the negative lens will have a lower index of refraction than the positive lenses. The positive lenses will usually have an index of refraction above 1.7, while the negative lenses will usually have values around 1.5. Lens design programs can vary the index of refraction during the optimization. Occasionally in a lens with three or more elements, the optimized design violates this

rule and one of the positive lenses turns out to have a lower index of refraction than the others. This may mean the design has more than enough Petzval correction, so the index of one of the elements is reduced in order to correct other aberrations, or it may mean that one of the positive lenses is no longer necessary. To remove such an element during the optimization process, distribute its net power (by adding or subtracting curvature to one or both neighboring surfaces) and optimize for a few iterations using curvatures and a few constraints to reinitialize the design before proceeding with the full optimization.

2.7 SPECIAL DESIGN REQUIREMENTS

This section discusses specific optical design requirements for different types of laser scanners.

2.7.1 Galvanometer Scanners

Galvanometer scanners are used extensively in laser scanning. Their principal disadvantage is that they are limited in writing velocity. Their many advantages from an optical perspective are:

- The scanning mirror can rotate about an axis in the plane of the mirror. The mirror can then be located at the entrance pupil of the lens system and its position does not move as the mirror rotates.
- The F–θ condition is often not required, for the shaft angular velocity of the mirror can be controlled electronically to provide uniform spot velocity.
- The galvanometer systems are suitable for X and Y scanning.

Galvanometer scanners provide the easiest way to design an XY scanning system. The two mirrors, however, have to be separated from each other, and this means the optical system has to work with two separated entrance pupils, with considerable distance between them. This in effect requires that the lens system be aberration corrected for a much larger aperture than the laser beam diameter. A system demanding both a large aperture and large field angle will have different degrees of distortion correction for the two directions of scan. In principle the distortions can be corrected electronically, but this adds considerable complexity to the equipment.

An alternate approach is to use a telescope (afocal) relay system with one scanner placed at the entrance pupil and the other placed at the exit pupil of the telescope relay. The telescope adds complexity and field curvature to the design, but also an intermediate image that can be useful in dealing with the field curvature. Systems requiring high resolution (a large number of image points over a given length) should avoid extra relay lenses that add Petzval field curvature, the aberration that often limits optical performance in scanning systems.

For precision scanning, any wobble the galvanometer mirror may have can be corrected with cylindrical optics, as described in the next section on polygon scanning.

2.7.2 Polygon Scanning

Some precision scanning system require extreme uniformity of scanning velocity, sometimes as low as 0.1%, with the addressability of a few microns. These requirements of high-speed scanning velocities force systems into high-speed rotating elements that scan at high-uniform velocity. Polygon and holographic scanners are most commonly used in these applications.

Special design requirements that must be considered in the design of lens systems for polygons that affect optical quality are scan line bow, beam displacement, and cross-scan errors.

2.7.2.1 Bow

The incoming and exiting beams must be located in a single plane that is perpendicular to the polygon rotation axis. Error in achieving this condition will displace the spot in the cross-scan direction by an amount that varies with the field angle. This results in a curved scan line, which is said to have *bow*. The spot displacement as a function of field angle is given by the equation

$$E = F \sin a \left(\frac{1}{\cos q - 1} \right) \tag{2.23}$$

where F is the focal length of the lens, θ is the field angle, and α is the angle between the incoming beam and the plane that is perpendicular to the rotation axis.

The optical axis of the focusing lens should be coincident with the center of the input laser beam, hereafter referred to as the *feed beam*. Any error will introduce bow. The bow introduced by the input beam not being in the plane perpendicular to the rotation can be compensated for, to some extent, by tilting the lens axis. Some system designers have suggested using an array of laser diodes to simultaneously print multiple rasters. Only one of the diodes can be exactly on the central axis, so all other diode beams will enter and exit the scanner out of the plane normal to the rotation axis, so bow will be introduced. The amount will increase for diodes farther away from the central beam. There is no simple remedy to this problem.

2.7.2.2 Beam Displacement

A second peculiarity of the polygon is that the facet rotation occurs around the polygon center rather than the facet face. This causes a facet displacement and a displacement of the collimated beam as the polygon rotates, as illustrated in Figure 2.15. This displacement of the incoming beam means that the lens must be well corrected over a larger aperture than the laser beam diameter. A comprehensive treatment of the center-of-scan locus for a rotating prismatic polygonal scanner is given in Reference 13.

2.7.2.3 Cross-Scan Errors

Polygons usually have pyramidal errors in the facets as well as some axis wobble. These errors cause cross-scan errors in the scan line. These errors must be corrected to a fraction of a line width, typically on the order of a one-fourth to a one-tenth of a line width. When designing a 2400 DPI (dots per inch) high-resolution system, the tolerable error can be less than 1 µm. A system with no cross-scan error correction using a 700-mm focal length lens would require pyramid errors no greater than 1.4 µ radians.
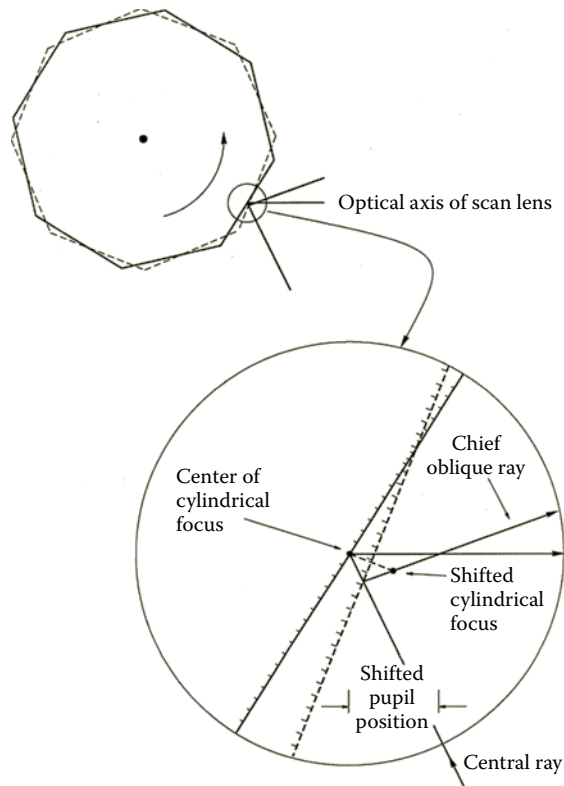
FIGURE 2.15
Facet rotation around the polygon center causes a translation of the facet, resulting in a beam displacement. (From Hopkins, R.E.; Stephenson, D. Optical systems for laser scanners. In *Optical Scanning*; Marshall, G.F., Ed.; Marcel Dekker: New York, 1991; 27–81. With permission.)

Correction methods for cross-scan errors due to polygon pyramidal errors can include deviation of the feed beam, the use of cylindrical and anamorphic lenses to focus on the polygon, an anamorphic collimated beam at the polygon, or use of a retro-reflecting prism to autocorrect.

Deviating the feed beam to anticipate the cross-scan errors at the polygon that are predictable and measurable can be accomplished by tilting a mirror, moving lens, or steering with an acousto-optic deflector (AOD). This method cannot correct for the random errors caused by polygon bearing wobble and is therefore limited in its application.

The diagram in Figure 2.16 shows how the use of cylindrical lenses can reduce the effects of wobble in the facet of a polygon. The top figure illustrates the in-scan plane, showing the length of the scan line. The lower section shows the cross-scan plane, where the laser beam is focused on the facet of the mirror by a cylindrical lens in the collimated beam. It then diverges as it enters the focusing lens. The scan lens with rotational symmetry cannot focus the cross-scan beam to the image plane without the addition of a cylindrical lens. The focal length and position of the cylindrical lens depend on the distance from the polygon facet to the all-spherical scan lens and the NA of the cylindrical lens that focuses a line image on the facet.

In order to form a round image in the scan plane, the beam in the cross-scan plane must focus with the same NA as in the in-scan plane. The ratio of the cross-scan NA at the
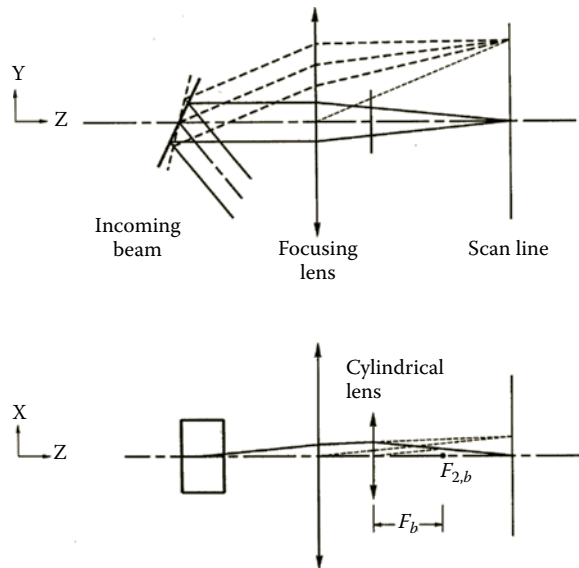
FIGURE 2.16
The use of a cylindrical lens to focus a line on the facet can reduce the cross-scan error caused by facet wobble. (From Hopkins, R.E.; Stephenson, D. *Optical systems* for laser scanners. In *Optical Scanning*; Marshall, G.F., Ed.; Marcel Dekker: New York, 1991; 27–81. With permission.)

facet to the cross-scan NA at the scanned image defines the cross-scan magnification of the lens. The selection of cylinder powers before the scan, between the scanner and scan lens, and/or between the scan lens and image plane will affect the sensitivity to wobble errors at the edges of the depth of field. Generally a cross-scan magnification of near 1:1 optimizes the correction in the presence of this polygon facet displacement from the line focus.

When the facet rotates to direct the light to the edge of the scan, the distance to the spherical lens increases. In the cross-scan plane the optical system is focusing the beam from a finite object distance. When the facet rotates to direct the light to the edge of the scan, the object distance increases so there is a conjugate change. As the scan spot moves from the center of scan, the object distance in the cross-scan plane also increases. The image conjugate distance is therefore shortened. The consequence of this is that astigmatism is introduced in the final image with the sagittal focal surface made inward curving. To compensate for this, the all-spherical focusing lens must be able to introduce enough positive astigmatism to eliminate the total astigmatism.

It has been shown[5] that placing a toroidal lens between the facet and the all-spherical lens may reduce this induced astigmatism. The toroidal surface in-scan radius of curvature should be located near the facet. In the cross-scan plane the curve should be adjusted to collimate the light. However, this solution bears with it the cost of special tooling, and imposes severe procurement, testing, and alignment challenges.

An anamorphic collimated beam at the polygon combined with a scan lens having a short cross-scan focal length and long in-scan focal length, as illustrated in Figure 2.17, can be used to reduce the effects of facet wobble. The reduction in sensitivity relative to no correction is simply the ratio of cross-scan to in-scan focal lengths, accomplished by adding cross-scan cylindrical lenses to modify the all-spherical in-scan lens to an inverse
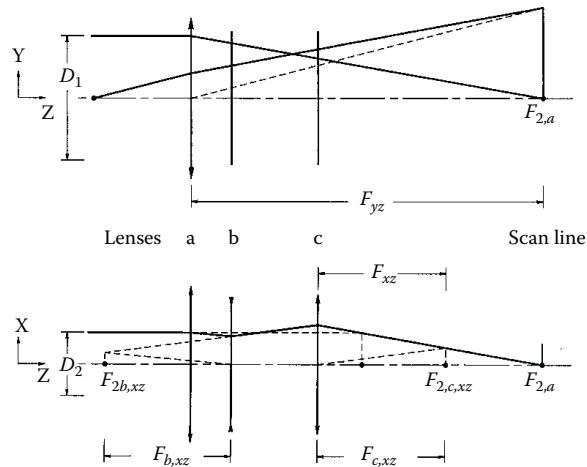
FIGURE 2.17
An anamorphic beam incident on the facet will also reduce cross-scan error. (From Hopkins, R.E.; Stephenson, D. Optical systems for laser scanners. In *Optical Scanning*; Marshall, G.F., Ed.; Marcel Dekker: New York, 1991; 27–81. With permission.)

telephoto cross-scan configuration. The feed beam is likewise compressed in the cross-scan plane to provide the necessary round beam converging on the image plane. The diagrams show the two focal lengths of the scan lens as F_{yz} and F_{xz}. In its simplest terms, the feed beam is compressed with an inverted cylinder beam expander and a comparable cylinder beam expander is placed before, distributed across, or after the scan lens. This system does introduce some conjugate shift astigmatism, but it eliminates the bow error, because collimated light is incident on the facet.

Placing the negative cylinder close to the all-spherical focusing lens and placing the positive cylinder as close to the image plane as is practical reduce the cylindrical lens powers. The position of the positive cylindrical lens, however, must consider such things as bubbles or defects on the surfaces of the lens. The beam size is extremely small when the lens is placed close to the focal plane and the entire beam can be blocked with a dust particle.

Systems using a retro-reflective prism (with 90° roof edge) that reflects the scanning beam back onto the facet face before it passes to the scan lens have been built to correct for facet wobble. The optical error introduced by the pyramidal or wobble error in the polygon is canceled on the second pass. Unfortunately the facet face has to be more than twice the aperture required to reflect the beam in a single reflection system to keep the retro-reflective beam on the facet. Consequently this configuration has low scan efficiency and limited uses.

2.7.2.4 Summary

Axis wobble and pyramidal error cause serious problems by introducing cross-scan errors. There are ways to reduce the cross-scan errors, but many other challenges are introduced. The use of cylinders results in procurement and alignment issues. The conjugate shift is difficult to visualize because the entire line image on the facet is not in focus and the analysis can be complex. The only way to determine accurately the combination of all the effects—pyramidal error in the facets, translation of the facet during its rotation, bow

tie effect, and the conjugate shift—is to raytrace the system and simulate the precise locations of the facet as it turns through the scanning positions. This can be done using the *multiconfiguration modes* available in most optical design programs. The multiconfiguration technique of design is discussed in more detail in the following sections.

2.7.3 Polygon Scan Efficiency

Figure 2.18 shows one facet of a polygon with feed beam for scanning. The parameters and relationships that determine the limits of scan efficiency and minimum size for a polygon scanner (assuming no facet tracking) are D = beam diameter, β = nominal feed beam offset angle at center of scan, ε = facet edge roll zone, $\alpha = 2\pi/$(no. of facets N) = angular extent of facet, and $\delta \sim [D/\cos(®) + \Sigma]/r$ = angular extent of beam plus roll zone, where the scan efficiency limit for a given polygon is

$$h_s = 1 - \frac{(d + e/r)}{a} \tag{2.24}$$

and the minimum polygon circumscribed radius is

$$r > \frac{[D/\cos(b) + e]}{[a * (1 - h_s)]}. \tag{2.25}$$

Given the circumscribed radius of the polygon, feed beam angle, and facet scan angle, and assuming the edge roll zone (unusable part of the facet clear aperture) is negligible, the maximum beam diameter that can be supported without vignetting is given by the equation

$$D < r\cos(b)[a * (1 - h_s)]. \tag{2.26}$$

The circumscribed cord defined by polygon facet less the cords defined by the feed beam footprint on the circumscribed circumference and roll zone will limit the useful polygon rotation.

FIGURE 2.18
Diagram of a polygon facet in its central position and parameters describing the maximum beam diameter that can be reflected with no vignetting through a scan angle at peak efficiency.

A sample polygon and scan lens design might have the following specifications: 2000 DPI with a 12.7 µm $1/e^2$ spot diameter; a wavelength of 0.6328 µm; a scan length L of 18 in (457.2 mm); an eight-sided polygon with a facet angle α of 0.7854 radians, a scan efficiency η_s of 60%, and a feed beam angle β of 30°.

The derived system parameters are: $F/\#$ of the lens for this spot diameter is $F/11$ (given by Equation 2.14); scan lens focal length F is 485.1 mm [given by $L/(2\alpha^*\eta_s)$]; and beam diameter D is 44.1 mm. The required circumscribed radius of the polygon is greater than 162 mm, with a facet face width of 2.8 times the diameter of the incoming beam.

To achieve a certain resolution (scan a given number of image points on a single scan line), there is a trade-off between the scan angle θ and the diameter of the feed beam (Equation 2.16). To achieve compactness with a smaller polygon, a smaller feed beam would be required along with a greater scan angle. The search for system compactness drives the field angle to larger and larger values. Scan angles above 20° increase the difficulty in correcting the $F/\#$. The conflict can be somewhat resolved by using a smaller angle of incidence ®, but then there may be interference between the incident beam and the lens mount. A compromise between these variables requires close cooperation between the optical and mechanical engineering effort.

Figure 2.19 illustrates the relationship between the number of polygon facets, polygon diameter, and scan efficiency. The previous example is illustrated in the eight facet plots. Polygons typically work at around 50% scan efficiency because of the feed beam diameter required by the resolution coupled with the rotating polygon scanners size limitations and cost considerations. A comprehensive treatment of the relationship between the incident beam, the scan axis, and the rotation axis of a prismatic polygonal scanner is given in Reference 14.

Note: Care should be taken in the specification of the *scan angle*. Without a clear definition it can be interpreted as the mechanical scan angle, the optical scan angle, or even the optical half scan angle.

2.7.4 Internal Drum Systems

As stated before, internal drum scanning systems are least demanding on the optical system, because the lenses do not have to cover a wide field. Most of the burden is shifted
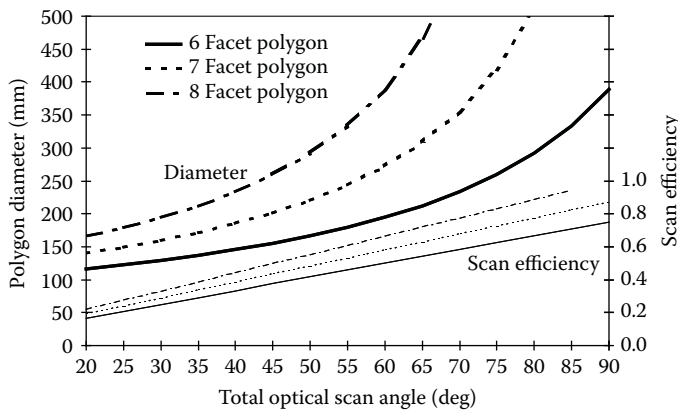
FIGURE 2.19
Polygon diameter and scan efficiency vs. total optical scan angle (D = 44.1 mm, $\varepsilon = 0$, $\beta = 30°$).

to the accurate mechanical alignment of the turning mirror. The concept of the internal drum scanner can be applied to a flatbed scanner by using a flat-field lens. The system then becomes the equivalent of a pyramidal scanner with only one facet. All of these systems have common alignment requirements.

The nominal position of the turning mirror is usually set at 45° (0.785 radians) with respect to the axis of rotation. It does not have to be exactly 45° as long as the collimated feed optical beam enters parallel to the axis of the rotation. The latter condition is needed to eliminate bow in the scan line. In a perfectly aligned system the ray that passes through the nodal points of the lens must meet the deflecting mirror on the axis of rotation. When the lens is placed in front of the turning mirror, the second nodal point of the lens must be on the rotation axis of the mirror. When the lens is placed after the turning mirror, its first nodal point must be on the ray that intersects the mirror on the rotation axis. There is some advantage in placing the lens between the mirror and the recording plane. In this position the lens has a shorter focal length, and the bow resulting from any error in the nodal point placement of the lens is reduced.

2.7.5 Holographic Scanning Systems

From an optical designer's perspective, holographic scanning systems have an advantage, as the need for wobble correction can be reduced significantly without resorting to cylindrical components. Conversely, they usually require some bow correction, and if used with laser diodes (which exhibit wavelength shifts), they require significant color correction. Line bow correction can be achieved by using a prism (or grating) component after the holographic scanning element and/or adding complex holograms, reflective and refractive optical components to the lens system. The prism component introduces bow to balance out the bow in the same way that a spectrographic prism adds curvature to the spectral lines. The lens can be tilted and decentered as an alternative method for reducing the bow. Holographic scanning systems are discussed in greater detail in Section 2.11.

2.8 LENS DESIGN MODELS

Regardless of the ultimate complexity of a scanning system, a simple model is often the best starting point for the design of a scan lens, and sometimes all that is needed. The exception is when adapting or tweaking a previous complex design model for minor changes in wavelength, scan length, or resolution. In a simple model, the actual method by which the beam deflection is introduced is not included in the lens design; the beam deflection method is assumed to introduce only angular motion; and it neglects any beam displacement that may occur due to the deflection method. The lens is modeled and optimized to perform at several field angles, in much the same way a standard photographic objective is optimized. The main differences are in the external placement of the aperture stop where the chief ray for each field passes through its center as if scanning and the optimization with distortion constraints for F–θ linearity. A detailed example demonstrating the simple model is developed in Section 2.8.1 with additional examples provided in Sections 2.9.1 through 2.9.7.

In practice, the reason that all parallel bundles in Figure 2.20a appear to pivot about the center of this external aperture stop surface is that a fixed beam is incident on a beam deflector rotating in proximity to the stop. If the mechanical rotation axis of this deflector intersects the plane of the mirror facet and the optical axis of the scan lens, then the simple model is accurate. This is the case for galvanometer-based systems, where the mechanical rotation axis is close enough to the plane of the mirror facet that the deviation from the simple model is negligible.

At some point in the design process it must be decided how rigorously the geometry of the moving deflector needs to be modeled. At the very least, the design should be analyzed for the actual angular motion with the rotation of the pupil and the displacement of the beam to determine as-built performance of a design and validate the effectiveness of the simple model. This can sometimes be accomplished without modeling the actual scanner, as described below.

Multifaceted holographic deflectors do not suffer from beam displacement, even though their mechanical rotation axis is some distance from the active region of the facets. Where dispersion from the hologon is not an issue, a simple model can suffice for the design of a lens for a hologon-based scan system, although complex truncation and multifacet illumination effects are ignored.

2.8.1 Anatomy of a Simple Scan Lens Design

The following is a description of a scan lens design, which begins with the design specifications outlined in Table 2.3 and describes the evolution of a design to meet these specifications.[15] The first five numbered specifications (scan line length, wavelength, resolution, image quality, and scan linearity) are the very minimum required to begin the optical design. The laser source for this exercise is a diode operating near-Gaussian TEM00 with a wavelength that can drift with temperature and power, and emit over a bandwidth of 25 nm. Parameters listed with no numbers are provided as reference or potential additional specifications.

Beginning with the resolution requirement of 300 DPI based on $1/e^2$ and a nominal wavelength of 780 nm, the ideal Gaussian waist size (pixel size) is

$$2w_0 = 84.7\,\mu\text{m}$$

with a NA defined by

$$NA = q_{1/2} = \frac{1}{pw_0}$$
$$= 0.006.$$

Setting the optical scan angle at ±15° (0.26 radians) for the initial (back-of-the-envelope) calculations, the required focal length for a 216 mm scan line is determined by

$$Fq = 216\,\text{mm}$$
$$F = \frac{216}{(2 \times 0.26)} = 415\,\text{mm}$$
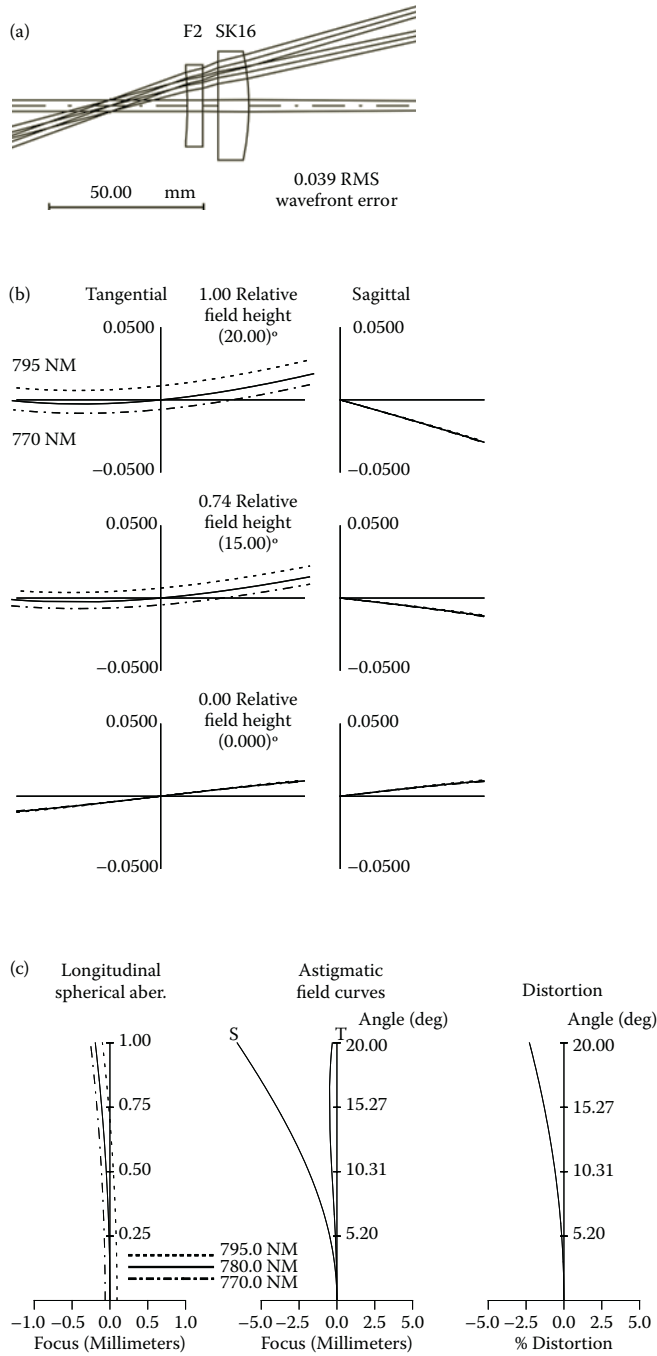
FIGURE 2.20
(a) Two-element starting scan lens design; (b) ray aberration plots for starting design; and (c) field performance plots for starting design.

TABLE 2.3

Specifications for a Scan Lens Example

Parameter	Specification or Goal
1. Image format (line length)	216 mm (8.5 in)
2. Wavelength	780 + 15, −10 nm
3. Resolution ($1/e^2$ based)	300 DPI (600 DPI goal)
4. Wavefront error	<1/20 wave RMS
5. Spot size and variation over scan	TBD ± 20% (± 10%, goal)
6. Scan linearity (F–∪ distortion)	<1%(<0.1%, <0.03%)
7. Depth of field	No specific requirement
8. Telecentricity	No scan length versus DOF control
9. Optical scan angle	TBD (±15 to 45°)
10. *F*-number	Defined by resolution
11. Effective focal length	… by resolution and scan angle
12. Overall length	<500 mm
13. Scanner clearance	>25 mm
14. Image clearance	>10 mm
15. Scanner requirements	TBD (type, size)
16. Packaging	TBD
17. Operating/storage temperature	TBD

TBD, to be determined (at a later date); DOF, depth of focus.

Experience has shown that the focal length arrived at in this first pass would likely result in a system that is too long (when considering scanner to lens clearance, the thickness of real lenses, and the image distance). To shorten the length of the optical system the optical scan angle is increased to ±20° (0.35 radians). The new required focal length then becomes

$$F = 309 \text{ mm}$$

and the required design aperture diameter is

$$EPD = F(2NA) = 3.7 \text{ mm}.$$

The simple two-lens configuration shown in Figure 2.20a comprises a concave-plano flint (Schott® F2) element and plano-convex crown (Schott SKI6) element (both by Schott Glass Technologies Inc., Duryea, PA) with powers appropriate for axial color correction selected as a starting point. Scaled for focal length and focused, the composite RMS wavefront error (weighted average over the field) is 0.038 waves. While at first glance this wavefront error appears to meet the requirements, further examination of the ray aberration and field performance plots illustrated in Figures 2.20(b) and (c) indicates substantial astigmatism limiting performance at the edge of the field and distortion that is far from F–θ.

The horizontal plot axis in Figure 2.20b is relative aperture and the vertical plot axis is the transverse ray error at the image plane. The slope of the shallow curve is a measure of focus shift and a change in slope over the relative field and between the sagittal and tangential curves is a measure of the field curvature and astigmatism. These ray aberration plots clearly show astigmatism (indicated by the slope difference between sagittal and tangential curves at 15 and 20° field angles) and lateral color (indicated by the displacement of

the tangential curves, a change in image height, for the extreme wavelengths at the 15 and 20° field angles).

The field performance plots in Figure 2.20c also indicate a very small amount of axial color and spherical aberration (displaced longitudinal curves for each wavelength of the left plot), substantial astigmatism (indicated by the departure of the sagittal field curve from the nearly flat tangential curve of the center plot), and the distortion that is closer to F–tan Θ than F–Θ (seen in the distortion curve on the right).

Optimizing this starting design with surface curvatures as variables for best spot performance while maintaining scan length with a constraint (but no distortion controls) yields the biconvex biconcave configuration illustrated in Figure 2.21. The RMS wavefront error improved after the first round of computer optimization iterations by a less than desirable balance of the astigmatism with higher-order aberrations (center field plots). The air interface between elements was deleted, leaving the three surfaces of a doublet for a second round of optimization (to test the design for a simpler solution), resulting in performance slightly better than the starting design for RMS wavefront error and astigmatism.

Adding glass variables such as index of refraction and dispersion and adding F–θ constraints, weighted rather than absolute for a more stable convergence, results in no significant change in performance, as illustrated in Figure 2.22. Splitting the power of the positive crown element between the doublet and an additional plano-convex lens provides more design variables for the next round of optimization iterations. The result is much better correction of both the astigmatism and F–θ distortion with an RMS wavefront error of 0.005 waves, but at the sacrifice of some axial color correction (a shift in focus with wavelength). This trade-off is acceptable if most of the wavelength variation is from diode to diode, where focus can be used to accommodate the different laser wavelengths, and the wavelength variations from changes in junction temperature due to power and ambient temperature are controlled to a few degrees. Selecting real glasses for the final design iterations results in the design and performance as illustrated in Figure 2.23, with no noticeable change in performance.
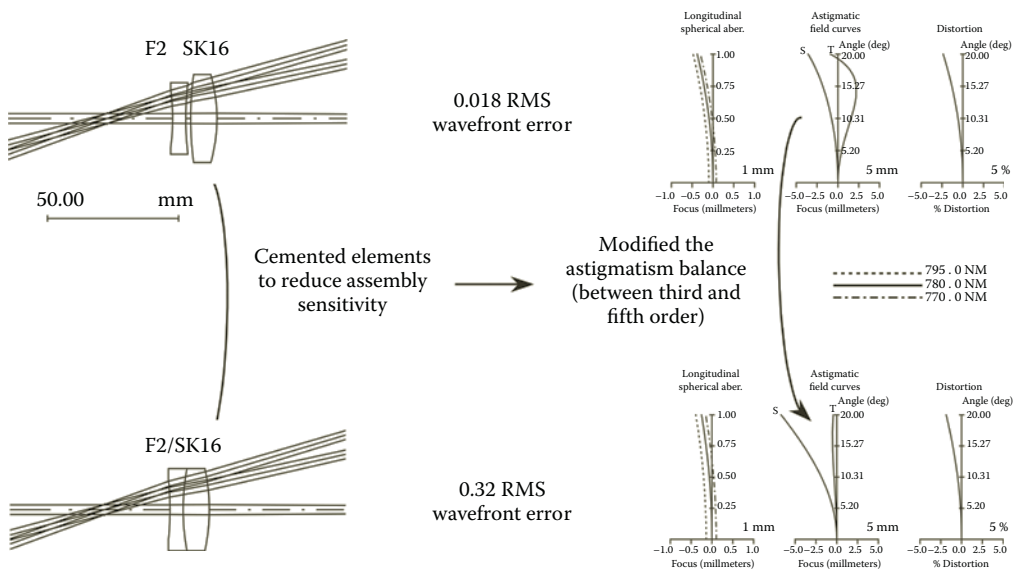
FIGURE 2.21
First and second design iterations with no distortion controls.
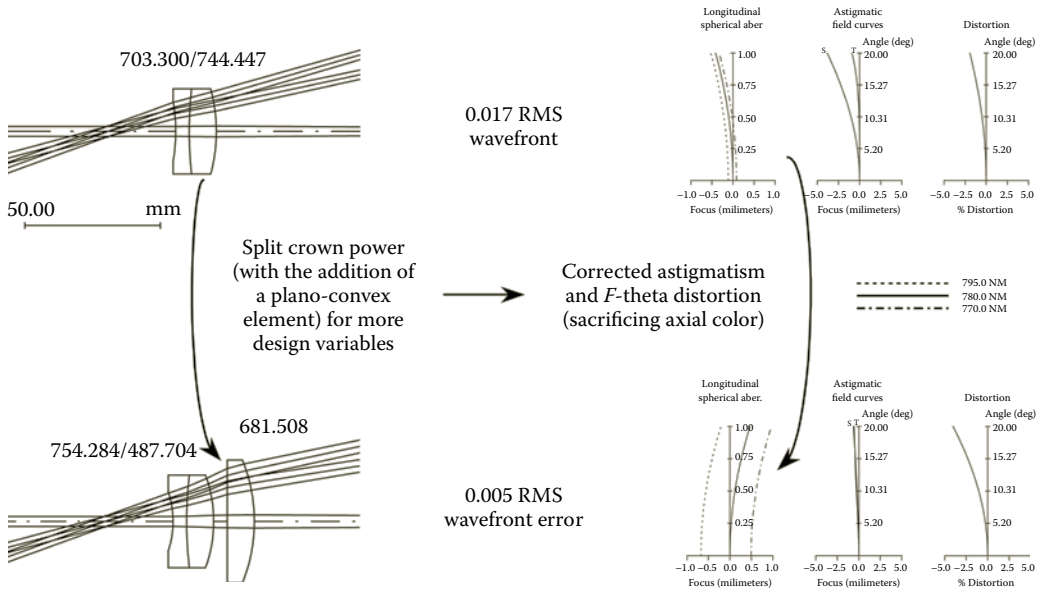
FIGURE 2.22
More design iterations (adding weighted $F-\theta$ distortion controls).
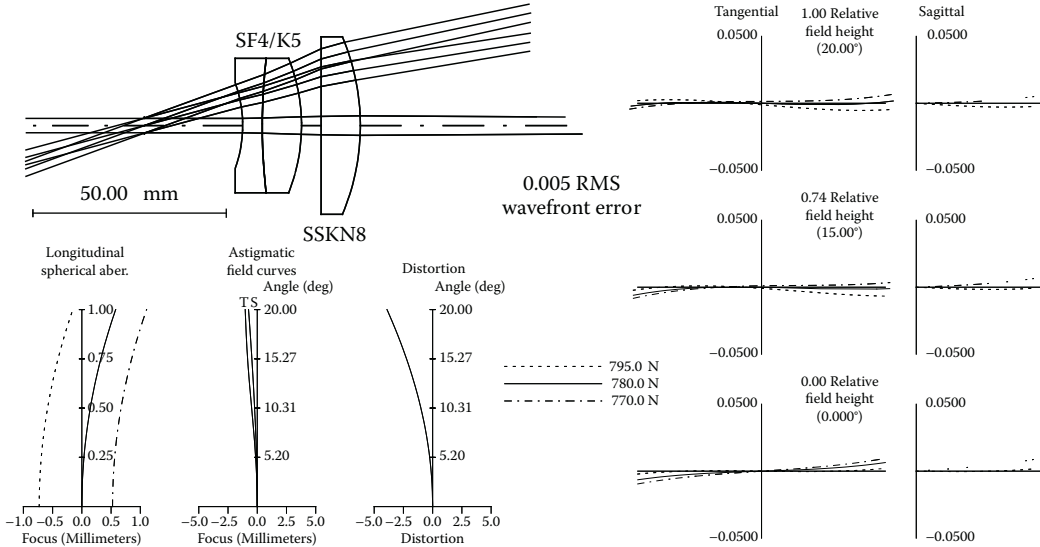
FIGURE 2.23
Final design iteration (with real glass types).

Raising the performance requirement target of this example design lens to the resolution goal of 600 DPI (double the original specification) requires twice the NA and hence twice the design aperture diameter. Design iterations for the higher resolution $F-\theta$ lens begins by adding back glass variables to deal with the pupil aberrations introduced by the larger
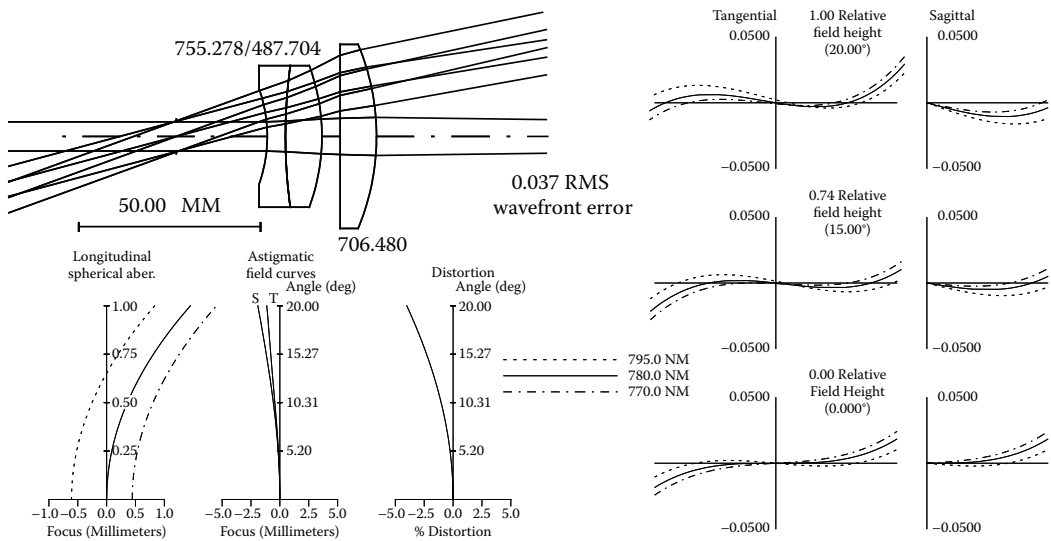
FIGURE 2.24
New design iterations (with previous *F–θ* lens and twice the design aperture diameter).

aperture, illustrated in Figure 2.24. The performance after these first iterations deteriorated to a wavefront error of 0.037 waves RMS, predominantly from spherical aberration (the tangential S-shaped curve) with a bit of coma at the edge of the field (indicated by the asymmetry in the full field tangential curve).

The performance target is raised yet again with an increase in the scan angle specification to ±30° from ±20°. The design for this exercise starts with the previous lens, scaled by the ratio of scan angles (20/30) from pupil to image including the design diameter. The performance after several more iterations and the selection of real glass types, illustrated in Figure 2.25, is improved to a 0.028 RMS wavefront error, by improving spherical aberration, axial color and astigmatism (with the help of the 2/3 scaling of the design aperture), and more linear distortion resulting in better *F–θ* correction. The final design prescription for the scan lens example is listed in Table 2.4 with performance specifications listed in Table 2.5. The calibrated *F–θ* scan distortion is plotted in Figure 2.26. It shows a linearity better than 0.1% over scan and better than 1% locally. The selection of LAFN23 in this latest design is not ideal for its availability or glass properties. Later iterations to finalize this preliminary design should explore glass types that optimize the availability, cost, and transmission along with image quality performance.

2.8.2 Multiconfiguration Using Tilted Surfaces

There is often no substitute for the introduction of a mirror surface in the lens design model to simulate the scan, where the mirror surface is tilted from one configuration (or "zoom position") to another to generate the angular scanning of the beam prior to the scan lens. Modeled in this way, there is no object field angle in the usual sense. The beam prior to the rotating mirror is stationary. When a beam of circular cross section reflects off the plane mirror, the reflected beam will have the same cross section.

That is not the case in a simple model involving object field angles and a fixed external circular stop, where bundles from off-axis field angles will be foreshortened in the
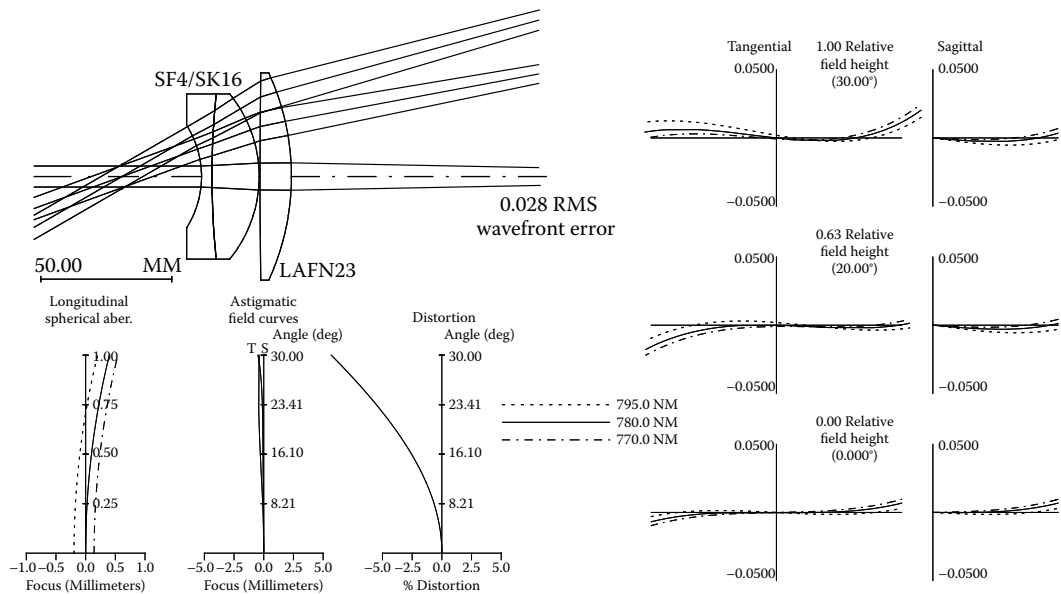
FIGURE 2.25
Final design iteration (with selected glasses).

TABLE 2.4

Final Design Prescription for the Scan Lens
Example

Surface	Radius	Thickness	Glass Type
1		30.3512	
2	−36.5662	4.0000	SF4
3	310.0920	18.0000	SK16
4	−49.8537	0.3276	
5	2541.4236	12.0000	LAFN23
6	−94.6674	270.3002	
Image		0.0462	

scan direction to an ellipse by the cosine of the field angle. As the field angle approaches 30° this effect becomes increasingly significant. At some stage in the design process the software can be tricked into enlarging the bundle in the scan direction to compensate for this foreshortening, but not all analyses may run using this work around. In particular, diffraction calculations typically rely on tracing a grid of rays that are limited by defined apertures in the optical system, which would defeat the intended effect of simple tricks.

A constant beam cross section at all scan angles can also be maintained by simply bending the optical axis at the entrance pupil in a multiconfiguration. This is often a good compromise in modeling complexity that does not require the use of reflective surfaces, but maintains the integrity of the scanned beam optical properties.

TABLE 2.5

Final Design Specifications for the Scan Lens Example

Parameter	Specification or Goal
1. Image format (line length)	216 mm (8.5 in)
2. Wavelength	770–795 nm
3. Resolution ($1/e^2$ based)	26 μm(~1000 DPI)
4. Wavefront error	<1/30 wave RMS
5. Scan linearity (F–θ distortion)	<1% (<0.2% over ± 25°)
6. Scanned field angle	±30°
7. Effective focal length	206 mm (calibrated)
8. F-number	$f/26$
9. Overall length	335 mm
10. Scanner clearance	25 mm
11. Image clearance	270 mm
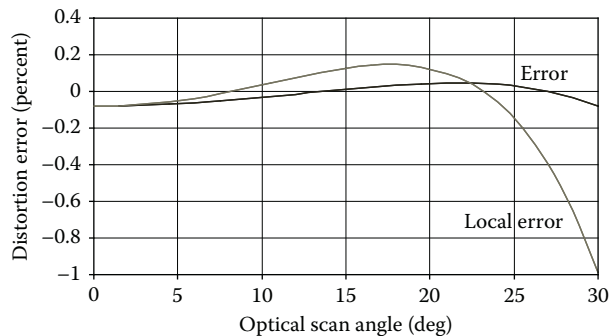
FIGURE 2.26
Final F–θ scan linearity.

Holographic deflectors may not faithfully emulate a tilted mirror. In such systems the beam cross section does change as a function of angle, where a circular input beam results in an elliptical output whose orientation changes with scan angle.

2.8.3 Multiconfiguration Reflective Polygon Model

A polygon in a lens design model is simply a mirror at the end of an arm. All rotation is performed about the end of the arm opposite the mirror. The length of the arm defined by the inscribed radius of the polygon, the location of the arm's rotation axis relative to the scan lens, and the amount of rotation about the pivot can all be (with care) optimized during the design. The facet shape is defined by an appropriate clear aperture specification on the mirror surface.

Because the mechanical rotation axis rarely intersects the optical axis of the scan lens and the mirror facet is far from the rotation axis, it is best just to apply a rigorous model that automatically accounts for the complicated mirror surface tilts, displacements, and aperture effects of the polygon scan complex geometry.

Specific pupil shifts and aperture effects could be computed and specified for each configuration, but this does not exploit the full potential of a multiconfiguration optical design program. When the model is defined in a general fashion, the actual constructional parameters of the polygon or parameters governing its interface with the feed beam and scan lens may be optimized simultaneously as the scan lens is being designed, particularly in the final stages of design. This often leads to better system solutions than simply combining devices in some preconceived way.

If the location of the entrance pupil is defined as where the chief ray intersects the optical axis of the scan lens, then the axial position of the pupil shifts with polygon rotation angle (and scan angle). This axial pupil shift requires the lens to be well corrected over an aperture larger than just the feed beam diameter. The effect is greater for systems with polygons having fewer facets, larger optical scan angles, and/or larger feed beam offset angles. High-aperture (large NA) scan lenses are especially susceptible to this effect.

Determining where the real entrance pupil is for each configuration, or how much larger of a beam diameter the lens should really be designed to accommodate, is difficult. It is especially difficult if the polygon is to be designed at the same time as the scan lens! When this pupil shift is included in the lens design model by rigorously modeling the polygon geometry, the effect on lens performance can be accurately assessed. More importantly, the lens being designed may be desensitized to expected pupil shifts and, if the polygon is being designed, the pupil shifts may be minimized. As maximum scan angle is approached, the facet size may be insufficient to reflect the entire beam. Asymmetrical truncation or vignetting occurs, which can modify the shape of spot at the image plane. Accurate aperture modeling is especially important for accurate diffraction-based spot profile calculations. In a rigorous polygon model, by putting aperture specifications on the surface that represents the reflective facet surface, all vignetting by the facet as a function of polygon rotation will be automatically accounted for.

By having a rigorous polygon model implemented, it is also possible to further evaluate what actually happens to the section of the beam that misses the facet. Classifying the rays that miss the facet as vignetted is really an oversimplification. In reality these rays will likely reflect off the tip and adjacent facet. This stray light beam may enter the lens and find its way to the image surface. Stray light problems can ruin a system. Double-pass systems are especially susceptible to this design flaw. The multiconfiguration setup can be used to evaluate stray light problems and suggest baffle designs.

2.8.4 Example Single-Pass Polygon Setup

The key to exploiting the multiconfiguration design method is to include the rotation axis when modeling the polygon. A brief overview of such a model begins with the definition of the expanded laser beam, followed by a first fold mirror, the defined polygon with its rotation axis, and finally the scan lens.

The following CODE V* sequence file (*CODE V is a trademark of Optical Research Associates, Pasadena, CA) for the lens illustrated in Figure 2.27, shows one way to set up a polygon and lens in a commercial lens design program. Standard catalog components were chosen: the six-sided polygon is Lincoln Laser's PO06-16-037, and the two-element sectioned scan lens is Melles Griot's LLS-090. The combination will create 300 DPI output using a laser diode source.

2.8.4.1 Multiconfiguration Code V Lens Prescription

```
RDM; LEN
TITLE "LINCOLN PO-6-16-37 MELLES GRIOT LLS-090 90 mm F/50 31.5-deg P-468"
EPD 1.8145
PUX        1.0       ;PUY1.0          ;PUI 0.135335
DIM M
WL         780
YAN        0.0
S0         0.0    0.1e20                                ! Surface 0      "A"
S          0.0    50.8                                  ! Surface 1
           STO
S          0.0    -25.4         REFL                    ! Surface 2
XDE        0.0;   YDE 0.0;      ZDE 0.0;       BEN
ADE        -31.5; BDE 0.0;      CDE 0.0;
S          0.0    0.0                                   ! Surface 3      "B"
XDE        0.0;   YDE 0.0;      ZDE 0.0;
ADE        63.0;  BDE 0.0;      CDE 0.0;
S          0.0    -16.892507                            ! Surface 4      "C"
S          0.0    0.0                                   ! Surface 5
XDE        0.0;   YDE 10.351742; ZDE 0.0;
ADE        -31.5; BDE 0.0;      CDE 0.0;
S          0.0    19.812                                ! Surface 6      "D"
XDE        0.0;   YDE 0.0;      ZDE 0.0;
ADE        -15.5; BDE 0.0;      CDE 0.0;
S          0.0    -19.812       REFL                    ! Surface 7
S          0.0    0.0                                   ! Surface 8      "E"
XDE        0.0;   YDE 0.0;      ZDE 0.0;       REV
ADE        -15.5; BDE 0.0;      CDE 0.0;
S          0.0    16.892507                             ! Surface 9      "F"
XDE        0.0;   YDE 10.351742; ZDE 0.0;      REV
ADE        -31.5; BDE 0.0;      CDE 0.0;
S          0.0    7.0                                   ! Surface 10     "G"
S          -49.606 4.5          SK16 _ SCH0TT           ! Surface 11
S          0.0    6.3 5                                 ! Surface 12
S          0.0    5.35          SFL6 _ SCH0TT           ! Surface 13
S          -38.633 104.340988                           ! Surface 14
PIM
SI         0.0    -0.633188
ZOOM       7                                            !                "H"
ZOOM ADE S6 -15.5 -11 -7.5 0 7.5 11 15.5
ZOOM ADE S8 -15.5 -11 -7.5 0 7.5 11 15.5
GO
CA                                                      !                "I"
CIR S2     2.5    ;CIRS2        EDG            2.5
REX S7     4.7625 ;REYS7        11.43
REX S7     EDG    4.7625;       REY S7         EDG      11.43
REX S11 5.0       ;REY S11      5.1
REX S12 5.0       ;REY S12      7.6
REX S13    5.0    ;REYS13       14.9
REX S14    5.0    ;REYS14       15.8
GO
```

25.00 MM
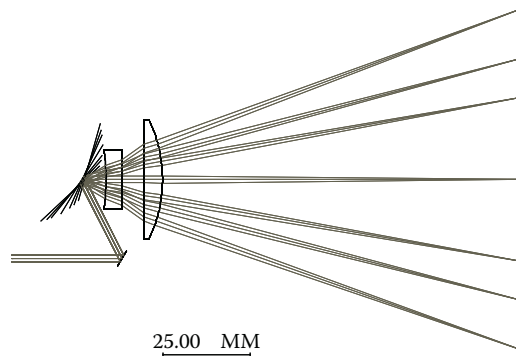
FIGURE 2.27
A multiconfiguration, single-pass polygon system.

2.8.4.2 Lens Prescription Model

Start with a collimated, expanded laser beam of the required diameter and fold its path with a −31.5° mirror tilt to obtain the desired feed angle of 63° with respect to the planned optical axis of the scan lens. The aperture stop should be defined prior to the polygon (preferably on surface 1). Any truncation of the Gaussian input beam should be done at the aperture stop. Do not flag the polygon surface as the aperture stop, because some software will automatically ray-aim each bundle to pass through the center of the stop surface.

Define a reference point that will be on the optical axis of the scan lens. It is convenient to have its location where the facet would intersect the axis when the polygon is rotated for on-axis evaluation. The surface should be tilted 63° so that any subsequent thickness would be along the optical axis.

Go to the polygon rotation center and tilt so that any subsequent thickness would be radial from the polygon center toward the facet surface. To get to the polygon center from the reference point, use a combination of surface axial and transverse decenters (traveling in right angles for simplicity). It is convenient to choose a tilt that will cause the polygon to be rotated into position for on-axis evaluation. Here, we must translate −16.9 mm away from the lens along the optical axis and then decenter up along Y with YDE = 10.4 mm. Tilting about X in the YZ-plane with ADE = −31.50° (63°/2) points the surface normal to the facet.

Before going to the polygon facet, any additional tilt about X (ADE) is specified. This is a multiconfiguration parameter: each configuration will have a different value specified for this additional tilt. Here, ADE 15.50 is specified to cause the polygon to rotate into position for maximum scan on the negative side. This is really the polygon shaft rotation angle; for nonpyramidal polygons the reflection angle (scan angle) changes at twice the rate of the shaft angle. Once the image surface is defined, the system will be defined to have seven configurations, and a different ADE value for this surface will be specified for each (step "H"). Translate to the facet using the polygon inscribed radius for thickness (19.8 mm). Now reflect. This reflection occurs at the first real surface that the beam encounters since the fold mirror. All other surfaces have been "dummy" surfaces where no reflection or refraction takes place. It is on this reflective surface that aperture restrictions may be defined to describe the shape of the facet. Use a thickness specification on this surface to

go back to the polygon center following the reflection (–19.8 mm), to maintain the integrity of the first-order optical path lengths.

Some commercial software programs (such as CODE V) have a "return" surface that one could now use to get back to the reference point defined in step B, prior to defining the scan lens. Here, a more conservative approach is taken that can be used with any software package. Undo the additional polygon shaft rotation that was done in step D. The REV flag in CODE V internally negates the angle.

Continuing to move back to the reference point defined in step B, undo the polygon shaft tilt that sets it for on-axis evaluation (ADE = 31.5°), decenter down along Y, back to the scan lens axis (YDE = –10.4 mm), and translate toward the lens along its axis to the reference. This places the mechanical axis back to the same location that it was in step B, before the reflection off the polygon. Here, the REV flag changes the signs of the tilt and decenter specifications and performs the tilt before the decenter.

Define the scan lens. The reference surface defined in step B, and returned to in steps E and F, is approximately the location of the entrance pupil. The thickness at the image surface is the focus shift from the paraxial image plane.

Having now specified a valid single-configuration system to the software, the system is redefined to have seven configurations (ZOOM 7) and the parameters that change from one configuration to another are listed. Owing to the way that the polygon was modeled, rotating it to a different position is simply a matter of changing the parameter that represents the shaft rotation angle. ADE is on surfaces 6 and 8.

Since these are catalog components, the clear apertures are available and are specified here. The rectangular aperture specifications for the polygon facet are given for surface 7.

2.8.5 Dual-Axis Scanning

When more than one galvanometer is used to generate a two-dimensional scan at the image plane it is usually necessary to use a multiconfiguration setup. Each galvanometer defines an apparent pupil and net effect of their physical separation creates a very astigmatic pupil, where X-scan and the Y-scan do not originate at the same location on the optical axis. In modeling these scanners, the optical axis or reflective surfaces representing the galvanometer mirrors may be tilted. The latter approach is usually worth the effort in order to visualize the problem and avoid mechanical interferences.

2.9 SELECTED LASER SCAN LENS DESIGNS

Scan lenses with laser beams have been considered almost from the time of the development of the first laser. Laboratory models for printing data transmitted from satellites were underway in the late 1960s. Commercial applications began coming out in the early 1970s, and laser printers became popular in the early 1980s. The range of applications is steadily increasing. With the development of new laser sources (including violet and UV) and high-precision manufacturing processes for complex surfaces (including plastics), the field of lens design offers new challenges and opportunities with room for new design concepts.

The lenses presented in this section were selected to show what appears to be the trend in development. The spot diameters are getting smaller, the scan lengths longer, and the speed of scanning higher. Some of the lenses near the bottom of the list are beginning to exhaust our present design and manufacturing capabilities. The newest requirements are reaching practical limits on the size of the optics and the cost of the fabrication and mounting of the optics. It appears that the future designs will have to incorporate mirrors and lenses with large diameters, and new methods for manufacturing segmented elements will be needed.

The lenses shown in this section start with some modest designs for the early scanners and progress to some of the latest designs. Two of the designs were obtained from patents. This does not mean that they are fully engineered designs. The rest of the designs are similar paper designs. This means that the designer has the problem "boxed in"—where all the aberrations are in tolerance and under control. The next phase begins the engineering task of preparing the lenses for manufacture, making sure that all the clear apertures will pass the rays and that the lenses are not too thick or too thin, checking availability and cost of the selected glass types as well as the experience of the shop working with those glasses. The design has to be reviewed to consider how the lenses are to be mounted. Some of the lenses may require precision bevels on the glass or a redesign may avoid this costly step. This section also includes a few comments about the designs with regard to practicality. Table 2.6 contains a summary of attributes for the selected lenses.

In the following lens descriptions all the spot diameters refer to the diameter of the spot at the $1/e^2$ irradiance level. The number of spots on the line is calculated assuming contiguous spots packed adjacent to each other at the $1/e^2$ irradiance level.

2.9.1 A 300 DPI Office Printer Lens (λ = 633 nm)

Figure 2.28 presents a patent (U.S. Patent 4,179,183, Tateoka, Minoura; December 18, 1979) assigned to Canon Kabushiki Kaisha. The patent contains a lengthy description of the design concepts used in developing a whole series of lenses. Fifteen designs are offered with the design data along with plots of spherical aberration, field curves, and the linearity of scan. The design shown is example 6 of the patent. The design data were set up and

TABLE 2.6

Summary of Attributes for the Selected Scan Lenses

Section	F	F/#	L	ROAL	RFWD	d	L/d	RBcr	RPR	REPR	NS/I	NO.el
9.1	300	60	328	1.4	0.13	70	4700	0.66	−11	−5	370	2
9.2	100	24	118	1.4	0.06	28	4300	0.34	−12	−12	920	3
9.3	400	20	310	1.6	0.17	23	13,000	0.49	−26	−30	1100	3
9.4	748	17	470	1.4	0.06	20	23,000	0.29	−15	−50	1200	3
9.5	55	5	29	2.1	0.44	5.8	5000	0.84	−32	−24	4300	3
9.6	52	2	20	4.2	0.39	4	5000	1.0	−56	−16	6350	14
9.7	125	24	70	2.3	0.8	20	3500	0.5	−4.5		1200	5

F, focal length of the lens (mm); $F/\#$, F-number (ratio F/D); L, total length of scan line (mm); ROAL, overall length from the entrance pupil to the image plane relative to the focal length (mm/mm); RFWD, front working distance relative to the focal length F; d, diameter of the image of a point at the $1/e^2$ irradiance level (µm); L/d, number of spots of a scan line; RBcr, paraxial chief ray bending relative to the input half angle; RPR, ratio of the Petzval radius to the lens focal length; REPR, estimated Petzval radius relative to the lens focal length from Equation 2.22; NS/I, number of spots per inch; and NO.el, number of lens elements.
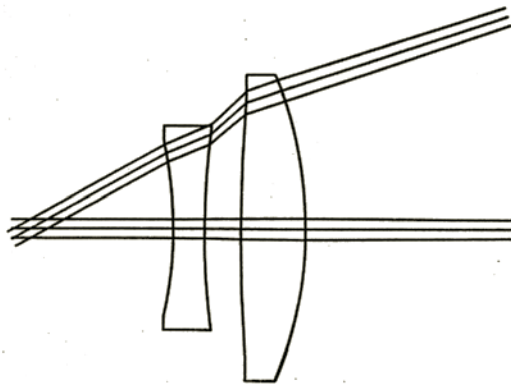
FIGURE 2.28
Lens 1: U.S. Patent 4,179,183 Tateoka, Minoura; F = 300 mm, F/60, L = 328 mm.

evaluated, and the results agree well with the patent. The focal length is given as 300 mm. The aberration curves appear to be given for the paraxial focal plane, and the linearity is shown to be within 0.6% over the scan. However, if one selects a calibrated focal length of 301.8 mm (11.8 in) and shifts the focus 2 mm (0.079 in) in back of the paraxial focus, it appears that the lens is well corrected to within $\lambda/4$ OPD (optical path difference) and linear to within 0.2%. A similar lens may have been used in early Canon laser printer engines. This lens has an exceedingly wide angle for a scan lens. It has a great advantage in the design of a compact scanner. This printer meets the needs of 300 DPI, which is quite satisfactory for high-quality typewriter printing of its time. The secret of the good performance of this lens is the airspace between the positive and negative lenses. There are strong refractions on the two inner surfaces of the lens, which means the airspace has to be held accurately and the lenses must be well centered.

2.9.2 Wide-Angle Scan Lens (λ = 633 nm)

The lens in Figure 2.29 has a 32° half-field angle. It has a careful balance of third-, fifth-, and seventh-order distortion, so that at the calibrated focal length it is corrected to be $F-\theta$ to within 0.2%. To do this the lens uses strong refraction on the fourth lens surface and refraction on the fifth surface to achieve the balance of the distortion curve. The airspace between these two surfaces controls the balance between the third- and fifth-order distortion. This would also be a relatively expensive lens to manufacture. The design may be found in U.S. Patent 4,269,478; it was designed by Haru Maeda and Yuko Kobayashi and assigned to Olympus Optical Co., Japan.

2.9.3 Semiwide Angle Scan Lens (λ = 633 nm)

The lens in Figure 2.30 shows how lowering the $F/\#$ to 20 and increasing the scan length increases the sizes of the lenses. This lens is a Melles Griot product designed by David Stephenson. It is capable of writing 1096 DPI and is linear to better than 25 μm. The large front element is 128 mm in diameter. It will transmit 2.8 times as many information points as the first lens (9.1). This lens requires modest manufacturing techniques, but, as shown in the diagram, the negative lens may be in contact with the adjacent positive lens. Either the airspace should be increased, or careful mounting has to be considered.
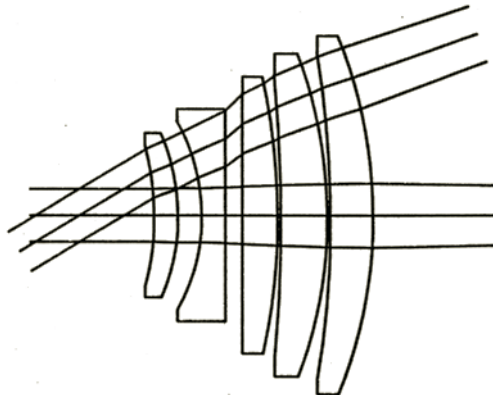
FIGURE 2.29
Lens 2: U.S. Patent 4,269,478 Maedo, Yuko; *F* = 100 mm, *F*/24, *L* = 118 mm.
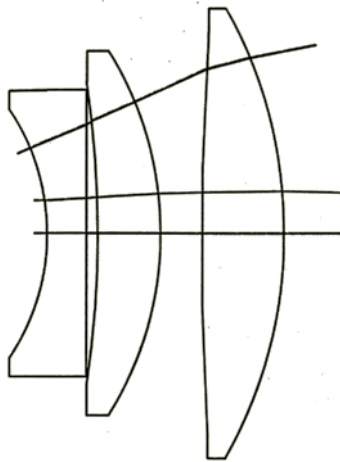
FIGURE 2.30
Lens 3: Melles Griot, Designer D. Stephenson; *F* = 400 mm, *F*/20, *L* = 300 mm.

2.9.4 Moderate Field Angle Lens with Long Scan Line (λ = 633 nm)

The lens in Figure 2.31 was designed by Robert E. Hopkins for a holographic scanner. It has a half scan angle of 18°, covers a 20 in scan length, and can write a total of 23,100 image points (about 1100 dpi). It has a short working distance between the holographic scan element and the first surface of the lens. This makes it more difficult to force the F–θ condition to remain within 0.1%. The working distance was kept short in order to keep the lens diameters as small as possible. The largest lens diameter is 110 mm. Lens performance could not be improved without making the elements considerably larger. Adding more lense elements does not help. The lens performs well in the design phase, but, since it is relatively fast for a scan lens, the small DOF makes the lens sensitive to manufacture and mounting. The only way to make the lens easier to build is to improve the Petzval field curvature, and this requires more separation between the positive and negative lenses. The
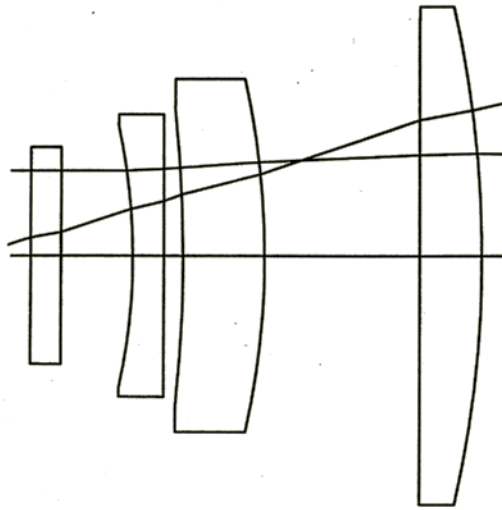
FIGURE 2.31
Lens 4: Designer R. Hopkins; F = 748 mm, F/17, L = 470 mm.

result is that the lens becomes longer and larger in diameter. Another way is to increase the distance between the scanner and the lens. This will also increase the lens size. We believe that this lens is close to the boundary of what can be done with a purely refracting lens for scanning a large number of image points. To extend the requirements will require larger lenses. It may be possible to combine small lenses with a large mirror close to the focal plane, but costs would have to be carefully considered. Photographic-quality shops can build this lens with attention to mounting.

2.9.5 Scan Lens for Light-Emitting Diode (λ = 800 nm)

In Figure 2.32 is another lens designed by Robert E. Hopkins to perform over the range of wavelengths from 770 to 830 nm. It was designed to meet a telecentricity tolerance of ±2°. The lens could not accommodate the full wavelength range without a slight focal shift. If the focal shift is provided for, the diodes may vary their wavelength from diode to diode over this wavelength region, and the lens will perform satisfactorily. This lens was not fully engineered for manufacture. It would be necessary to consider carefully how the negative–positive glass-to-glass contact combination would be mounted.

2.9.6 High-Precision Scan Lens Corrected for Two Wavelengths (λ = 1064 and 950 nm)

The Melles Griot lens in Figure 2.33 was designed for a galvanometer XY scanner system. It is capable of positioning a 4-µm spot anywhere within a 20-mm diameter circle. The spot is addressed with 1064-nm energy with simultaneous viewing of the object at 910–990 nm. The complexity of the design is primarily due to the need for the two-wavelength operation, especially the broad band around 946 nm. The thick cemented lenses require glass types with different dispersions. This lens requires precision fabrication and assembly to realize the full design potential.
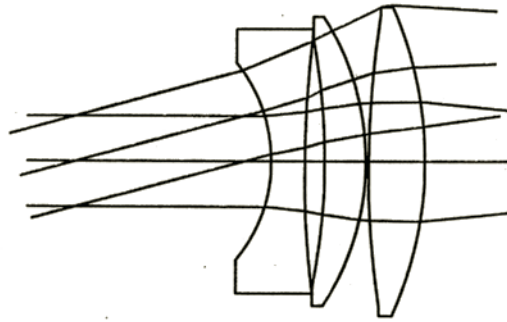
FIGURE 2.32
Lens 5: Designer R. Hopkins; F = 55 mm, F/5, L = 29 mm.
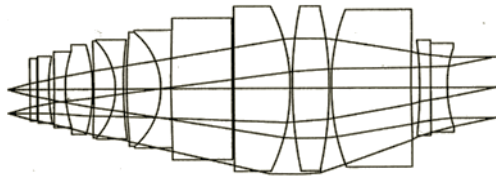
FIGURE 2.33
Lens 6: Designer D. Stephenson; F = 50 mm, F/2, L = 20 mm.
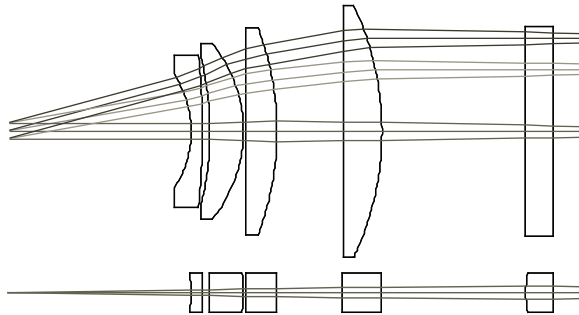
FIGURE 2.34
Lens 7: Designer S. Sagan; F = 125 mm, F/24, L = 70 mm.

2.9.7 High-Resolution Telecentric Scan Lens (λ = 408 nm)

The lens in Figure 2.34 was designed for a violet laser diode to image 1200 DPI over a small telecentric field. The design comprises three spherical elements and two cross cylinders to provide optical cross-scan correction and a precise telecentric field well corrected for F–Θ distortion. This lens form resembles Lens 9.5 in Figure 2.32, before the addition of the cross-scan correction. The careful selection of optical glasses for the violet and UV is particularly important for transmission and dispersion. Many of the new eco-friendly glasses can have significant absorption around 400 nm and below. For example, the internal transmittance

at 400 nm for a 10-mm thick Schott SF4 element is 0.954, while the new glass, N-SF4 internal transmittance is 0.79.

2.10 SCAN LENS MANUFACTURING, QUALITY CONTROL, AND FINAL TESTING

The first two designs shown in the previous section require tolerances that are similar to quality photographic lenses. Compared to the lens diameters, the beam sizes are small, and so surface quality is generally not difficult to meet unless scan linearity and distortion are critical performance parameters. Designs 2 and 3 do not require the highest quality precision lenses, but will require an acceptable level of assembly precision. Because the scan line uses only one cross-sectional sweep across the lenses, the yield of acceptable lenses can be improved by rotating the lenses in their cell, avoiding defects in the individual elements by finding the best line of scan. However, this requires that each lens be appropriately mounted into the scanning system, highlighting the need for good communications between the assemblers and the lens builders.

Lenses 9.5, 9.6, and 9.7 require lens fabricators capable of a precision build to achieve the expected performance of design. Surfaces need better than quarter-wave surface quality, and must be precision centered and mounted. Precision equipment will be required to maintain quality control through the many steps to fabricate and mount such lenses.

Precision lenses need special equipment for testing the performance of the finished lens assembly, using appropriate null tests and/or scanning of the image with a detector to measure the spot diameter in the focal plane, but the challenge is finding that plane. Attention must also be paid to the straightness of the detector measurement plane and motion relative to the lens axis. Typically the collimated beam should be directed into the lens exactly the way it will be introduced into the final scanning system. If the image beam intersects the image plane at an angle, the entire beam should pass through the scanner and to the detector. If the detector needs to be rotated relative to the image plane, a correction of the as-measured spot size to an as-used spot size has to be made. If the image is relayed with a microscope objective, the NA should be large enough to collect all the image cone angles.

Because scan lenses are usually designed to form diffraction-limited images, it is recommended that the lenses be tested using a laser beam with uniform intensity across the design aperture of the lens. The image of the point source should be diffraction-limited and its expected dimensions predictable. The image can be viewed visually or measured for profile and diameters via scanning spot or slit. Departures from a spherical wavefront as small as a one-tenth of a wavelength are easily detected. It is also possible to detect the effects of excessive scattered light.

2.11 HOLOGRAPHIC LASER SCANNING SYSTEMS

Holographic scanning systems were first developed in the late 1960s, in part through government-sponsored research for image scanning (reading) of high-resolution aerial photographs. Its application to image scanning and printing for high-resolution business

graphics followed in the 1970s with efforts dominated by IBM and Xerox.[16] Subsequent developments in both the holographic process (design and fabrication) and laser technology (from commercializing of the HeNe laser to low-cost diode lasers) have helped broaden its application into commercial and industrial systems.

Applications include: low-resolution point-of-sale barcode scanners; precision noncontact dimensional measurement, inspection and control of the production of high-tech optical fibers, medical extrusions and electrical cables; medium resolution (300–600 DPI) desk-top printers; and high-resolution (1200 DPI and up) direct-to-press marking engines.

Advantages of holographic scanners over traditional polygon mirror scanners include lower mass, less windage, and reduced sensitivity to scan disc errors such as jitter and wobble. Advancements in replication methods to fabricate scanning discs with surface holograms have also helped lower cost.

Disadvantages of holographic scanners include limited operating spectral bandwidth, deviation from classical optical design methods, and the introduction of cross-scan errors in simple design configurations (using a simple grating at compound angles). These issues must be managed during the design process through the configuration of the optical design and its ability to balance the image aberrations.

This section deals with the optical design of holographic scanning systems, beginning with the basics of a rotating holographic scanner and then developing its use with other components into more complex systems.[17]

2.11.1 Scanning with a Plane Linear Grating

A simple scanning system in terms of both the complexity of the holographic optical element (HOE) and its configuration is illustrated in Figure 2.35.[18] This system comprises a collimated laser beam incident on a hologram (nonplane grating the product of the interference of beams in a two-point construction). The performance characteristics of such a rudimentary scanner (line straightness, length, scan linearity, etc.) will depend on the angle of incidence at the hologram and the deviation by the hologram. These performance
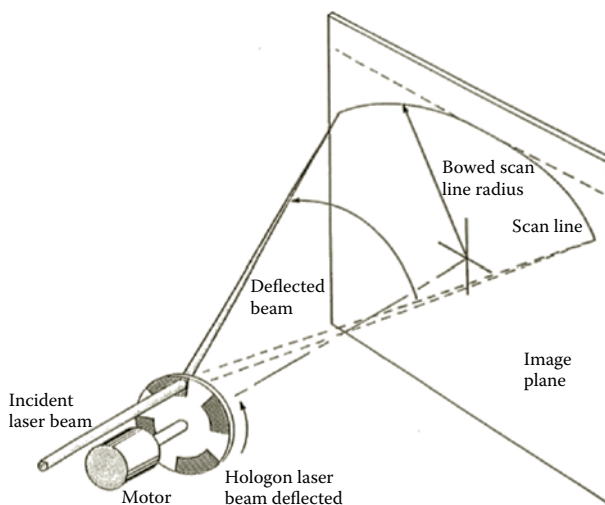
FIGURE 2.35
Cindrich-type holographic scanning. (From Kramer, C.J. Holographic deflector for graphic arts system. In *Optical Scanning*; Marshall, G.F., Ed.; Marcel Dekker: New York, 1991; 240.)

characteristics can be best understood through modeling of a simple plane linear diffraction grating (PLDG) rotated to generate the scan, the PLDG being equivalent to a hologram constructed from two collimated beams.

Deviation of the beam by the grating (for a grating perpendicular to a plane containing the incident beam and scan disc rotation axis) is the sum of the input incidence angle θ_i and output exit angle θ_0, as illustrated in Figure 2.36. These angles are derived from the grating equation[23] for a given grating or hologram fringe spacing d, the wavelength of the light λ_0, and the diffraction order m,

$$\sin q_i + \sin q_0 = \frac{m \lambda_0}{d}. \tag{2.27}$$

As with other types of scanning systems, errors such as line bow, scan disc wobble, eccentricity, axis longitudinal vibration, and disc tilt and wedge can affect scanned image position.

2.11.2 Line Bow and Scan Linearity

The typical purpose of a laser scanning system is to generate a straight line of points by moving a focused beam across a focal plane at a linear rate relative to the scanner rotation. A method for deriving a nearly straight line scan from a PLDG in a disc-like configuration was developed independently by C. J. Kramer in the United States and M. V. Antipin and N. G. Kiselev in the former Soviet Union.[16]

The scanning configurations they developed operate at the Bragg condition where the nominal input angle θ_i and output angle θ_0 are nearly 45°. The Bragg condition is where the input beam and diffracted output beam are at equal angles relative to the diffracting surface. Operating near this Bragg condition minimizes the effects of scan disc wobble and operating near 45° minimizes line bow (cross-scan departure from straightness). For the first diffraction order $m = 1$, the grating or hologram fringe spacing d would be given by the reduced equation

$$d = \frac{\lambda_0}{\sqrt{2}}. \tag{2.28}$$

The actual optimum angle will depend on the scan length and the degree of line bow correction desired. The dependence of the scan line bow and scan linearity (in-scan position
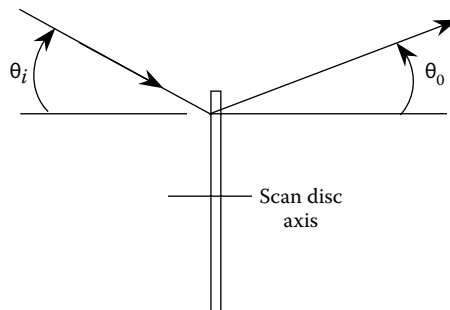
FIGURE 2.36
Scan disc input and output angles.

error relative to the scan angle) on the Bragg angle is illustrated in Figures 2.37a and b for a design wavelength of 786 nm. In a 45° monochromatic corrected configuration, the chromatic variation of the line bow and scan length for a ±1-nm change in wavelength results in the departures illustrated in Figures 2.38a and b, respectively. These errors (line bow, scan linearity, and their chromatic variation) are significant compared to the resolution of the printer system and must be balanced in the designs to allow the use of cost-effective laser diodes with diffractive optical components.

2.11.3 Effect of Scan Disc Wobble

In general, HOEs are best designed to operate in the Bragg regime to minimize the effects of scan disc wobble. Wobble is the random tilt of the scan disc rotation axis due to bearing errors. The error in the diffraction angle ε resulting from a wobble angle δ indicated in
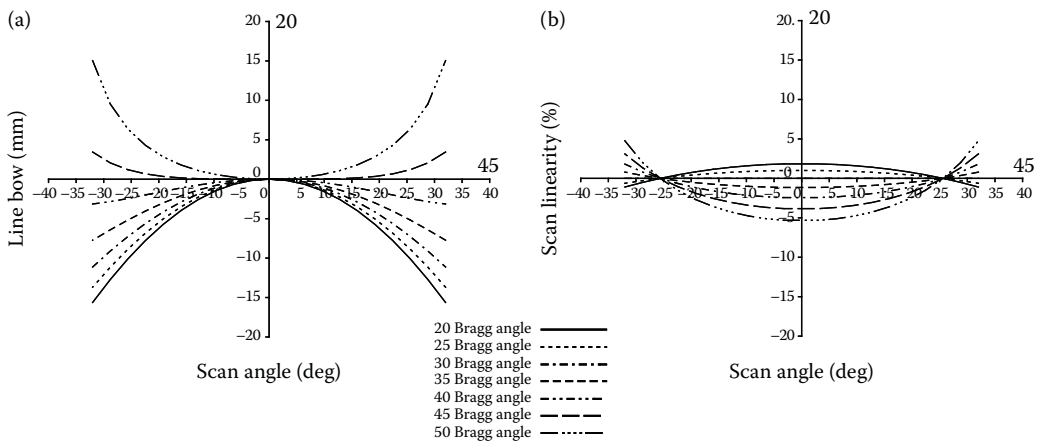
FIGURE 2.37
Effect of Bragg angle on line bow (a) and scan linearity (b).
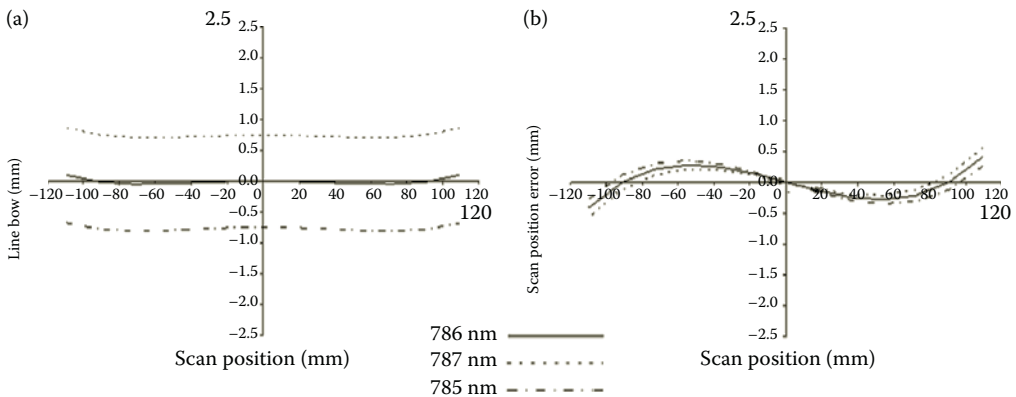
FIGURE 2.38
Chromatic variation of line bow (a) and scan length (b).
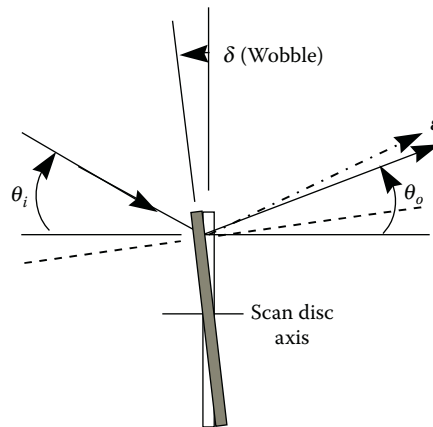
FIGURE 2.39
Scan disc wobble and beam deviation.

Figure 2.39 is given by the modified grating equation

$$e = \arcsin\left[\frac{m l_0}{d} - \sin(q_i + d)\right] + d - q_0. \tag{2.29}$$

This angular error produces a cross-scan displacement in the scanned beam, affecting the position of a measurement being made or the position of a point being written. That displacement error is a product of the angle error and the focal length of the projection optics. In systems with traditional reflective scanners (such as a polygon), this angular error is twice the wobble tilt error unless it is optically compensated using anamorphic optical methods, which greatly increases complexity and cost. Scan jitter in the scanner rotation will likewise generate twice the in-scan position errors. The precision of the scan degrades as the motor bearings wear and mirror wobble increases.

In a holographic scanning system, the cross-scan image error can be minimized to hundredths or even thousandths of the disc wobble error and the in-scan image error is approximately equal to the disc rotation (jitter) error. In a classical polygon scanning system the image errors are twice the wobble and jitter errors. The effect on output angle error of a tilt in the scan disc rotation axis is shown in the curves of Figure 2.40 derived from the modified grating Equation 2.29. Two Bragg angle and two near-Bragg angle design configurations operating near the minimum line bow condition of 45° and arbitrary angle of 22° are plotted.

2.12 NONCONTACT DIMENSIONAL MEASUREMENT SYSTEM USING HOLOGRAPHIC SCANNING

Optoelectronic systems for noncontact dimensional measurements have been available to industry since the early 1970s. Major applications include the production and precision inspection of linear products like wire, cable, hose, tubing, optical fiber, and metal, plastic and rubber extruded shapes.
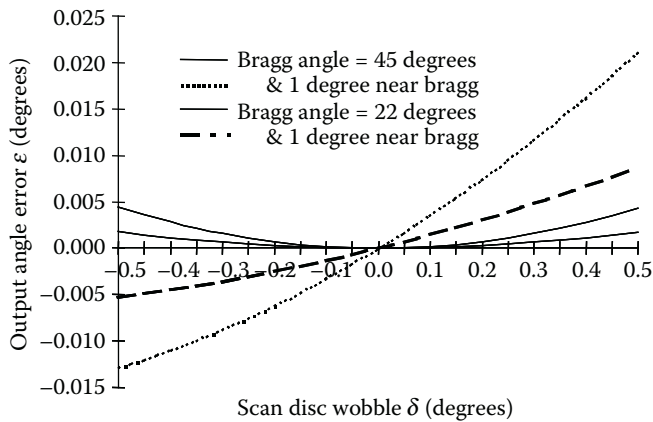
FIGURE 2.40
Effect of input angle and disc wobble on output scan angle.

Linear measurement instruments have used two basic technologies to perform the measurement. The first systems used a linearly scanned laser beam and a collection system with a single photodetector. As the telecentric beam moves across the measuring zone, the object to be measured blocks the laser beam from the collection system for a period of time. By knowing the speed of the scan and the time the beam is blocked, the dimension of the object can be calculated by a microprocessor and then displayed. Later, as video array technology developed, systems using a collimated light source and a linear charge coupled device (CCD) or photodiode array were introduced. These systems use an incandescent lamp or an LED with a collimating lens to produce a highly collimated beam of light that shines across the measuring zone. The object to be measured creates a shadow that is cast on a linear array of light-sensing elements. The count of dark elements is scaled and then displayed.

In most scanning systems, a laser beam is scanned by reflecting the beam off a motor-driven polygon mirror. The precision of the scan degrades as the motor bearings wear, increasing mirror wobble and scan jitter. Otherwise, the overall accuracy of a laser scanning system can be quite good, because resolution is based on a measurement of time, which can be made very precisely.

The electronic interface to a scanning dimensional measurement system is well understood and has been used in various applications for years. Assuming that the scanning beam travels through the measuring zone at a constant, nonvarying speed, improvements in system performance are limited to (1) maximizing the reference clock speed; (2) further dividing the reference clock speed by delay lines, capacitive charging, or other electronic techniques; and (3) minimizing the beam on/off detection error.

Video array systems that use incandescent lamps drive the lamp very hard, which produces substantial heat and reduces the lamp life. Systems that use a solid-state source, such as an LED, are very efficient and require much less power. The quality of the shadow image depends on the entire optical system. Relatively large apertures are required to collect a sufficient number of photons to achieve the needed signal-to-noise ratio. The physical dimensions and element size in the video array limit the resolution and achievable accuracy of the system. System reliability, however, is typically high, and the mean time between failures can be long if a solid-state light source is used. An LED/video array system is all solid state and has no moving parts to wear out.

The application of holographic scanning to noncontact dimensional measuring systems provides the opportunity to take advantage of the positive points in scanning systems and the high reliability of the all-solid-state video array systems, while avoiding some of the problems found in each.

A polygon mirror is manufactured one facet at a time, but a holographic disc can be replicated, like a compact disc or CD-ROM. A holographic disc can be produced with 20 to 30 facets at a fraction of the cost of producing a comparably high-precision polygon mirror with as many facets. Furthermore, other holographic components that might be used in the system, such as a prescan hologram, can also be replicated.

2.12.1 Speed, Accuracy, and Reliability Issues

As line speeds have increased and tolerances have narrowed, the need to provide not only diameter measurements but also flaw detection has grown dramatically. To provide fast response in process control and surface flaw detection, a measurement system must make many scans per second. The number of scans per second that can be made by a measurement system is determined by the number of facets, the speed of the motor, and the data rate capability of the analog to digital (A/D) converter. With the high-speed electronics that are available today, the motor speed and thus motor lifetime and cost are the limiting factors when designing high-speed measurement devices. Typically, polygon mirror scanners have one-third the number of facets (or fewer) than a holographic disc, so the motor must run three times faster (or more) to produce the same number of scans per second as the holographic scanner. Consequently, the trade-off between speed and lifetime/cost of the motor for holographic disc scanners is considerably more attractive than polygon mirror scanners.

Traditional laser-based systems scan from 200–600 times/s/axis and provide limited single-scan information. The holographic disc in the Holix® Gage by Target Systems Inc., Salt Lake City, UT, has 22 segments as compared to 2–8 sides on a typical polygon mirror, with single-scan-based flaw detection possible at 2833 scans per second per axis.[19]

The ability to scan faster means more diameter measurements can be made over a given length of the test object in a given time interval. With this resolution increase, the system is more likely to detect surface flaws, as illustrated in Figure 2.41. In these holographic scanning systems there is no need to average groups of scans to compensate for the surface irregularities in a manufactured polygon mirror. This is a major advancement in the ability to detect small surface flaws that can be missed by traditional gages.

In reflective scanners, any tilt error in the mirror can cause twice the error in the output beam. In transmission holographic scanners, the beam is diffracted rather than reflected, and the beam error on output is much smaller than the tilt of the disc axis.

37 Scans @ 1833 scans/sec.   5 Scans @ 240 scans/sec.
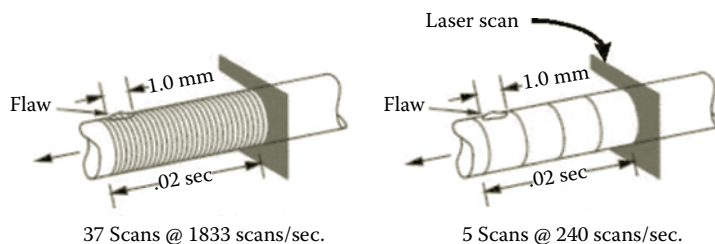
FIGURE 2.41
Detection of surface flaws with multiple scans.

As the number of facets increase on a polygon mirror, another problem can develop. Because of the finite size of the laser beam and the required scan angle, there is a minimum size for each facet on the mirror. As the size of each facet is increased, the distance from the center of rotation to the facet surface increases (the polygon mirror gets larger in diameter). As the facet surface moves farther from the center of rotation, the virtual location of the scan center point shifts during the scan (pupil shift). This means the scan center point is not maintained at the focal point of the scan lens (which is designed to generate a telecentric scan) for the entire duration of the scan. The telecentricity errors from this shift in the pupil reduce the measurement accuracy or the depth of the measurement zone. Holographic disc scanners have no pupil shift, and therefore, are not limited by these errors.

2.12.2 Optical System Configuration

The optical system illustrated in Figure 2.42 offers scan and laser spot size performance in the measurement zone that, together with a proprietary processing algorithm, can provide repeatable measurements to within one micro-inch. The optical design comprises only the basic components required to scan a line: a laser diode, a collimating lens, a prescan HOE, a scanning HOE, a parabolic mirror (the scan lens in this system), a collecting lens, and fold mirrors to provide the desired packaging.

The laser diode is mounted in a metal block that is thermally isolated from the instrument frame. A thermoelectric cooler (TEC) or heater and temperature controller are used to maintain the operating temperature of the laser diode over a narrow range of ~0.5 °C. Temperature control is required to prevent "mode hops" in the laser diode as the temperature changes. The wavelength shift of a mode hop can be 0.5–1 nm with temperature change, as illustrated in Figure 2.43. The temperature control setpoint is selected to center the laser diode between mode hops and prevent the diode from changing modes and abruptly shifting the emission wavelength. This is a critical feature of the system because diffractive elements are very sensitive to changes in wavelength. Although corrected for a wavelength drift in the ±0.1-nm class, like many other HOE-based optical systems, it cannot accommodate mode hops.

The diverging beam emitted from the laser diode is apertured and quasi-collimated by a typical laser diode collimator. The ratio of the collimator focal length to the parabolic mirror
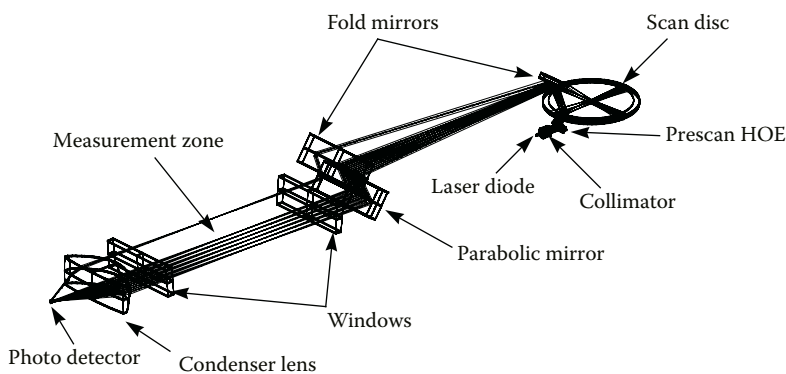
FIGURE 2.42
Single axis HOE optical system configuration.
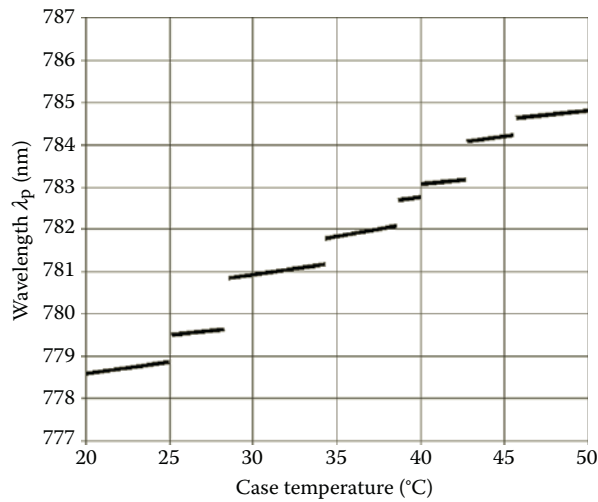
FIGURE 2.43
Laser diode characteristic wavelength versus temperature dependence (SHARP 5 mW, 780 nm, 11·29° divergence).

focal length sets the magnification for the projected laser spot width in the measurement zone. The beam is then diffracted by a stationary prescan HOE and diffracted again by a rotating scan disc HOE. The first-order diffracted beam exits the scan HOE at an angle about 22° off normal. As the disc rotates, the orientation of the diffractive structure in the scan HOE changes, causing the beam to scan from side to side. The ratio between disc rotation angle and scan angle is approximately 1:1. The zero-order beam is not diffracted by the HOE and continues straight to a photodiode that is used to monitor disc orientation.

The use of replication to manufacture production holographic discs shown in Figure 2.44 provides an economical way to reproduce discs with other features in addition to a large number of facets. Disc sectors are two-point construction or multiterm $(x, y, \ldots, x^n y^m)$ phase HOEs. All of the HOE component or surfaces can be produced by photo resist on glass substrates, embossing (replication) using polymers, or injection/compression molding in plastics.

The discs are designed with a gap between two adjacent facets that is much narrower than a holographic facet, and that has no grating on it. When the gap rotates over the laser beam, the beam is not diffracted and continues straight to a photodiode. This signal is used to identify the orientation of the disc. The data processor can use this synchronization pulse to associate a unique set of calibration coefficients with each facet. This allows variations in optical parameters from one facet to another to be calibrated out.

The diffracted first-order beam is folded by two mirrors, reflected by the scan mirror, and then continues out through a window to the measurement zone. The parabolic mirror serves two purposes. First, it provides a nearly telecentric scan. Secondly, it focuses the beam so that the minimum in-scan waist is in the center of the measurement zone. After the measurement zone, the beam passes through a second window and is collected by a condenser lens that focuses it onto a single photodetector. The collecting lens is an off-the-shelf condenser lens with an aspheric surface to minimize the spherical aberration generated by the nearly $F/1$ operating condition. The ratio of the condenser to scan lens focal length determines the size of the beam (an image of the beam at the scan disc) on the photodetector.
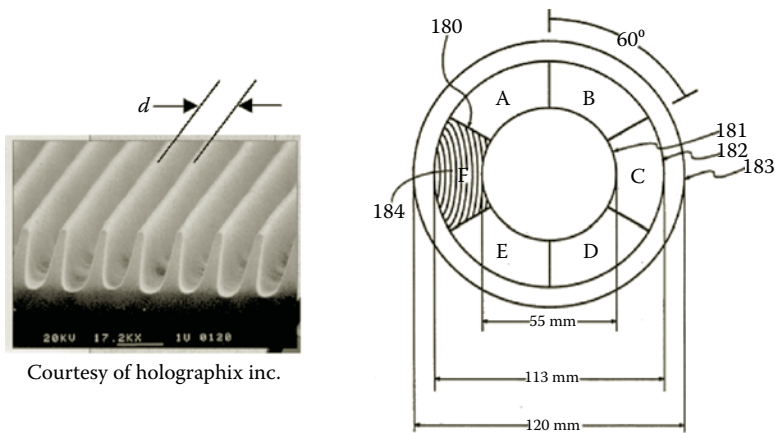
FIGURE 2.44
Replicated holographic scan disc. (Courtesy of Holographix, Inc.).

The inherent chromatic variations of scan length and scan bow in a holographic scanner are corrected to acceptable levels by the Holographix patented configuration of Figure 2.45.[20] The HOE scan disc introduces both line bow and chromatic aberrations (in- and cross-scan). The prescan HOE is used to introduce additional cross-scan chromatic error, which, when coupled with the bow correction provided by the tilted curved mirror (a rotationally symmetric parabola for the noncontact dimensional measurement system), produces a chromatically corrected in-scan beam. The cross-scan corrector hologram in the noncontact dimensional measurement system can be eliminated at the expense of cross-scan chromatic correction, illustrated in Figure 2.46. However, the magnitude of the cross-scan error is less than 100 microns, and does not affect the diameter measurement.

The foregoing description is of a single measurement axis instrument. For many applications a single instrument with two measurement axes is used. This configuration includes two lasers and collimating systems that use a single scan disc at separate locations (at 90°) on the radius. The resulting two scanning beams continue and pass through separate mirrors and optics to produce two orthogonal telecentric scans. Separate collection lenses and photodetectors complete the optical paths. This system, with two measuring axes, is used to measure the nonroundness of round objects, two dimensions of nonround objects, and to view the object's surface from more directions for more complete surface flaw detection. The discs are designed with facets of a specific size, so systems that use a single disc to produce multiple scan axes will be compatible with the orientation of the optical paths.

2.12.3 Optical Performance

The single measurement axis instrument can be optimized to provide a relatively large measurement zone while maintaining telecentricity and a consistent beam size in the scan direction. The profile and consistency of the in-scan beam width are shown in Figure 2.47 for a 50-mm gage system over a ±20-mm measurement zone. A key to providing this performance is orientation of the laser diode astigmatism with the slow axis parallel to the scan, also optimizing scan efficiency. The $1/e^2$ in-scan beam width over the measurement zone is approximately 210 microns, with centroid stability better than one micron and a
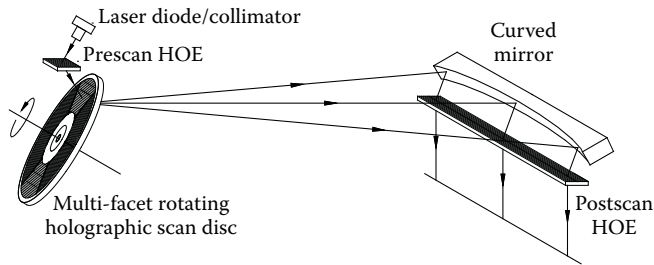
FIGURE 2.45
Holographix, Inc., U.S. patent 5,182,659.
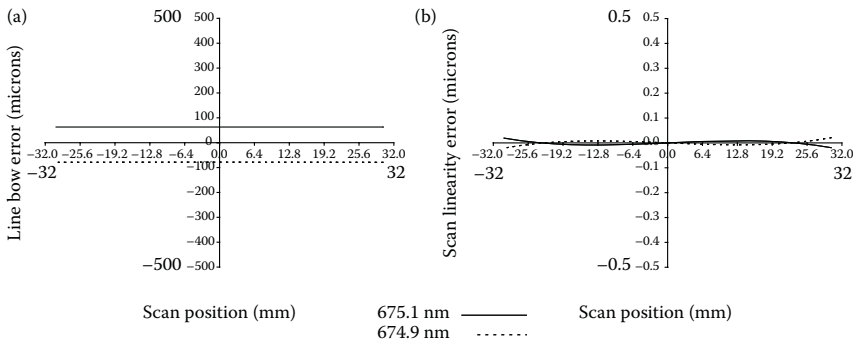
FIGURE 2.46
Chromatic variations of line bow (a) and scan position (b).
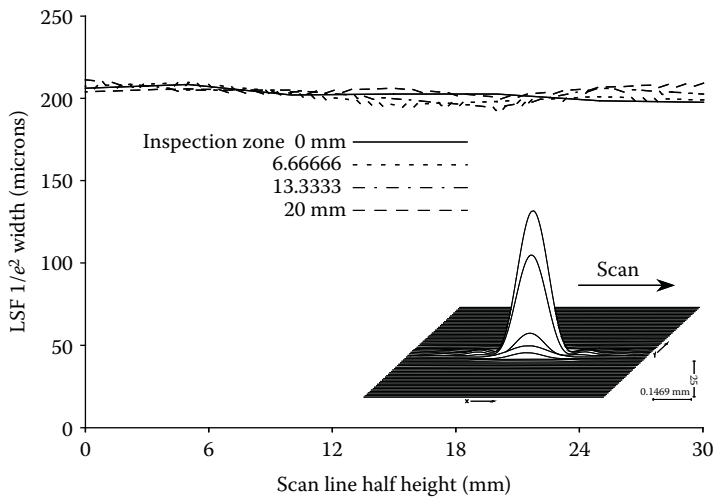
FIGURE 2.47
Line-spread function-based in-scan spot width.

monochromatic telecentricity better than 0.04 mrad (as-designed). The profile of the elliptical spot is Gaussian with small diffraction side lobes that are the result of truncation by the collimating lens aperture. The cross-scan spot size is typically much smaller, but is not controlled except by the first-order configuration. The design is optimized for an operating wavelength of 675 ± 0.1 nm anywhere within a setpoint of 670–680 nm. The subtle variations in the spot width are static for each system. Also static is the scan nonlinearity (distortion). These static variations are calibrated over the inspection zone to provide measurement repeatability of 30 micro-inches for the 50 mm and 3 micro-inches for the 7 mm Holix gages.

2.13 HOLOGRAPHIC LASER PRINTING SYSTEMS

The configuration of a lower performance scanning system based on the Holographix patent is shown in Figure 2.48. This design is for a 300 DPI system comprising a laser diode source, a collimating lens, a prescan HOE, a holographic scan disc, a tilted concave cylindrical mirror, a postscan HOE, and assorted fold mirrors for packaging. This configuration provides a method for balancing errors of line bow, scan linearity, and their significant chromatic variations to the resolution requirements of the printer system, allowing the use of cost-effective laser diodes coupled with diffractive optical components.

This configuration uses a scan disc HOE at other than the minimum line bow condition (to introduce a prescribed amount of line bow) that both focuses and scans the beam. A prescan HOE is used to introduce additional chromatic cross-scan error that, coupled with the bow correction provided by a tilted curved mirror, produces a corrected in-scan beam. The postscan HOE completes the correction by balancing the cross-scan compounded errors and focusing the beam. The unique configuration yields a system with significantly and acceptably reduced sensitivity to changes in laser wavelength (due to mode hops during scan, wavelength drift over temperature, and variation from diode to diode).

Typical system performance specifications (beam size, line bow, scan linearity, and change in line bow and scan linearity over wavelength and scan) are listed in Table 2.7. As the system's performance requirements increase, tighter control of the system aberrations (particularly field curvature and astigmatism) is required. Additional design variables
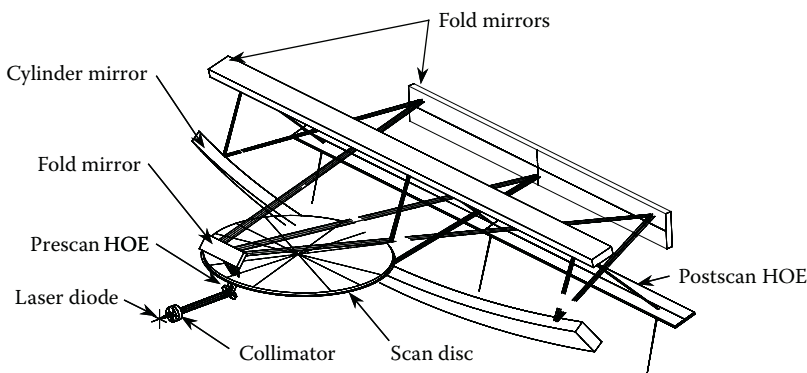
FIGURE 2.48
Holographix laser printer optical system.

TABLE 2.7

Typical Performance Specifications for Holographic Laser Printer

Parameter	Specification
1. System configuration	• Commercial laser diode
	• Collimator
	• Prescan HOE
	• 65-mm scan disc with HOE
	• Cylindrical mirror
	• Postscan HOE
2. Laser diode	
(i) Center wavelength	670, 780, 786 (nm)
(ii) Wavelength drift	±1 nm
(iii) FWHM divergence	11H · 29V degrees
(iv) Astigmatism	7 microns
3. Wavelength accommodation	±10 nm
4. Beam diameter at focal plane (nominal, best focus)	300 DPI:$(1/e^2)$
	In-scan: 80 + 10 microns
	Cross-scan: 100 + 20 microns
	600 DPI: $(1/e^2)$
	In-scan: 50 ± 10 microns
	Cross-scan: 50 ± 10 microns
	1200 DPI: (FWHM)
	In-scan: 20 ± 5 microns
	Cross-scan: 25 ± 5 microns
5. Scan line	
(i) Total length	300/600 DPI: 216 mm
	1200 DPI: 230 mm
(ii) Linearity (w.r.t. scan disc rotation)	300/600 DPI: ±1%
	1200 DPI: ±0.03%
(iii) Chromatic variation of line length (over ± 1 nm)	300 DPI: <20 microns
	600 DPI: <5 microns
	1200 DPI: <5 microns
(iv) Bowing (microns)	300/600 DPI: <300 microns
	1200 DPI: <25 microns
(v) Chromatic variation of line bow (over ± 1 nm)	300 DPI: <20 microns
	600 DPI: <10 microns
	1200 DPI: <5 microns
(vi) Telecentricity	<4°

such as higher-order terms on the scan disc and postscan HOEs provide the degrees of freedom necessary to achieve the line bow, linearity, beam quality, and chromatic variations specifications. These laser scanning system designs are nearly telecentric in image space to avoid position errors over the in-use DOF.

Since the original 300 DPI systems were developed, these designs have been continually refined and improved. The latest 600 DPI designs actually have a larger DOF, smaller package, reduced material cost, increased ease of assembly and alignment, and better chromatic

correction. These benefits have been obtained using an optimization process that increases the complexity of the recording parameters of the HOEs while decreasing the complexity of the rest of the system.

Over the years, Holographix has developed proprietary alignment and recording techniques to fabricate complex HOEs for both transmission and reflection. As complexity has increased, so has the cost of recording of the "master" HOE. However, with the development of HOE mastering and replication techniques in which several thousand replicas can be made from one master, the increased cost of the masters becomes insignificant. Development of holographic scanning systems has included refinement of production/replication processes to achieve higher diffraction efficiencies at lower cost. Diffraction efficiencies of replicated HOEs averaging 80% provide for high system throughput.

The optical design and tolerancing of systems with HOEs is nontrivial. Optimization of the optical system requires controlling (and sometimes limiting) the HOE degrees of freedom to minimize a feedback interaction between HOEs. The tolerancing of the optical designs for as-built performance requires the tolerancing of each HOE construction setup and then introducing the as-built HOEs into the tolerancing of the optical design. This multiple configuration/layer tolerancing can be modeled in Code V to predict the as-built performance of these complicated systems.

2.14 CLOSING COMMENTS

The significant trends in refinement of laser scanning systems are ever-increasing scan lengths and larger number of spots per inch. The design examples show present-day practical boundaries for scan lenses. These boundaries continue to slowly expand with new concepts and ever-increasing precision, using more combinations of refractive lenses, diffractive elements, and mirrors with anamorphic power—both nearer the scanner and closer to the image plane. Methods for economically manufacturing these elements continue to be developed, allowing for larger diameter refracting lenses and reduced cost lens segments and mirrors.

If not limited by the optical invariant, optical systems are generally limited by the precision of lens fabrication, assembly, alignment, and testing. Availability of fabrication houses with special equipment for unconventional surface types such as cylinders and toroids, in combination with spherical and aspheric surfaces, is limited. As usual, the market to pay for the special equipment and tooling has to be large enough to support the investment.

ACKNOWLEDGMENTS

REFERENCES

1. Hopkins, R.E.; Stephenson, D. Optical systems for laser scanners. In *Optical Scanning*; Marshall, G.F., Ed.; Marcel Dekker: New York, 1991; 27–81.
2. Beiser, L. *Unified Optical Scanning Technology*; John Wiley & Sons: New York, 2003.
3. Siegman, A.E. *Lasers*; University Science Books: Mill Valley, California, 1986.
4. Melles Griot. *Laser Scan Lens Guide*; Melles Griot: Rochester, NY, 1987.
5. Fleischer; Latta; Rabedeau. *IBM Journal of Research and Development,* 1977, *21*(5), 479.
6. Hopkins, R.E.; Hanau, R. *MIL-HDBK-141*; *Defense Supply Agency*: Washington, DC, 1962.
7. Kingslake, R. *Optical System Design*; Academic Press: New York, 1983.
8. Hopkins, R.E.; Buzawa, M.J. *Optics for Laser Scanning*, SPIE 1976, *15*(2), 123.
9. Kingslake, R. *Lens Design Fundamentals*; Academic Press: New York, 1978.
10. Smith, W.J. *Modern Optical Engineering*; McGraw-Hill: New York, 1966.
11. Welford, W.T. *Aberrations of Symmetrical Optical Systems*; Academic Press: London, 1974.
12. Levi, L. *Applied Optics, A Guide to Optical System Design/Volume 1*; John Wiley & Sons: New York, 1968; 419.
13. Marshall, G.F. Center-of-scan locus of an oscillating or rotating mirror. *In Recording Systems: High-Resolution Cameras and Recording Devices and Laser Scanning and Recording Systems*, Proc. SPIE Vol. 1987; Beiser, L., Lenz, R.K., Eds.; 1987; 221–232.
14. Marshall, G.F. Geometrical determination of the positional relationship between the incident beam, the scan-axis, and the rotation axis of a prismatic polygonal scanner. In *Optical Scanning 2002*, SPIE Proc. Vol. 4773; Sagan, S., Marshall, G., Beiser, L., Eds.; 2002; 38–51.
15. Sagan, S.F. *Optical Design for Scanning Systems*; SPIE Short Course SC33, February 1997.
16. Beiser, L. *Holographic Scanning*; John Wiley & Sons: New York, 1988.
17. Sagan, S.F.; Rowe, D.M. Holographic laser imaging systems. SPIE Proceedings 1995, *2383*, 398.
18. Kramer, C.J. Holographic deflector for graphic arts system. In *Optical Scanning*; Marshall, G.F., Ed.; Marcel Dekker: New York, 1991; 240.
19. Sagan, S.F.; Rosso, R.S.; Rowe, D.M. Non-contact dimensional measurement system using holographic scanning. *Proceedings of SPIE*, 1997, *3131*, 224–231.
20. Clay, B.R.; Rowe, D.M. Holographic Recording and Scanning System and Method. U.S. Patent 5,182,659, January 26, 1993.
21. Wetherell, W.B. The calculation of image quality. In *Applied Optics and Optical Engineering*; Academic Press: New York, 1980; Vol. 7.
22. Thompson, K.P. *Methods for Optical Design and Analysis—Seminar Notes*; Optical Research Associates: California, 1993.
23. O'Shea, D.C. *Elements of Modern Optical Design*; John Wiley & Sons: New York, 1985; 277.

3

Image Quality for Scanning and Digital Imaging Systems

Donald R. Lehmbeck

Xerox Corporation, Webster, New York, USA (retired)
College of Imaging Arts and Sciences, Rochester Institute of Technology,
Rochester, New York, USA (Adjunct Faculty)
Imaging Quality Technology Consulting, Penfield, New York, USA
Torrey Pines Research, Fairport, New York, USA

John C. Urbach

Portola Valley, California, USA (deceased)

CONTENTS

3.1 INTRODUCTION

3.1.1 Imaging Science for Scanned Imaging Systems

This chapter presents some of the basic concepts of image quality and their application to scanned imaging systems. In this revised edition, we have added more on tonal rendition including system plots and halftones, new approximations to MTF's and revised pointers to the current industry standards in image quality, as well as more reference data and charts while reducing content on binary imaging and overall quality. New references and other technical details have been added throughout.

The emphasis in this chapter will be on the input scanner. Output scanners will be dealt with mainly by inference since many input scanner considerations and metrics are directly applicable to the rest of a complete electronic scanned imaging system. Expanded discussions of halftone methods, tone reproduction, and nonuniformity are examples of output scanners and systems implications. The chapter is organized as 10 major sections moving from the basic concepts and phenomena of image scanning and color, through practical aspects of image quality, to performance of input scanners that produce multilevel (gray) signals and then the special but common case of binary scanned images. This is followed by sections on very specific topics: various summary measures of imaging performance and specialized image processing. To assist the reader, psychophysical measurement methods used to evaluate image quality and some reference data and charts have been added.

3.1.1.1 Scope

We, like so many others, follow in the path pioneered over a half century ago by the classic 1934 paper of Mertz and Gray.[1] Without going into the full mathematical detail of that paper and many of its successors, we attempt to bring to bear some of the modern approaches that have been developed both in image quality assessment and in scanned image characterization. Many diverse technologies used in scanned imaging systems are addressed throughout the rest of this book. We cannot address the explicit effects of any of these on quality because they provide an enormous array of choices and trade-offs. Building on a more general foundation of imaging science, we shall attempt to provide a framework in which to sort out the many image quality engineering and technology issues that depend on these choices.

It is our intent not to show that one scanner or technique is better than another, but to describe the methods by which a scanning system can be evaluated to compare to other systems and to assess the technologies used in them. This chapter therefore deals primarily with such matters as the sharpness or graininess of an image and not with such hardware issues as the surface finish of an aluminum mirror, uniformity of a drive motor, or the efficiency of charge transfer in a charge-coupled device (CCD) imager.

Scanning is considered here in the general context of electronic imaging. An electronic imaging system can be considered as composed of 10 basic parts[2] illustrated in Figure 3.1 as a flow chart of fairly self explanatory terms. Both digital photography and scanning use the same type of CCD or CMOS sensors, that is, detectors. Both create images in two-dimensional pixel format. For both, the processor is on the sensor, in hardware resident on the system and also off-line in computers. Both systems generate two-dimensional prints or displays of images using one-dimensional output applying them to one-dimensionally electronic/computer stored bit streams. Both systems use optical systems and input radiation to create the captured image including arrays of color filters to create colored images.
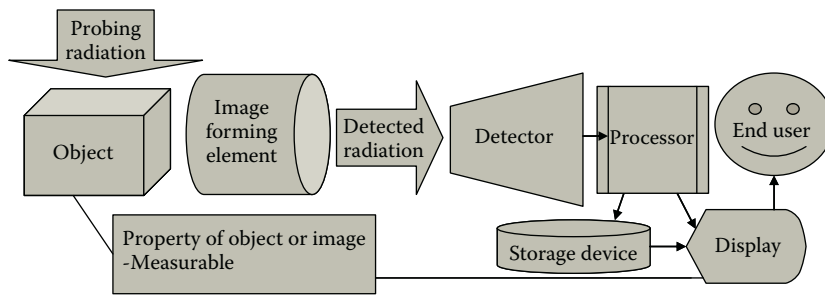
FIGURE 3.1
The fundamental elements of any imaging system arranged in a flow diagram that approximates a typical scanner or digital camera.

Some input scanners use reduction optics much as a camera in macromode but some use selfoc lens arrays which nearly contact the reflection original.

The primary difference between digital photography and input scanning is that the sensor in most photography is a fixed two-dimensional array of photosites (i.e., one-pixel sensors), while in scanning the array is synthesized by moving a long line of photosites one-pixel wide (i.e., a one-dimensional array or possibly three lines one for each color) over as much of a document as is needed. This has an effect in the scanning electronics—speed of the real-time circuits and optomechanical structures—that might create errors in positioning the line of sensors. This creates a difference from two-dimensional arrays making it appear as if the synthesized array was nonuniform ... Similar nonuniformities from one-dimensional moving arrays also occur in the printers which serve both scanners and other digital imaging devices. The nonuniformities are described in Sections 3.3.2 and 3.3.3 (pp. 167–174). Much of the chapter, therefore, applies equally well to digital photography and scanning.

3.1.1.2 The Literature

Considerable research, development, and engineering have occurred over the last two decades and only a very small portion could be referenced in the following pages. A few general references of note are provided as References 2–18 and elementary tutorials in References 19–23. Other more specific work of importance that may interest the reader includes: the vast technology of image processing,[17] many papers focused on specific problems in scanner image quality (see titles),[24–26] digital halftoning,[27,28] color imaging,[29–32] and various forms of image quality assessment.[33–39]

While the focus here is on imaging modules and imaging systems, scanners may, of course, be used for purposes other than imaging, such as digital data recording, from Bar Codes for example. We believe that the imaging science principles used here are sufficiently general to enable the reader with a different application of a scanning system to infer appropriate knowledge and techniques for these other applications.

3.1.1.3 Types of Scanners

All input scanners convert one- or (usually) two-dimensional image irradiance patterns into time-varying electrical signals. Image integrating and sampling systems, such as those found in many forms of electronic cameras and electronic copying devices, have sensors such as a CCD array. The signals produced by these scanners can be in one of two

general forms, either (a) binary output (a string of on and off pulses), or (b) gray-scale output (a series of electrical signals whose magnitude varies continuously).

The term *digital* here refers to a system in which each picture element (pixel) must occupy a discrete spatial location; an analog system is one in which a signal level varies continuously with time, without distinguishable boundaries between individual picture elements. A two-dimensional analog system is usually only analog in the more rapid direction of scanning and is discrete or "digital" in the slower direction, which is made up of individual raster lines. Television typically works in this fashion. In one form of solid-state scanner, the array of sensors is actually two-dimensional with no moving parts. Each individual detector is read out in a time sequence, progressing one raster line at a time within the two-dimensional matrix of sensors.

In other systems a solid-state device, arranged as a single row of photosites or sensors, is used to detect information one raster line at a time. In these systems either the original image is moved past the stationary sensor array, or the sensor array is scanned across the image to obtain information in the slow scan direction.

Cameras in digital photography employ totally digital solid-state two-dimensional sampling arrays. In some sense they represent commonly encountered forms of input scanners. The reader should be able to infer many things about the other forms of scanners and digital cameras from examples discussed in this chapter.

3.1.2 The Context for Scanned Image Quality Evaluation

Building blocks for developing a basic understanding of image quality in scanning systems are shown in Figure 3.2. The major elements of a generalized scanning system are on the left, with the evaluation and analysis components on the right. This chapter will deal with all of these elements and it is therefore necessary to see how they all interact.

The general configuration of scanning systems often requires two separate scanning elements. One is an input scanner to capture, as an electronic digital image, an input analog optical signal from an original scene (object), shown here as a hard copy input, such as a photograph. The second scanning element is an output scanner that converts a digital signal, either from the input scanner or from computer-generated or stored image data, into analog optical signals. These signals are rendered suitable for writing or recording on some radiation-sensitive medium to create a visible image, shown here as hard copy output. The properties of this visible image are the immediate focus of image quality analysis. It may be photographic, electrophotographic, or something created by a variety of unconventional imaging processes. The output scanner and recording process may also be replaced by a direct marking device, such as a thermal, electrographic, or ink jet printer, which contains no optical scanning technology and therefore lies outside the scope of this volume. Nonetheless, its final image is also subject to the same quality considerations that we treat here.

It is to be noted that the quality of the output image is affected by several intermediate steps of image processing. Some of these are associated with correcting for the input scanner or the input original, while others are associated with the output scanner and output writing process. These are mentioned briefly throughout, with the digital halftoning process, described in Section 3.2.2.3, cited as a major example of a correction for the output writing. Losses or improvements associated with some forms of data communication, and compression are very important in a practical sense, especially for color. These are briefly reviewed in Section 3.7.1. Additional processing to meet user preferences or to enable some particular application of the image must also be considered a part of the image quality evaluation. A few examples are given throughout. A comprehensive treatment of image
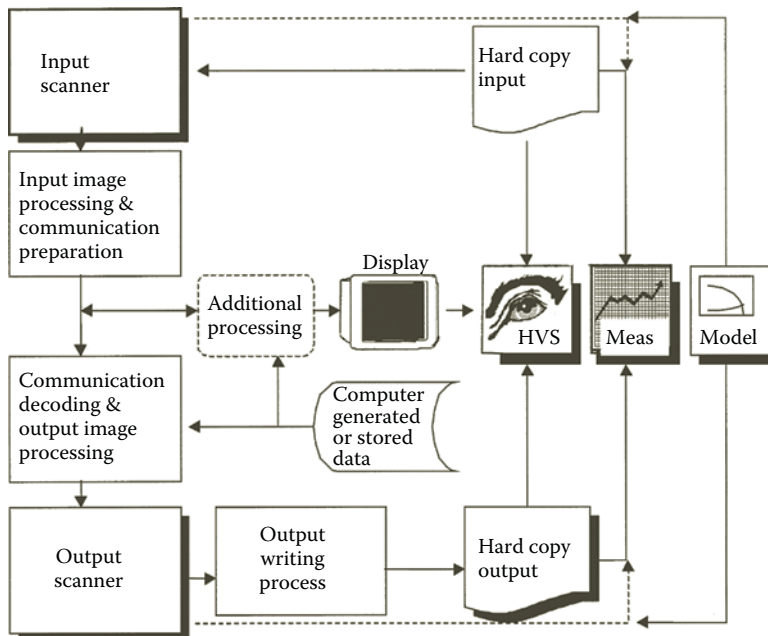
FIGURE 3.2
The elements of scanned imaging systems as they interact with the major methods of evaluating image quality. "HVS" refers to the human visual system. "Meas" refers to methods to measure both hard copy and electronic images and "Models" refers to predicting the imaging systems performance, not evaluating the images *per se.*

processing is beyond the scope of this chapter but several references are given at the end of this chapter to help the reader learn more about this critical area of scanned imaging.

The assessment of quality in the output image may take the form of evaluation by the human visual system (HVS) and the use of psychometric scaling (see Section 3.8) or by measurement with instruments as described in parts of Sections 3.3–3.5. One can also evaluate measured characteristics of the scanners and integrated systems or model them to try to predict, on average, the quality of images produced by these system elements. (Both of these hardware characterizations are also described in parts of Sections 3.3–3.5.) The description of overall image quality (Section 3.6) tends to focus on the models of systems and their elements, not the images themselves. For some purposes, for example, judging the quality of a copier, the comparison between the input and output images is the most important way of looking at image quality, whether it be by visual or measurement means. For other applications it is only the output image that counts. In some cases, the most common visual comparison is between the partially processed image, as can only be seen on the display, and either the input original or the hard copy output. In most cases, the evaluation criteria depend on the intended use of the image. A display of the scanned image in a binary (black or white) imaging mode reveals some interesting effects that carry through the system and often surprise the unsuspecting observer. These are covered in Section 3.5. Physical and visual measurements evaluate output and input images, hence the arrows in Figure 3.1 flow from hard copy toward these evaluation blocks. Models, however, are used mostly to synthesize imaging systems and components and may be used to predict or simulate performance and output. Hence the "model" arrows flow toward the system components.

The nonscanner components for electronic image processing and the analog writing process play a major role in determining quality and hence will be unavoidably included in

any realistic HVS or measurement evaluation of the quality of a scanned image or imaging system. Models of systems and components, on the other hand, often ignore the effects of these components and the reader is cautioned to be aware of this distinction when designing, analyzing, or selecting systems from the literature.

A model has been described by P. Engeldrum[12,40–42] called the Image Quality Circle, which ties all of these evaluations together and expands them into a logical framework to evaluate any imaging system. This is shown in Figure 3.3 as the circular path connecting the oval and box shapes, along with the three major assessment categories from Figure 3.2, namely the HVS, Measurements, and Models. In his model, the HVS category above is expanded to show a type of model he calls "visual algorithms," which predict human perceived attributes of images from physical image parameters. Examples of perceptions would include such visual subjective sensations as darkness, sharpness, or graininess (i.e., "nesses"). These are connected to physical measurements of densities, edge profiles, or halftone noise, respectively, made on the images used to evoke these subjective responses. In Engeldrum's analysis, the rest of what we call the HVS and brain combination includes "image quality models," which predict customer preferences based on relationships among the perceived attributes. This purely subjective dimension of individuals is often not included in the "brain" functions normally associated with HVS, therefore it is
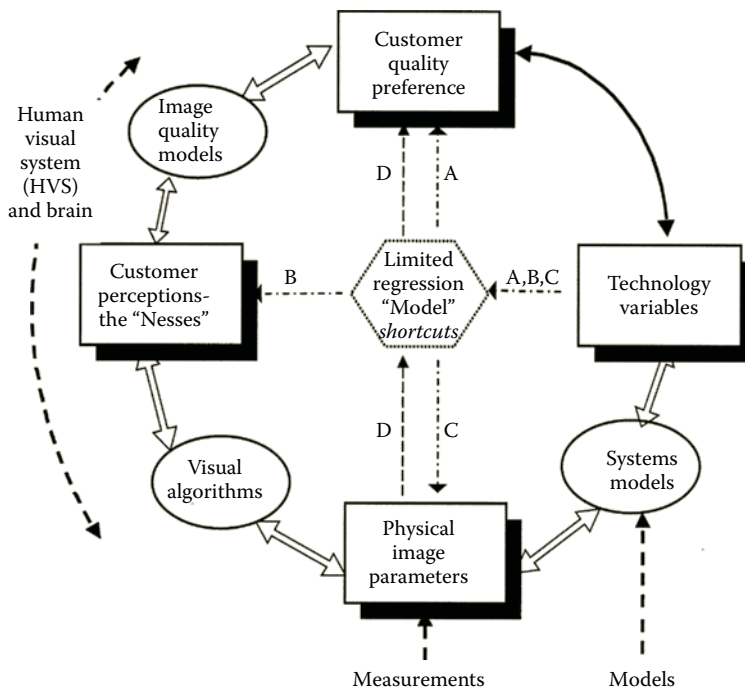
FIGURE 3.3

An overall framework for image quality assessment, composed of the elements connected by the outline arrows, known as the "Image Quality Circle" (adapted from Engeldrum, P.G. *Psychometric Scaling: A Toolkit for Imaging Systems Development;* Imcotek Press: Winchester, MA, 2000 and Engeldrum, P.G. *Chapter 2 Psychometric Scaling: A Toolkit for Imaging Systems Development;* IMCOTEK Press: Winchester, MA, 2000; 5–17.) and the inner "spokes" which illustrate four commonly used, but limited, regression model shortcuts as paths A, B, C, and D. The latter were not proposed by Engeldrum as part of the Image Quality Circle model, but added here to illustrate how selected examples given in Section 3.6 fit the framework. The connection to HVS, measurement, and model elements of Figure 3.2 are indicated by the labels and heavy dashed lines that surround the figure.

mentioned explicitly here. The methodologies to enable these types of analysis generally fall into the realm of psychometrics (quantifying human psychological or subjective reactions). They will be reviewed in Section 3.8.

Many authors (Section 3.6) have attempted to short-circuit this framework, following the dashed "spokes" we have added to the circle in Figure 3.3. These create regression models using psychometrics that directly connect physical parameters (path D) or technology variables with overall image quality models (path A) or preferences (path C). These have been partially successful, but, having left out some of the steps around the circle, they are very limited, often applying only to the circumstances used in their particular experiment. When these circumstances apply, however, such abbreviated methods are valuable. Following all the steps around the circle leads to a more complete understanding and more general models that can be adapted to a variety of situations where preferences and circumstances may be very different. The reader needs to be aware of this and judge the extent of any particular model's applicability to the problem at hand.

3.2 BASIC CONCEPTS AND EFFECTS

3.2.1 Fundamental Principles of Digital Imaging

The basic electronic imaging system performs a series of image transformations sketched in Figure 3.4. An object such as a photograph or a page with lines and text on it is converted from its analog nature to a digital form by a *raster input scanner* (RIS). It becomes "digital" in distance where microscopic regions of the image are each captured separately as discrete pixels; that is, it is *sampled*! It is then quantized, in other words, digitized in level, and is subsequently processed with various strictly digital techniques. This digital image is transformed into information that can be displayed or transmitted, edited, or merged with other information by the *electronic and software subsystem* (ESS). Subsequently a *raster output scanner* (ROS) converts the digital image into an analog form; that is, it is *reconstructed,* typically through modulating light falling on some type of photosensitive material. The latter, working through analog chemical or physical processes, converts the analog optical image into a reflectance pattern on paper, or into some other display as the final output image.

What follows assumes optical output conversion, but direct-marking processes, involving no optics (e.g., ink jet, thermal transfer, etc.) can be treated similarly. Therefore, while one often thinks of electronic imaging or scanned imaging as a digital process, we are really concerned in this chapter with the imaging equivalent of analog to digital (A/D) and digital to analog (D/A) processes. The digital processes occur between as image processing. In fact that is where we become familiar with the scanned imaging characteristics because that is one place where we can take a look at a representation of the image, that is, in a computer.

3.2.1.1 Structure of Digital Images

Before considering all the system and subsystem effects, let us turn our attention to the microscopic structure of this process, paying particular attention to the A/D and sampling domain of the input scanner. Sampled electronic images were first studied in a comprehensive way by Mertz and Gray.[1]
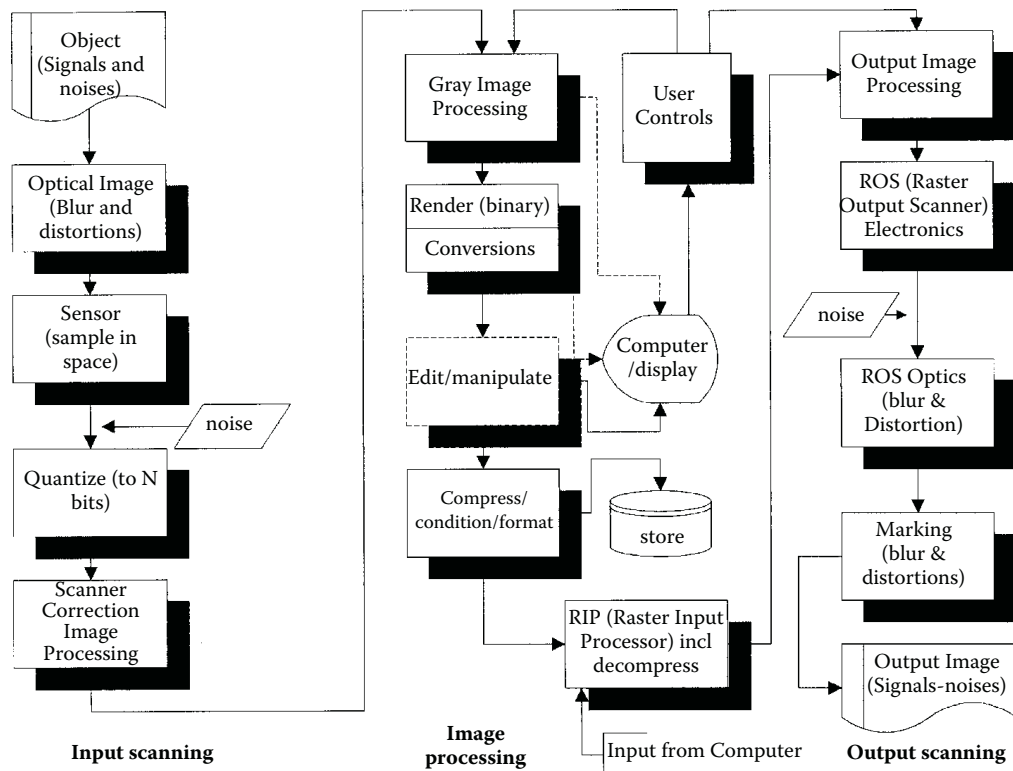
FIGURE 3.4
Steps in typical scanning electronic reprographic system showing basic imaging effects.

To understand how sampling works, let us examine Figure 3.5. It illustrates four different aspects of the input scanning image transformations. Figure 3.5a shows the microscopic reflectance profile representative of an input object: there is a sharp edge on the left, a "fuzzy" edge (ramp), and a narrow line. Figure 3.5b shows the optical image, which is a blurred version of the input object. Note that the relative heights of the two pulses are now different and the edges are sloping that were previously straight. Figure 3.5c represents the blurred image with a series of discrete signals, each being centered at the position of the arrows. This process is referred to as *sampling.*

Each sample in Figure 3.5c has some particular height or gray value associated with it (scale at right). When these individual samples can be read as a direct voltage or current, that is they can have any level whatsoever, then the system is analog. When an element in the sensor output circuit creates a finite number of gray levels such as 10, 128, or even 1000, then the signal is said to be *quantized.* (When a finite number of levels is employed and is very large, the quantized signal resembles the analog case.) Being both sampled and quantized in a form that can be manipulated by a digital processor makes the image *digital.* Each of these individual samples of the image is a *picture element,* often referred to as a *pixel* or pel. A sampled and *multilevel* (>2) quantized image is often referred to as a *grayscale image* (a term also used in a different context to describe a continuous tone analog image). When the quantization is limited to *two levels,* it is termed a *binary image.* Image processing algorithms that manipulate these different kinds of images can be "bit constrained" to the

FIGURE 3.5
Formation of binary images, illustrating how a single, blurred electronic image of a small continuous tone test object could yield many different binary images depending on the threshold selected. (a) 1–reflectance profile of the test pattern, (b) blurred analog electronic image of the test pattern in relative response units (e.g. relative millivolts), (c) Sampling into pixels where each arrow at top represents a pixel location. The magnitude of each arrow at bottom represents the response at that pixel location. The % reflectance and the assigned grey level response are given at the right where the larger grey level represents the blacker (lower reflectance) parts, (d) Each row represents a different threshold for pixels shown in (c), each black dot represents a black image pixel, with thresholds for each row identified at the right.

number of levels appropriate to the image bit depth (another expression for the number of levels), that is, integer arithmetic. This is effectively equivalent to many digital image processing circuits. Alternatively, algorithms may be floating point arithmetic, the results of which are quite different from the bit constrained operations.

  A common and simple form of image processing is the conversion from a gray to a binary image as represented in Figure 3.5d of Figure 3.5. In this process a threshold is set at some particular gray level, and any pixel at or above that level is converted to white or black. Any pixel whose gray value is below that level is converted to the other signal, that is, black or white, respectively. Four threshold levels are shown in Figure 3.5c by arrows on the gray-level scale at the right. Results are depicted in Figure 3.5d as four rows, each being a raster from the different binary images, one for each of the four thresholds.

In Figure 3.5d, each black pixel is represented by a dot, and each white pixel is represented by the lack of a dot. (It is common to depict pixels as series of contiguous squares in a lattice representing the space of the image. They are better thought of as points in time and space that can have any number of dimensions, attributes, and properties.)

Each row of dot patterns shows one line of a sampled binary image. These patterns are associated with the location of the sampling arrows, shown in Figure 3.5c, the shape of the blur, and the location of the features of the original document. Notice at the 85% threshold, the narrow line is now represented by two pixels (i.e., it has grown), but the wider and darker pulse has not changed in its representation. It is still five-pixel wide. Notice that the narrow pulse grew in an asymmetric fashion and that the wider pulse, which was asymmetric to begin with, grew in a symmetric fashion. These are quite characteristic of the problems encountered in digitizing an analog document into a finite number of pixels and gray levels. It can be seen that creating a thresholded binary image is a highly nonlinear process. The unique imaging characteristics resulting from thresholding are discussed in detail in Section 3.5.

Figure 3.6 represents the same type of process using a real image. The plot is the gray profile of the cross section of a small letter "I" for a single scan line. The width of the letter is denoted at various gray levels, indicated here by the label "threshold" to indicate where one could select the potential black to white transition level. The reader can see that the width of the binary image can vary anywhere from one to seven pixels, depending on the selection of threshold.

Figure 3.7a returns to the same information shown in Figure 3.5, except that here we have doubled the frequency with which we sampled the original blurred optical image. There are now twice as many pixels, and their variation in height is more gradual. In this particular instance, increased resolution is responsible for the binary case detecting the narrow pulse at a lower level (closer to 0% threshold). This illustration shows the general results that one would expect from increasing the spatial density at which one samples the image; that is, one sees somewhat finer detail in both the gray and the binary images with higher sampling frequency.
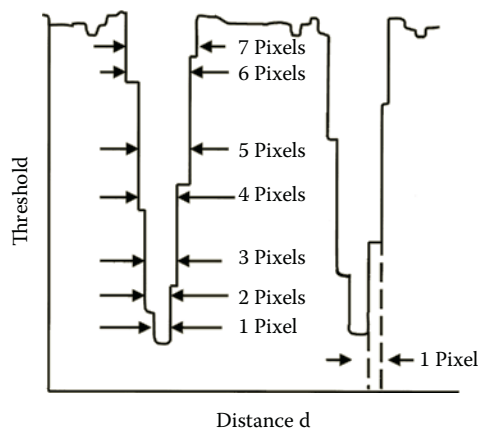
FIGURE 3.6
An actual scanned example of a gray scan line across the center of a letter "I." A different representation of the effect shown in step (c) in Figure 3.5. Here the sample points are displayed as contiguous pixels. The width of one pixel is indicated. The image is from a 400 dpi scan of approximately a six-point Roman font.

(a) Higher density sampling (e.g. Double above)

Each dot
represents
black pixel

Thresholds
as before

←— 5
←— 50
←— 65 New threshold
       to "just detect
←— 75 narrow line"
←— 85

(b) The effect of sampling phase

(c) The effect of blur

Phase A

0%
80
100%

Phase B

0%
70
100%

Phase "A"
Phase "B"

"A"

0
60
100

"B"

0
30
100
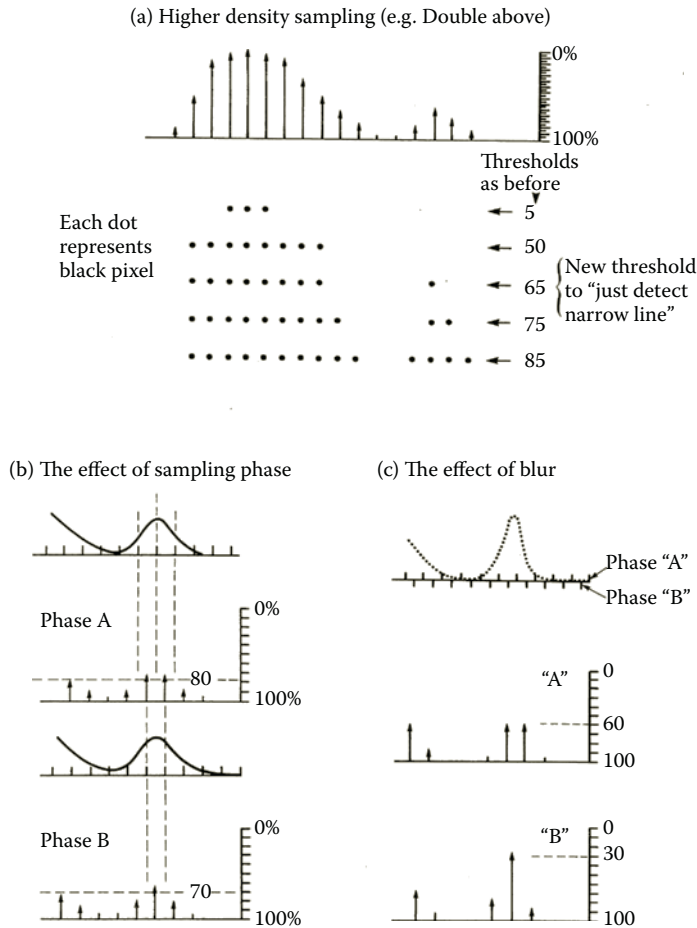
FIGURE 3.7
The effects of (a) doubling the resolution, (b) changing sampling phase, (c) sharpening the optical image.

This is, however, not always the case when examining every portion of the microstructure. Let us look more closely at the narrower of the two pulses (Figure 3.7b). Here we see the sampling occurring at two locations, shifted slightly with respect to each other. These are said to be at different sampling phases. In phase A the pulse has been sampled in such a way that the separate pixels near the peak are identical to each other in their intensity, and in phase B one of the pixels is shown centered on the peak. When looking at the threshold required to detect the information in phase A and phase B, different results are obtained for a binary representation of these images. Phase B would show the detection of the pulse at a lower threshold (closer to ideal) and phase A, when it detects the pulse, would show it as wider, namely as two pixels in width.

Consider an effect of this type in the case of an input document scanner, such as that used for facsimile or electronic copying. While the sampling array in many input scanners is constant with respect to the document platen, the location of the document on the platen is random. Also the locations of the details of any particular document within the format of the sheet of paper are random. Thus the phase of sampling with respect to detail is random and the type of effects illustrated in Figure 3.7 would occur randomly over a page.

There is no possibility that a document covered with some form of uniform detail can look absolutely uniform in a sampled image. If the imaging system produces binary results, it will consistently exhibit errors on the order of one pixel and occasionally two pixels of edge position and line width. The same is true of a typically quantized gray image, except now the errors are primarily in magnitude and may, at higher sampling densities, be less objectionable. In fact, an analog gray imaging process, sampling at a sufficiently high frequency, would render an image with no visible error (see the next subsection). Continuing with the same basic illustration, let us consider the effect of blur. In Figure 3.7c we have sketched a less blurred image in the region of the narrower pulse and now show two sampling phases A and B, as before, separated by half a pixel width. Two things should be noted. First, with higher sharpness (i.e., less blur), the threshold at which detection occurs is higher. Secondly, the effect of sampling phase is much larger with the sharper image. Highly magnified images in Figure 3.8 illustrate some of these effects.

3.2.1.2 The Sampling Theorem and Spatial Relationships

By means of these illustrations we have shown the effects of sampling frequency, sampling phase, and blur at an elementary level. We now turn our attention to the more formal description of these effects in what is known as the sampling theorem. For these purposes we assume that the reader has some understanding of the concepts of Fourier analysis or at least the frequency-domain way of describing time or space, such as in the frequency analysis of audio equipment. In this approach, distance in millimeters is transformed to frequency in cycles per millimeter (cycles/mm). A pattern of bars spaced 1 mm apart would result in 1 cycle/mm as the fundamental frequency of the pattern. If the bars were represented by a square wave, the Fourier series showing the pattern's various harmonics would constitute the frequency-domain equivalent.

Figure 3.9 has been constructed from such a point of view. In Figure 3.9a we see a single-raster profile of an analog input document (i.e., an object) represented by the function $f(x)$.
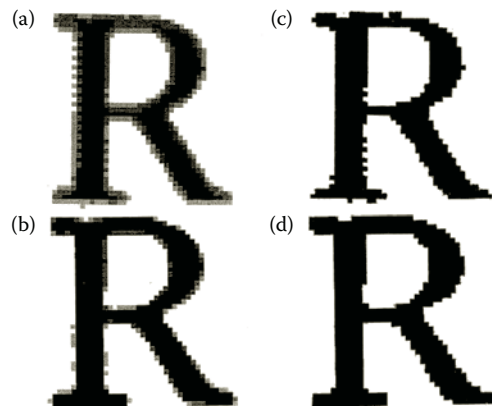
FIGURE 3.8
Digital images of a 10-point letter "R" scanned at 400 dpi showing quantization and sharpening effects. Parts (a) and (c) were made with normal sharpness for typical optical systems and parts (b) and (d) show electronic enhancement of the sharpness (see Figure 3.32). Parts (a) and (b) are made with 2 bits/pixel, that is, four levels including white, black and two levels of gray. Parts (c) and (d) are 1 bit/pixel images, that is, binary with only black and white where the threshold was set between the two levels of gray used in (a) and (b). Note the thickening of some strokes in the shaper image and the increased raggedness of the edges in the binary images. Some parts of the sharp binary images are also less ragged.
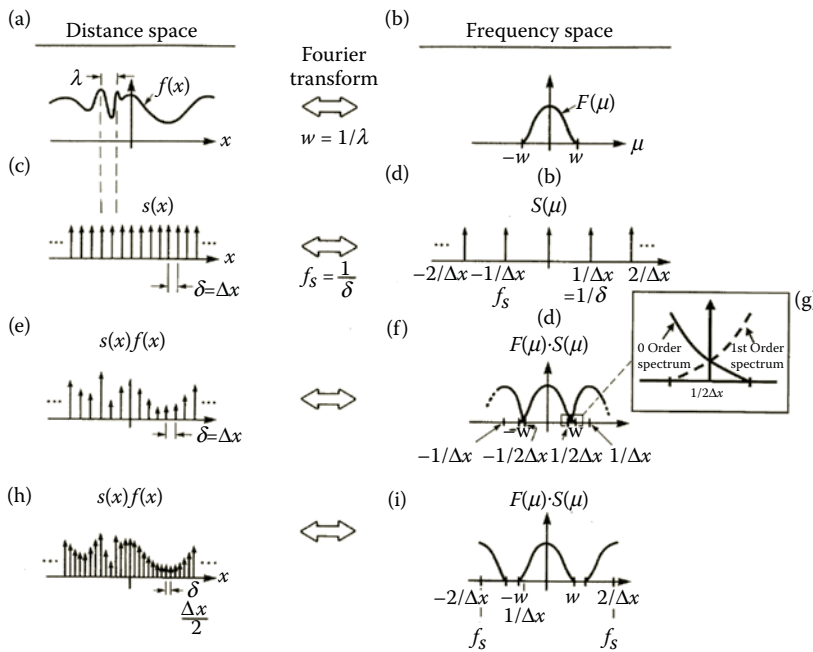
FIGURE 3.9
The Fourier transformation of images and the effects of sampling frequency. The origin and prevention of aliasing: (a) original object; (b) spectrum of object; (c) sampling function; (d) spectrum of sampling function; (e) sampled object; (f) spectrum of sampled object; (g) detail of sampled object spectrum; (h) object sampled at double frequency; and (i) spectrum of object sampled at double frequency. (Adapted from Gonzalez, R.C.; Wintz, P. *Digital Image Processing;* Addison, Wesley: Reading, MA, 1977; 36–114.)

This is a signal extending in principle to +∞ and contains, upon analysis, many different frequencies. It could be thought of as a very long microreflectance profile across an original document. Its spectral components, that is, the relative amplitudes of sine waves that fit this distribution of intensities, are plotted as $F(\mu)$ in Figure 3.9b. Note that there is a maximum frequency in this plot of amplitude versus frequency, at w. It is equal to the reciprocal of λ (the wavelength of the finest detail) shown in Figure 3.9a. This is the highest frequency that was measured in the input document. The frequency w is known as the bandwidth limit of the input document. Therefore the input document is said to be band limited. This limit is often imposed by the width of a scanning aperture that is performing the sampling in a real system.

We now wish to take this analog signal and convert it into a sampled image. We multiply it by $s(x)$, a series of narrow impulses separated by Δx as shown in Figure 3.9c. The product of $s(x)$ and $y\ f(x)$ is the sampled image, and that is shown in Figure 3.9e. To examine this process in frequency space, we need to find the frequency composition of the series of impulses that we used for sampling. The resulting spectrum is shown in Figure 3.9d. It is, itself, a series of impulses whose frequency locations are spaced at $1/\Delta x$ apart. For the optical scientist this may be thought of as a spectrum, with each impulse representing a different order; thus the spike at $1/\Delta x$ represents the first-order spectrum, and the spike at zero represents the zero-order spectrum. Because we multiplied in distance space in order to come up with this sampled image, in frequency space, according to the convolution theorem, we must convolve the spectrum of the input document with the spectrum of the

sampling function to arrive at the spectrum of the sampled image. The result of this convolution is shown in Figure 3.9f.

Now we can see the relationship between the spectral content of the input document and the spacing of the sampling required in order to record that document. Because the spectrum of the document was convolved with the sampling spectrum, the negative side of the input document spectrum $F(\mu)$ folds back from the first-order over the positive side of the zero-order document spectrum. Where these two cross is exactly halfway between the zero- and first-order peaks. It is a frequency $(1/2\Delta x)$ known as the Nyquist frequency. If we look at the region in Figure 3.9g between zero and the Nyquist frequency, the region reserved for the zero-order information, we see that there is "contamination" from the negative side of the first order down to the frequency $[(1 - \Delta x) - w]$, where w is the band limit of the signal. Any frequency above that point contains information from both the zero and the first order and is therefore corrupted or mixed, often referred to as *aliased.*

Should one desire to avoid the problem of aliasing, one must sample at a finer sampling interval, as shown in Figure 3.9h. Here the spacing is one-half that of the earlier sketches, and therefore the sampling frequency is twice as high. This also doubles the Nyquist frequency. This merely separates the spectra by spreading them out by a factor of 2. Since there is no overlap of zero and first orders in this example, one can recover the original signal quite easily by simply filtering out the higher frequencies representing the orders other than zero. This is illustrated in Figure 3.10, where a rectangular function of width $\pm w$ and amplitude 1 is multiplied by the sampled image spectra, resulting in recovery of the original signal spectra. When inversely Fourier transformed, this would give the original signal back [compare Figures 3.10e and 3.9a].

We can now restate Shannon's[43] formal sampling theorem, [sometimes referred to as the Whittaker–Shannon Sampling Theorem (R. Loce, personal communication, 2001)] in terms
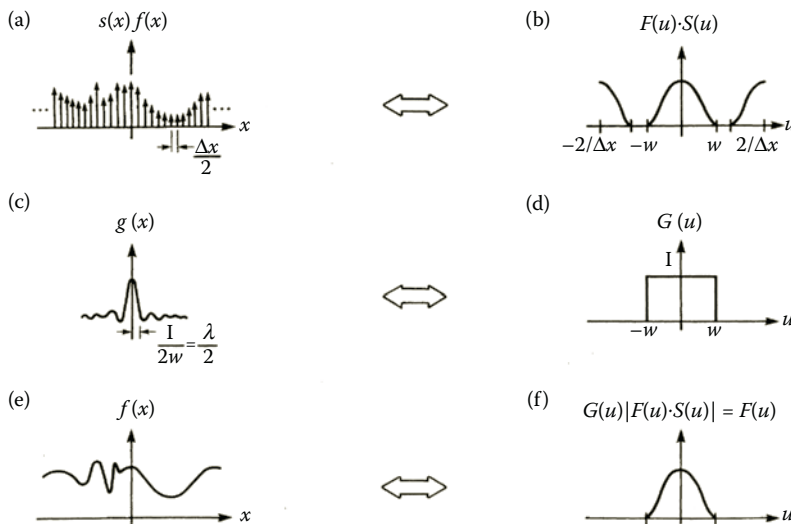
FIGURE 3.10
Recovery of original object from properly sampled imaging process: (a) object sampled at double frequency (from Figure 3.9h); (b) spectrum of "a" (from Figure 3.9i); (c) spread function for rectangular frequency filter function; (d) rectangular frequency function; (e) recovered object function; (f) recovered object spectrum. (Adapted from Gonzalez, R.C.; Wintz, P. *Digital Image Processing;* Addison, Wesley: Reading, MA, 1977; 36–114.)

that apply to sampled imaging: if a function $f(x)$ representing either an original object or the optical/aerial image being digitized contains no frequencies higher than w cycles/mm (this means that the signal is band limited at w), it is completely determined by giving its values at a series of points $<1/2w$ mm apart. It is formally required that there be no quantization or other noise and that this series be infinitely long; otherwise windowing effects at the boundaries of smaller images may cause some additional problems (e.g., digital perturbations from the presence of sharp edges at the ends of the image). In practice, it needs to be long enough to render such windowing effects negligible.

It is clear from this that any process such as imaging by a lens between the document and the actual sampling, say by a CCD sensor, can band limit the information and ensure accurate effects of sampling with respect to aliasing. However, if the process of band limiting the signal in order to prevent aliasing causes the document to lose information that was important visually, then the system is producing restrictions that would be interpreted as excessive blur in the optical image. Another way to improve on this situation is, of course, to increase the sampling frequency, that is, decrease the distance between samples.

We have shown in Figure 3.10 that the process of recovering the original spectrum is accomplished by a filter having a rectangular shape in frequency space (Figure 3.10d). This filter is known as the reconstruction filter and represents an idealized reconstruction process. The rectangular function has a $(\sin x)/x$ inverse transform in distance space (Figure 3.10c), whose zero crossings are at $\pm N\Delta x$ from the origin where $N = 1,2, \ldots$ Rectangular and other filters with flat modulation transfer functions (MTFs) are difficult to realize in incoherent systems. This comes about because of the need for negative light in the sidelobes (in distance space). A reconstruction filter need not be precisely rectangular in order to work. It should be relatively flat and at a value near 1.0 over the bandwidth of the signal being reconstructed (also difficult and often impossible to achieve). It must not transmit any energy from the two first-order spectra. If the sampling resolution is very high and the bandwidth of the signal is relatively low, then the freedom to design the edge of this reconstruction filter is relatively great and therefore this edge does not need to be as square. From a practical point of view the filter is often the MTF of the output scanner, typically a laser beam scanner, and is not usually a rectangular function but more of a Gaussian shape. A nonrectangular filter, such as that provided by a Gaussian laser beam scanner, alters the shape of the spectrum that it is trying to recover. Because the spectrum is multiplied by the reconstructing MTF, this causes some additional attenuation in the high frequencies, and a trade-off is normally required in practical designs.

3.2.1.3 Gray Level Quantization: Some Limiting Effects

Now that we have seen how the spatial or distance dimension of an input image may be digitized into discrete pixels, we explore image quantization into a finite number of discrete gray levels. From a practical standpoint this quantization is accomplished by an A/D converter, which quantizes the signal into a number of gray levels, usually some power of 2. A popular quantization is 256 levels, that is, 8 bits, which lends itself to many computer applications and standard digital hardware. There may be good reasons for other quantizations, higher or lower, to optimize a design or a system. (See Reference 21, pp. 213–227 for some practical applications and tests.)

From an overall systems engineering perspective, one needs to understand the limits on the useful number of quantization levels. This should be based upon noise in the input as seen by the system (inbound limit) or upon the ultimate output goal of how many distinguishable gray levels can be seen by the human eye (outbound limit). Both approaches

have been explored in the literature and involve complex calculations and experimental measurements.

Use of the *HVS response* with various halftoning methods represents an *outbound limit* approach to defining practical quantization limits for scanned imaging. The "visual limit" results shown in Figure 3.11[44] plot the number of visually distinguishable gray levels against the spatial frequency at which they can be seen. This curve was derived from a very conservative estimate of the visual system frequency response and may be thought of as an upper limit on the number of gray levels required by the eye. Plotted on the same curve are performance characteristics for 20 pixels/mm (500 pixels/in) digital imaging systems that produce 3 bits/pixel and 1 bit/pixel (binary) images. These were obtained by use of a generalized algorithm to create halftone patterns (see Section 3.2.2.3 and Reference 45) at different spatial frequencies. The binary limit curve, added here to Roetling's, graph, shows the number of effective gray levels for each frequency whose period is two halftone cells wide. The 3-bit limit assumes each halftone cell contributes 2^3 gray values, including black and white.

Roetling[44] integrated the visual response curve to find an average of 2.8 bits/pixel as a good upper bound for the eye itself. Note that his general halftoning approach, using 3 bits/pixel and 20 pixels/mm (500 pixels/in) also approximates the visual limit in the important midfrequency region. Specialized halftoning techniques[7,45] may produce different and often more gray levels per pixel at the lower frequencies.

Another approach to setting quantization limits is to examine the noise in the input, assuming in so doing that the quantization is *input bound* and not output bound by the visual process as in the foregoing approach. A range of photographic input was selected as examples of a practical lower limit (best) on input noise. The basic principle for describing the useful number M of gray levels in a photograph involves quantizing its density scale into steps whose size is based on the noise (granularity) of that photographic image[46] when scanned by the digital imaging process. In simplified terms this can be described as

$$M = \frac{L}{2k s_{\mathrm{a}}} \tag{3.1}$$
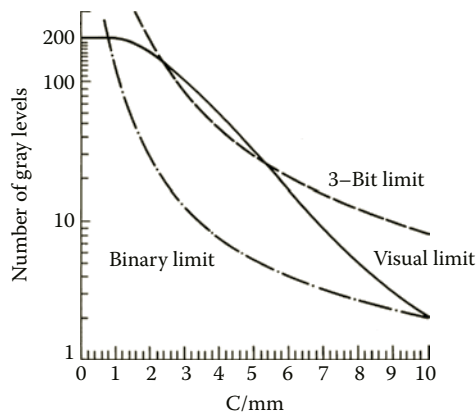
FIGURE 3.11
Example of outbound quantization limits, using visually distinguishable number of gray levels versus spatial frequency, with corresponding 1 (binary) (adapted from Roetling, P.G. Visual performance and image coding. *Proceedings of the Society of Photo-Optical Instrumentation Engineers on Image Processing,* Vol. 74, 1976; 195–199.) and 3 bit/pixel limits.

where L = the density range of the image, σ_a = measured standard deviation of density using aperture area = A, and k = the number of standard deviations in each distinguishable level.

The question being addressed by this type of quantization is how reliably one wants to be able to determine the specific tone in a given part of the input picture from a reading of a single pixel. For some purposes, where the scanned image is used to extract radiometric information from a picture,[46] the reliability must be high, for other cases such as simply copying a scene for artistic purposes it can be much lower. To precisely control a digital halftone process (see later) it must be fairly high.

Photographic noise is approximately random uncorrelated noise. To a first order, photographic noise (granularity) is the standard deviation of the density fluctuations. It is directly proportional to the square root of the effective detection area,[47,48] a of a measuring instrument or scanner-sensor, that is, Selwyn's law:

$$s_a = S(2a)^{1/2} \tag{3.2}$$

where S is a proportionality constant defined as the Selwyn granularity. It is also proportional to the square root of the mean density, that is, Siedentopf's relationship,[47,49] in an ideal film system. In practical cases, as is done here, the density relationship must be empirically determined. Figure 3.12 shows the number of distinguishable gray levels reported in the literature by various authors for various classes of films obtained by directly measuring granularity as a function of density. They are reported at apertures that are approximately equivalent in size to the smallest detail the film could resolve, that is, the diameter of the film spread function. For a real world example, assume that 35-mm film images are enlarged perfectly by a high-quality 3.3 × enlarger. The conversion to the number of distinguishable gray levels per pixel is based on assuming Selwyn's law, a reliability of 99.7% (± 3λ or $k = 6$) and that any nonlinear relationship between granularity and density scales as the aperture size changes. The actual scanner aperture is reduced by 3.3 × in its two dimensions to resemble directly scanning the film.

Four specific films were selected, each representative of a different class, three of which are black and white films: (1) an extremely fine-grained microfilm; (2) a fine-grained amateur film; and (3) a high-speed amateur film.[50] A special purpose color film was also included.[51] Despite now being obsolete, these films still represent a reasonable cross section of photographic materials. A 3.3 × enlargement was selected as typical of consumer practice, roughly giving a 3.5" × 5" print from a 35-mm negative. The reciprocal of this magnification is used to scale the scanner aperture back to film dimensions. Two popular scanner resolutions of 600 and 300 dpi were selected. The corresponding sensor "aperture" widths in μm, scaled to the film, are noted in parentheses in the key at the top of each figure. The width is the inverse of the sampling period. A third scanner aperture, equivalent to that in the Roetling visual calculations, was used for one case, that is, a 20 samples/mm (500 samples/in) scanning system with an aperture of 50 × 50 μm (2 × 2 mils). The fourth situation, called "Film @ max" describes the number of levels resulting from scanning the film with an aperture that matches the blur (spread function) for the film, given in the film category label in parentheses at the bottom of each figure. These approximate calculations are an oversimplification of the photographic and enlarging processes, ignoring significant nonlinearities and blurring effects, but they provide a rough first-order analysis.

Examination of the charts suggests that a practical range of inbound quantization limits (IQLs) for pictorial images is approximately anywhere from 2 to 4 bits/pixel (microfilm is
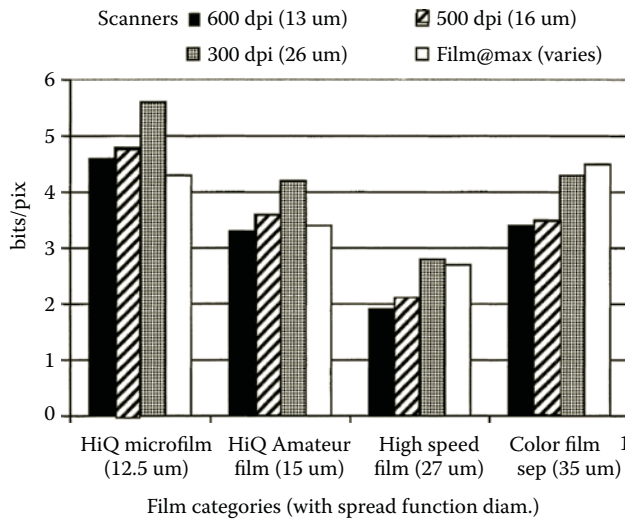
FIGURE 3.12

Example of inbound quantization limits, using the number of distinguishable gray levels, in bits/pixel, for input consisting of 3.3 × enlargements (3 × 5 in prints of 35 mm film) from four example films (adapted from Altman, J.H.; Zweig, H.J. Effect of spread function on the storage of information on photographic emulsions. *Photog. Sci. Eng.* 1963, 7, 173–177 and Lehmbeck, D.R. Experimental study of the information storing properties of extended range film. *Photog. Sci. Eng.* 1967, 11, 270–278.) scanned by four generic types of systems indicated by their scanning resolutions. Color film is for a single separation, others are black and white films. The limiting blur in μm for the first three scanners is given in the parentheses after the scan frequency. It is the sensor aperture width scaled to the film size. The fourth scanner has variable resolution set by a scaled aperture width adjusted to equal the width of each film blur function (spread function), shown in parentheses with the film type. Assumes a 99.7% confidence on distinguishability using Equation 3.1 (i.e., with $k = 6$).

not made for pictorials). For typical high-quality reproduction, then, an input bound limit is a little over 3 bits/pixel at 600 dpi using the three standard deviation criterion. This compares with the rate of 2.8 bits/pixel found by Roetling for a visual outbound quantization limit (OQL). Recent work by Vaysman and Fairchild,[52] limited to an upper frequency of 300 dpi by their printer selection, also found, through psychophysical studies, that 3 bits/pixel/color was a useful system optimum for reproducing color pictures.

One may ask, then, why are there so many input scanners operating at 8, 10, or even 12 bits/pixel? First of all there are many reasons to modify these calculations for specific situations such as larger tolerances on probabilities for distinguishing differences less reliably, considering larger sampling apertures for certain rendering/viewing methods, different frequency weightings and many others that would result in more inbound or outbound gray levels (Reference 53, p. 198).

A very practical reason, however, is that actual hardwired scanners cannot adapt to detail and granularity in originals and change performance striving for these optimums in the way they were calculated. For slower, computationally intensive, off-line image processing which can adapt to the information in small regions of the image (as in the case of JPEG and other lossy compressions—see Section 3.7.1) one can, in essence, approximate the limits just discussed. (As an example 10–20 × JPEG compression for 8-bit images works well and is approx. 4+ bits, leaving 3+ bits for the resulting image.)

Actual hardwired real-time scanners have to assume the worst case (e.g., 200 gray levels—see Figure 3.11). This is rounded up to 256 or 8 bits. However, the 200 gray levels

are not equally spaced in linear units. They are essentially spaced as equal increments on an L^* scale (See Equations 3.5 and 3.8.)

Thus for an input density of 2.0 or an L^* of ~9 and a difference of .5 L^* (=1/200 of full L^* scale) a linear difference of ~.0006 is called for. This is 1 part out of 1700, or more than 10 bits (1024) and would require an 11-bit system (not a common A to D circuit). From a practical perspective that suggests 12 bits (4096 levels) which allows for some enhancement of high density areas. For those believing that a ΔL^* of 1 is just noticeable (true for certain conditions) the above situation calls for approx. 870 levels and 10 bits is satisfactory. Another alternative to stay with 8 bits and meet the visual requirements is to distort the linear sensor response via scanner electronics to approximate L^* (some digital cameras and scanners do this) prior to the final digital output

Being aware of the inbound limits, the system options and the outbound limits as an endpoint gives a framework for robust engineering and optimization of image quality in a systems context. Information capacity approaches extend these concepts (see Section 3.6.7).

3.2.2 Basic System Effects

3.2.2.1 Blur

Blur, that is, the spreading of the microscopic image structure, is a significant factor in determining the information in an image and therefore its quality. In the input scanner, blur is caused by the optical system, the size and properties of the light-sensing element, other electronic elements, and by mechanical and timing factors involved in motion. This blur determines whether the system is aliased. Roughly speaking, if the image of a point (the profile of which is called the point spread function) spreads over twice the sampling interval, the system is unaliased. The spreading also determines the contrast of fine details in the gray video image prior to processing. The cascading of these elements can be described conveniently by a series of spatial frequency responses [see later under MTFs for a detailed discussion] or other metrics that relate generally to the sharpness of optical images. It can be compensated for, in certain aspects, by subsequent electronic or computer image processing.

Blur in an output scanner is caused by the size of the writing spot, for example, the laser beam waist at focus, by modulation techniques and by the spreading of the image in any marking process such as xerography or photographic film. It is also affected by motion of the beam relative to the data rate and by the rate of motion of the light-sensitive receptor material. Output scanner blur more directly affects the appearance of sharpness in the final hard copy image that is presented to the HVS than does blur in the input scanner. Overall enhancement of the electronic input scanned image can, however, draw visual attention to details of the output image unaffected by blur limitations of either scanner.

Blur for the total system, from input scanner through various types of image processing to output scanner and then to marks on paper, is not easily cascaded, because the intervening processing of the image information is extremely nonlinear. This nonlinearity may give rise to such effects as a blurred input image looking very sharp on the edges of a binary output print because of the small spot size and low blur of the marking process. In such a case, however, the edges of square corners look rounded and fine detail such as serifs in text or textures in photographs may be lost. Conversely, a sharp input scan printed by a system with a large blurring spot would appear to have fuzzy edges, but the edge noise due to sampling would have been blurred together and would be less visible than in the first case. Moiré, from aliased images of periodic subjects caused by

low blur relative to the sample spacing, however, would still be present in spite of output blur. (Note, superposition of periodic patterns such as a halftoned document (see Section 3.2.2.3) and the sampling grid of a scanner results in new and often striking periodic patterns in the image commonly called Moiré patterns (see Bryngdahl[54]). Once aliased, no amount of subsequent processing can remove this periodic aliasing effect from an image.) The popular technologies called "anti aliasing" deal with a different effect of undersampling, namely that binary line images exhibit strong visible staircase or jaggie effects on slanted lines when the output blur and sampling are insufficient for the visual system. These techniques nonlinearly "find" the stair steps and locally add gray pixels to reduce the visibility of the jaggie (see Section 3.7.2 and Figure 3.43).[55] Aliasing is also known as spurious response.[13]

It is apparent, then, that blur can have both positive and negative impacts on the overall image quality and requires a careful trade-off analysis when designing scanners.

3.2.2.2 System Response

There are four ways in which electronic imaging systems display or print tonal information to the eye or transmit tonal information through the system:

1. By producing a signal of varying strength at each pixel, using either amplitude or pulse-width modulation.

2. By turning each pixel on or off (a two-level or binary system; see Section 3.5).

3. By use of a halftoning approach, which is a special case of binary imaging. Here, the threshold for the white–black decision is varied in some structured way over very small regions of the image, simulating continuous response. Many, often elaborate, methods exist for varying the structure; some involve multiple pixel interactions (such as error diffusion; see the end of Section 3.2.2.3) and others use subpixels (such as high addressability, extensions of the techniques mentioned in Section 3.7.2).

4. By hybrid halftoning combining the halftone concept in (3) with the variable gray pixels from (1) (e.g., see References 44 and 45).

From a hardware point of view, the systems are either designed to carry gray information on a pixel-by-pixel basis or to carry binary (two-level) information on a pixel-by-pixel basis. Because a two-level imaging system is not very satisfactory in many applications, some context is added to the information flow in order to obtain pseudogray using the halftoning approach.

Macroscopic tone reproduction is the fundamental characteristic used to describe all imaging systems' responses, whether they are analog or digital. For an input scanner it is characterized by a plot of an appropriate, macroscopic output response, as a function of some representation of the input light level. The output may characteristically be volts or digital gray levels for a digital input scanner and intensity or perhaps darkness or density of the final marks-on-paper image for an output scanner. The correct choice of units depends upon the application for which the system response is being described. There are often debates as to whether such response curves should be in units of density or optical intensity, brightness, visual lightness or darkness, gray level, and so on. For purposes of illustration, see Figure 3.13.
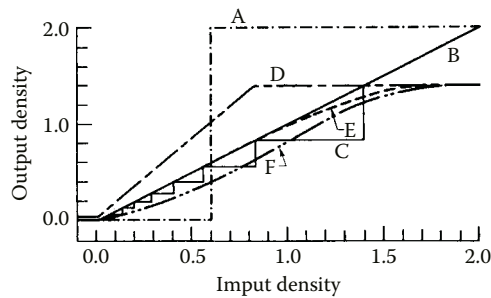
FIGURE 3.13
Some representative input/output density relationships for (A) binary imaging response; (B) linear imaging response; (C) stepwise linear response; (D) saturation–limited linear response; (E) linear response with gradual roll-off to saturation; (F) idealized response curve for best overall acceptability.

Here we have chosen to use the conventional photographic characterization of output density plotted against input density using normalized densities. Curve A shows the case of a binary imaging system in which the output is white or zero density up to an input density of 0.6, at which point it becomes black or 2.0 output density. Curve B shows what happens when a system responds linearly in a continuous fashion to input density. As the input is equal to the output here, this system would be linear in reflectance, irradiance, or even Munsell value (visual lightness units).

Curve C shows a classic abridged gray (severely limited number of levels) system attempting to write linearly but with only eight levels of gray. This response becomes a series of small steps, but because of the choice of density units, which are logarithmic, the sizes of the steps are very different. Had we plotted output reflectance as a function of input reflectance, the sizes of the steps would have been equal. However, the visual system that usually looks at these tones operates in a more or less logarithmic or power fashion, hence the density plot is more representative of the visual effect for this image. Had we chosen to quantize in 256 gray levels, each step shown would have been broken down into 32 smaller substeps, thereby approximating very closely the continuous curve for B.

When designing the system tone reproduction, there are many choices available for the proper shape of this curve. The binary curve, as in A, is ideal for the case of reproducing high-contrast information because it allows the minimum and maximum input densities considerable variation without any change to the overall system response.

For reproducing continuous tone pictures, there are many different shapes for the relationship between input and output, two of which are shown in Figure 3.13. If, for example, the input document is relatively low contrast, ranging from 0 to 0.8 density, and the output process is capable of creating higher densities such as 1.4, then the curve represented by D would provide a satisfactory solution for many applications. However, it would create an increase in contrast represented by the increase in the slope of the curve relative to B, where B gives one-for-one tone reproductions at all densities. Curve D is clipped at an input density greater than 0.8. This means that any densities greater than that could not be distinguished and would all print at an output density of 1.4.

In many conventional imaging situations the input density range exceeds that of the output density. The system designer is confronted with the problem of dealing with this mismatch of dynamic ranges. One approach is to make the system respond linearly to density up to the output limit; for example, following curve B up to an output density

of 1.4 and then following curve D. This generally produces unsatisfactory results in the shadow regions for the reasons given earlier for curve D. One general rule is to follow the linear response curve in the highlight region and then to roll off gradually to the maximum density in the shadow regions starting perhaps at a 0.8 output density point for the nonlinear portion of the curve as shown by curve E. Curve F represents an idealized case approximating a very precisely specified version arrived at by Jorgenson.[56] He found the "S"-shaped curve resembling F to be a psychologically preferred curve among a large number of the curves he tried for lithographic applications. Note that it is lighter in the highlights and has a midtone region where the slope parallels that of the linear response. It then rolls off much as the previous case toward the maximum output density at a point where the input density reaches its upper limit.

3.2.2.3 Halftone System Response

One of the advantages of digital imaging systems is the ability to completely control the shape of these curves to allow the individual user to find the optimum relationship for a particular photograph in a particular application. This can be achieved through the mechanism of digital halftoning as described below. Historically important studies of tone reproduction, largely for photographic and graphic arts applications, include those of Jones and Nelson,[57] Jones,[58] Bartleson and Breneman,[59] and two excellent review articles, covering many others, by Nelson.[60,61] Many recent advances in the technology of digital halftoning have been collected by Eschbach.[7]

The halftoning process can be understood by examination of Figure 3.14. In the top of this illustration two types of functions are plotted against distance x, which has been marked off into increments one pixel in width. The first functions are three uniform reflectance levels, Rl, R2, and R3. The second function $T(x)$ is a plot of threshold versus distance, which looks like a series of up and down staircases, that produces the halftone pattern. Any pixels whose reflectance is equal to or above the threshold is turned on, and any that is below the threshold for that pixel is turned off.

Also sketched in Figure 3.14 are the results for the thresholding process for R1 on the second line and then for R2 and R3 on the third line. The last two are indistinguishable for this particular set of thresholding curves. It can be seen from this that the reflectance information is changed into width information and thus that the method of halftoning is a mechanism for creating dot growth or spatial pulse-width modulation over an area of several pixels. Typically, such threshold patterns (i.e., screens) are laid out two-dimensionally. An example is shown in Figure 3.15.

This thresholding scheme emulates the printer's 45° screen angle, which is considered to be favorable from a visual standpoint because the 45° screen is less visible (oblique effect[4]) than the same 90° screen. Other screen angles may also be conveniently generated by a single string of thresholds and a shift factor that varies from raster to raster.[62,63] The numbers in each cell in the matrix represent the threshold required in a 32-gray-level system to turn the system on or off. The sequence of thresholds is referred to as the dot growth pattern. At the bottom, four thresholded halftone dots (Parts b–e) are shown for illustration. There are a total of 64 pixels in the array but only 32 unique levels. This screen can be represented by 32 values in a 4 × 8 pixel array plus a shift factor of four pixels for the lower set of 32, which enables the 45° screen appearance as illustrated. It may also be represented by 64 values in a single 8 × 8 pixel array, but this would be a 90° screen. It is also possible to alternate the thresholding sequence between the two 4 × 8 arrays, where the growth pattern in each array is most commonly in a spiral pattern, resulting in two unique sets of

Analysis of the halftone process
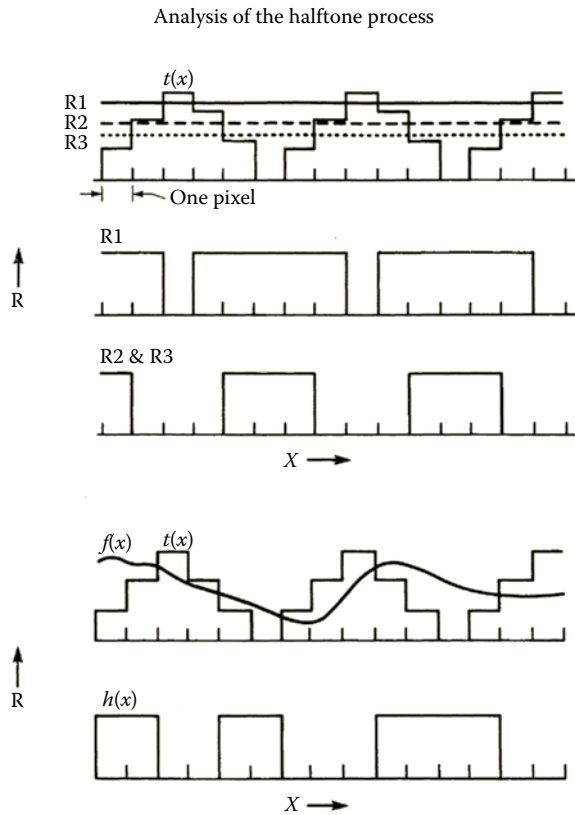
FIGURE 3.14
Illustration of halftoning process. Each graph is a plot of reflectance *R* versus distance *X*. *T(x)* is the profile of one raster of the halftone threshold pattern, where image values above the pattern are turned on (creates black in system shown) by the halftone thresholding process. Rl, R2, and R3 represent three uniform images of different average reflectances shown at the top as uniform input and in the middle of the chart as profiles of halftone dots after halftone thresholding. *f(x)* represents an image of varying input reflectance and *t(x)* is a different threshold pattern. *h(x)* is the resulting halftone dot profile, with dots represented, here, as blocks of different width illustrating image variation.

32 thresholds for an equivalent of 64 different levels and preserving the screen frequency as shown. This screen is called a "double dot." The concept is sometimes extended to four unique dot growth patterns and hence is named a "quad dot." Certain percent area coverage dot patterns in these complex multicentered dot structures generate very visible and often objectionable patterns.

The halftone matrix described in Figure 3.15 represented 32 specific thresholds in a specific layout. There are many alternatives to the size and shape of the matrix, the levels chosen, the spatial sequence in which the thresholds occur, and arrangements of multiple, uniquely different matrices in a grouping called a super cell. Here there are many different cells (more than the four in a quad dot) varying slightly in shape and each may contain a slightly different number of pixels. This gives its designer even more gray levels since there are more cells and each may contain unique thresholds. There are also more available angles due to the size and shape differences of the individual cells giving the centers of the collection of all the supercells more precision to form a new screen angle. See Figure 3.16.

	Q1				Q2		
13	6	7	14	20	27	26	19
12	1	2	8	28	32	31	25
5	4	3	9	21	29	30	24
16	11	10	15	17	22	23	18
20	27	26	19	13	6	7	14
28	323	31	25	12	1	2	8
21	29	30	24	5	4	3	9
17	22	23	18	16	11	10	15

Q3   Q4

(a)

(b)   (c)   (d)   (e)
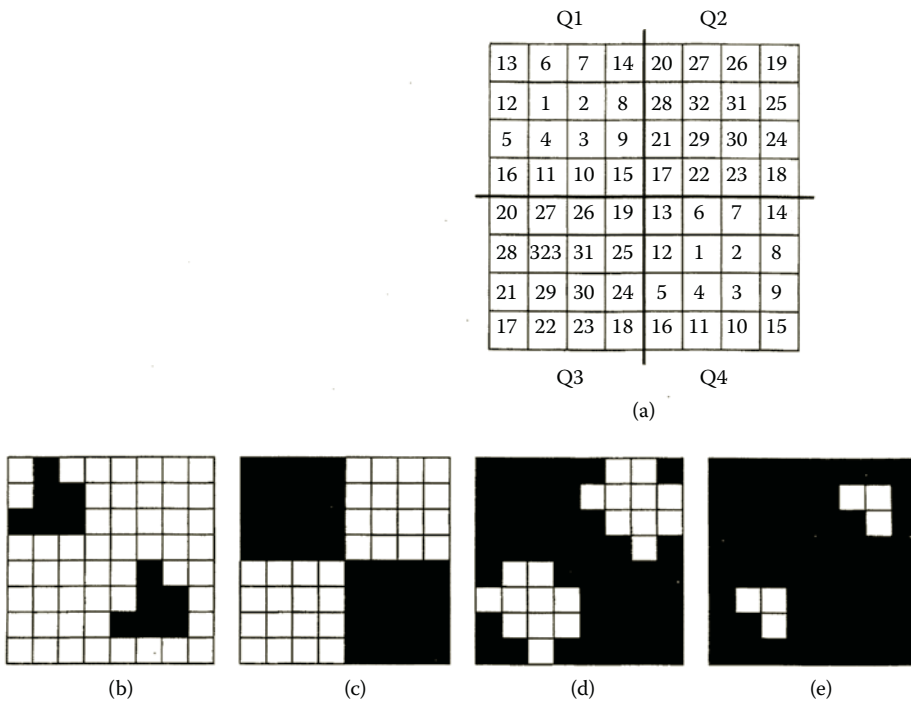
FIGURE 3.15

Example of two-dimensional quantized halftone pattern, with illustrations of resulting halftone dots at various density levels. (a) 8 × 8 spiral halftone matrix; (b) density = 0.10 or 20% fill 12/64 pixels; (c) density = 0.30 or 50% fill 32/64 pixels (d) density = 0.50 or 68% fill 44/64 pixels; (e) density = 1.00 or 90% fill 58/64 pixels.

The careful selection of these factors gives good control over the shape of the apparent tone reproduction curve, granularity, textures, and sharpness in an image. The halftone system's ability to resolve structures finer than the halftone screen array or cell size has been described as "partial dotting" by Roetling[64] and others and is an important and often misunderstood factor in image quality studies (Reference 53, p. 163; Reference 65, p. 403). It is the result of the high-resolution pixel-by-pixel comparison of the threshold matrix and the image detail which allows high-contrast image detail to pass through the halftone matrix, nearly unchanged.

There are also many other methods for converting binary images into pseudogray images using digital halftoning methods of a more complex form.[66,67] These include alternative dot structures, that is, different patterns of sequences in alternating repeat patterns, random halftoning, and techniques known as error diffusion. In his book *Digital Halftoning,* Ulichney[68] describes five general categories of halftoning techniques:

1. Dithering with white noise (including mezzotint)
2. Clustered dot ordered dither
3. Dispersed dot ordered dither (including "Bayer's dither")
4. Ordered dither on asymmetric grids
5. Dithering with blue noise (actually error diffusion)
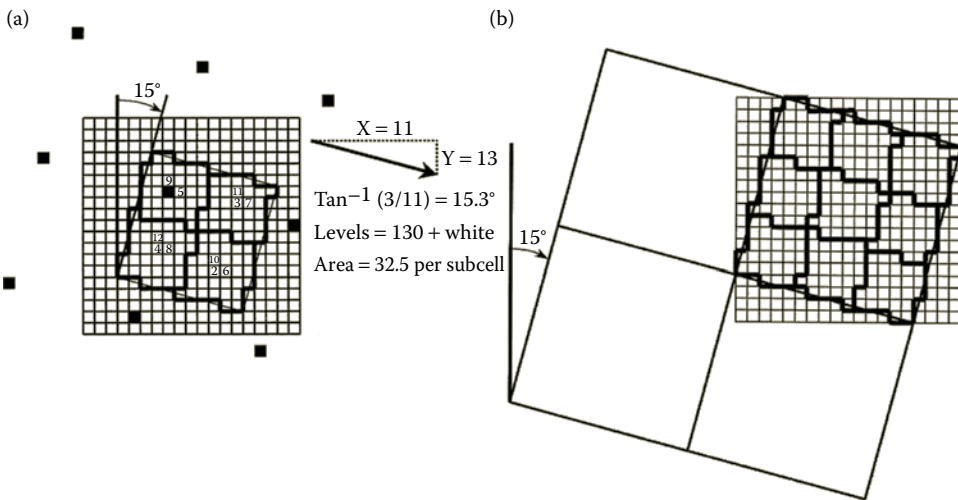
FIGURE 3.16
Examples of multicentered dots: (a) a classic quad dot showing the first three thresholds in each individual cell and the large black dots showing the repeat pattern centers at 15.255° and (b) a nine center "supercell" where the cell shape and size varies: from L to R 26, 27,27; 27,29,25; 27,27,26 pixels and the angle is 14.9°. (Reproduced with permission of the publisher, CRC Press, Boca Raton, FL, from Reference 16, p. 412 in Chapter 6 by Haines, Wang and Knox, 2003.)

He states that "spatial dithering is another name often given to the concept of digital halftoning. It is perfectly equivalent, and refers to any algorithmic process that creates the illusion of continuous tone images from the judicious arrangement of binary picture elements." The process described in Figures 3.15 and 3.16 falls into the category of a clustered dot ordered dither method (category 2) as a classical rectangular grid on a 45° base.

There is no universally best technique among these. Each has its own strengths and weaknesses in different applications. The reader is cautioned that there are many important aspects of the general halftoning process that could not be covered here. (See Reference 45 for a summary of digital halftoning technology and many references, and Reference 69 for many practical aspects of conventional halftoning for color reproduction.) For example, the densities described in Figure 3.15 only apply to the case of perfect reproduction of the illustrated pixel maps on nonlight-scattering material using perfect, totally black inks. In reality, each pattern of pixels must be individually calibrated for any given marking process. The spatial distribution interacts with various noise and blurring characteristics of output systems to render the mathematics of counting pixels to determine precise density relationships highly erroneous under most conditions. This is even true for the use of halftoning in conventional lithographic processes, due to the scattering of light in white paper and the optical interaction of ink and paper. These affect the way the input scanner "sees" a lithographic halftone original. Some of these relationships have been addressed in the literature, both in a correction factor sense[69,70] and in a spatial frequency sense.[71–73] All of these methods involve various ways of calculating the effect that lateral light scattering through the paper has on the light reemerging from the paper between the dots.

The effects of blur from the writing and marking processes involved in generating the halftone, many of which may be asymmetric, require individual density calibrations for each of the dot patterns and each of the dithering methods that can be used to generate these halftone patterns. The control afforded through the digital halftoning process by the

careful selection of these patterns and methods enables the creation of any desired shape for the tone reproduction curve for a given picture, marking process, or application.

3.2.2.4 Noise

Noise can take on many forms in an electronic imaging system. First there is the noise inherent in the digital process. This is generally referred to as either sampling noise associated with the location of the pixels or quantization noise associated with the number of discrete levels. Examples of both have been considered in the earlier discussion. Next there is electronic noise associated with the electronic components from the sensor to the amplification and correction circuits. As we move through the system, the digital components are generally thought to be error-free and therefore there is usually no such noise associated with them.

Next, in a typical electronic system, we find the ROS itself, often a laser beam scanner. If the system is writing a binary file, then the noise associated with this subsystem is generally connected with pointing of the beam at the imaging material and is described as jitter, pixel placement error, or raster distortion of some form (see the next subsection). Under certain circumstances, exposure variation produces noise, even in a binary process. For systems with gray information, there is also the possibility that the signals driving the modulation of exposure may be in error, so that the ROS can also generate noise similar to that of granularity in photographs or streaks if the error occurs repeatedly in one orientation. Finally we come to the marking process, which converts the laser exposure from the ROS into a visible signal. Marking process noise, which generally occurs as a result of the discrete and random nature of the marking particles, generates granularity.

An electronic imaging system may enhance or attenuate the noise generated earlier in the process. Systems that tend to enhance detail with various types of filters or adaptive schemes are also likely to enhance noise. There are, however, processes (see Section 3.7.2) that search through the digital image identifying errors and substitute an error-free pattern for the one that shows a mistake.[74,55] These are sometimes referred to as noise removal filters.

Noise may be characterized in many different ways, but in general it is some form of statistical distribution of the errors that occur when an error-free input signal is sent into the system. In the case of imaging systems, an error-free signal is one that is absolutely uniform, given a noise-free, uniform input. Examples would include a sheet of white microscopically uniform paper on the platen of an input scanner, or a uniform series of laser-on pulses to a laser beam scanner, or a uniform raster pattern out of a perfect laser beam scanner writing onto the light-sensitive material in a particular marking device. A typical way to measure noise for these systems would be to evaluate the standard deviation of the output signal in whatever units characterize it. A slightly more complete analysis would break this down into a spatial frequency or time-frequency distribution of fluctuations. For example, in a photographic film a uniform exposure would be used to generate images whose granularity was measured as the root-mean-square fluctuation of density. For a laser beam scanner it would be the root-mean-square fluctuation in radiance at the pixel level for all raster lines.

In general, certain factors that affect the signal aspect of an imaging system positively, affect the noise characteristics of that imaging system negatively. For example, in scanning photographic film, the larger the sampled area, as in the case of the microdensitometer aperture, the lower the granularity [Equation 3.1]. At the same time, the image information is more blurred, therefore producing a lower contrast and smaller signal level. In general the

signal level increases with aperture area and the noise level (as measured by the standard deviation of that signal level) decreases linearly with the square root of the aperture area or the linear dimension of a square aperture. It is therefore very important when designing a scanning system to understand whether the image information is being noise limited by some fundamentals associated with the input document or test object or by some other component in the overall system itself. An attempt to improve bandwidth, or otherwise refine the signal, by enhancing some parts of the system may, in general, do nothing to improve the overall image information, if it is noise in the input that is limiting and that is being equally "enhanced." Also, if the noise in the output writing material is limiting, then improvements upstream in the system may reach a point of diminishing returns.

In designing an overall electronic imaging system it should be kept in mind that noises add throughout the system, generally in the sense of an RSS (root of the sum of the squares) calculation. The signal attenuating and amplifying aspects, on the other hand, tend to multiply throughout the system. If the output of one subsystem becomes the input of another subsystem, the noise in the former is treated as if it were a signal in the latter. This means that noise in the individual elements must be appropriately mapped from one system to the other, taking into account various amplifications and nonlinearities. In a complex system this may not be easy; however, keeping an accurate accounting of noise can be a great advantage in diagnosing the final overall image quality. We expand on the quantitative characterization of these various forms of signal and noise in the subsequent parts of this chapter.

3.2.2.5 Color Imaging

Color imaging in general and especially digital color imaging have received considerable attention in the literature in recent years.[5,6,14,29,30] An elementary treatment is given below covering a few major points important to scanning and image quality. See References 16 or 30 for a recent broad overview and literature survey of digital color imaging, and Reference 75 for a classic review of more traditional color reproduction systems and colorimetry.

3.2.2.5.1 Fundamentals

There are two basic methods of creating images, including digital images, in color, called additive and subtractive methods.

In an *additive color* system one creates the appropriate color image pixels by combining red (R), green (G), or blue (B) microsized lights, that is, pixels of varying intensities. Roughly equal amounts of each produce the sensation of "white" light on viewing. This applies to many self-luminous displays such as a CRT/TV or liquid crystal displays. The pixels must be small enough that the eye blurs them together. The eye detects these signals using sensors called "cones" in the retina. These are associated with the HVS sensations of red, green, and blue.

In the second method of color imaging, called *subtractive color,* light is removed from otherwise white light by filters that subtract the above components one at a time. Red is removed by a cyan (C) filter, green by a magenta (M) filter, and blue by a yellow (Y) filter. For an imaging system, these filters are created by an imagewise distribution of transparent colorants created (e.g., pixel by pixel) in varying amounts. They are laid down color layer by color layer. The "white" light may come from a projector as in the case of transparencies or from white room light reflected by a white sheet of paper with the imagewise distribution of transparent colorants bonded to it. Here the subtraction occurs once on the way to the paper and then a second time after reflection on the way to the eye. Color photographic reflection prints and color offset halftone printing both use this method.

A digital color imaging system, designed to capture the colors of an original object, breaks down light reflected (or transmitted) from the object into its R, G, and B components by a variety of possible methods. It uses separate red, green, and blue image capture systems and channels of image processing, which are eventually combined to form a full color image.

The visual response involves far more than just the absorption of light. It involves the human neurological system and many special processes in the brain. The complexity of this can be appreciated by observing the results of simple color matching experiments, in which an observer adjusts the intensities of three color primaries until their mixture appears to match a test color. Such experiments, using monochromatic test colors, lead to the development of a set of color matching functions for specific sets of colored light sources and specific observer conditions. Certain monochromatic colors require the subtraction of colored light (addition of the light to the color being matched) in order to create a match. Color matching experiments are described extensively in the literature[4,10,14,29] and provide the foundation to the science of colorimetry.

Two such sets of color matching functions are shown in Figures 3.17a and b. The first set reports experimental results using narrow band monochromatic primaries. Note the large negative lobe on the third curve of "a," showing the region where "negative light" is needed, that is, where the light must be added to the color under test to produce a match. The second set has become a universally accepted representation defining the CIEs (Commission Internationale de l'Eclairage) 1931 2° Standard Colorimetric Observer. It is a linear transformation of standardized color matching data, carefully averaged over many observers and is representative of 92% of the human population having normal color vision. This set of functions provides the standardization for much of the science of the measurement of color, in other words, important colorimetry standards.

This overly simplistic description goes beyond the scope of this chapter to explain. Ideally the information recorded by a color scanner should be equivalent to that seen by an observer. In reality, the transparent colorant materials used to create images are not perfect. Significant failures stem from the nonideal shapes of the spectral sensitivities of the capturing device and the nonideal shapes of the spectral reflectance or transmittance of the colorants. Practical limitations in fabricating systems and noise also restrict the accuracy of color recording for most scanners. Ideal spectral shapes of sensitivities and filters would allow the system designer to better approximate the HVS color response. For example, an input original composed of conventional subtractive primaries such as real magenta (green absorbing) ink, not only absorbs green light, but also absorbs some blue light. Different magentas have different proportions of this unwanted absorption. Similar unwanted absorptions exist in most cyan and, to a lesser extent, in most yellow colorants. These unwanted characteristics limit the ability of complete input and output systems to reproduce the full range of natural colors accurately. Significant work has been carried out to define quality measures for evaluating the color quality of color recording instruments and scanning devices.[32, 6 (ch. 5), 21 (ch. 19). See also Table 3.9: INCITS-WI, ANSI–IT8].

3.2.2.5.2 Colorimetry and Chromaticity Diagrams

This leads to two large problem areas in color image quality needing quantification, namely: (1) that the color gamuts of real imaging systems are limited; and (2) that colors which appear to match under one set of conditions appear different by some amount under another set of circumstances. This is conveniently described by a color analysis tool from the discipline of colorimetry (the science of color measurement) called a *chromaticity diagram,* shown in Figure 3.18. It describes color in a quantitative way. It can be seen, in this
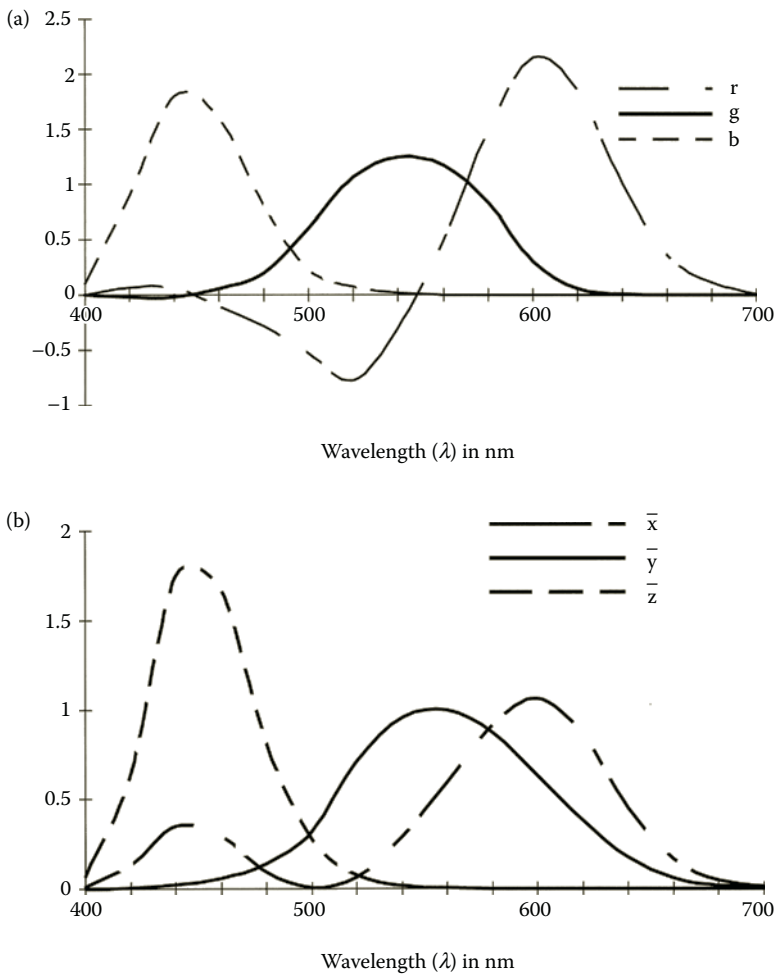
FIGURE 3.17
Color matching functions: (a) example of a directly measured result (adapted from Giorgianni, E.J.; Madden, T.E. *Digital Color Management Encoding Solutions;* Addison-Wesley: Reading, MA, 1998.); (b) a transformed result chosen as the CIE Standard Observer for 2° field of view.

illustration, that the monitor display is capable of showing different colors from a particular color printer. It is also possible, with this diagram, to show the color of an original. Note that a color gamut is the range of colors that can be produced by the device of interest as specified in some three- or more dimensional color space. It is important to note that a two-dimensional representation, like that shown here, while very helpful, is only a part of the whole three-dimensional color space. Variations derived from the chromaticity diagram, and the equations that define it, however, provide a basis for much of the literature that describes color image quality today. It is designed to facilitate description of small color differences, for example, between an original and a reproduced color or two different reproductions of the same color.

The reader must be warned, however, that the actual perception of colors involves many psychophysical and psychological factors beyond those depicted in this diagram.[4] It is, however, a useful starting point. It describes any color in an image or a source and is

often the starting point in many of the thousands of publications on color imaging. There are also many different transformations of basic chromaticity diagram, a few primary examples of which we will describe here.

For the purposes of this chapter the basic equations used to derive the chromaticity diagram and to transform it provide an introduction to color image quality measurement. The outer, horseshoe-shaped curve, known as the "spectral locus," represents the most saturated colors possible, those formed by monochromatic sources at different wavelengths. All other possible colors lie inside this locus. Whites or neutrals by definition are the least saturated colors, and lie nearer the center of the horseshoe-shaped area. The colors of selected broad-spectrum light sources A, B and C are shown later in Figure 3.47 (along with the equal energy white point E from Figure 3.18, plotted here for reference) using a more precise chromaticity diagram. Saturation (a perceptual attribute) of any color patch (transparent or reflection) can be estimated on this chart by a physical measure called excitation purity. It can be seen as the relative distance from the given illumination of the patch to the horseshoe limit curve along a vector. The dominant wavelength (approximate correlate with perceptual attribute of hue) is given by the intersection of that vector with the spectral locus. The lightness of the color is a third dimension, not shown, but is on an axis perpendicular to the plane of the diagram (coming out of the page). Use of dominant wavelength and purity to describe colors in the *x, y* version of the chromaticity diagram is shown in Figure 3.46 in Section 3.9. Different light sources may be used but standard source "C" (See Figure 3.47) was chosen here. These correlates are only approximate because lines of constant hue are slightly curved in these spaces.
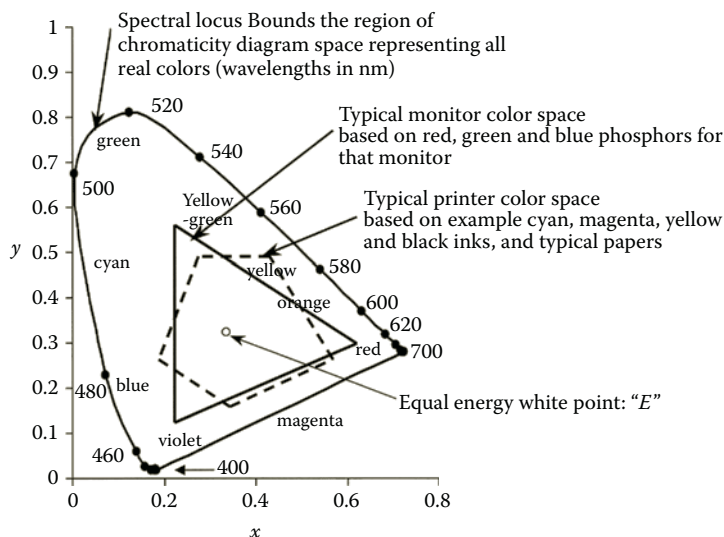
FIGURE 3.18
The *x, y* chromaticity diagram. Variations derived from it, and the equations that define it, provide a basis for much of the literature that describes color image quality today. It is designed to facilitate description of small color differences such as an original and a reproduced color or two different reproductions of the same color. Examples of the differences between possible colors at a given lightness formed in two different media, a printer and monitor, are shown (adapted from Adams, R.M.; Weisberg, J.B. *The GATF Practical Guide to Color Management;* GATF Press, Graphic Arts Technical Foundation: Pittsburgh, PA, 2000, which cites data from X-Rite Inc). A more precise chromaticity diagram is shown in Figure 3.46.

To understand the chromaticity coordinates, x and y, return to Figure 3.17b. From these curves for $\bar{x}, \bar{y}, \bar{z}$, the spectral power of the light source $S(\lambda)$, and the spectral reflectance (or transmittance) of the object $R(\lambda)$, one can calculate

$$X = k \sum_{\lambda=380}^{780} S(\lambda)R(\lambda)\bar{x}(\lambda) \tag{3.3a}$$

$$Y = k \sum_{\lambda=380}^{780} S(\lambda)R(\lambda)\bar{y}(\lambda) \tag{3.3b}$$

$$Z = k \sum_{\lambda=380}^{780} S(\lambda)R(\lambda)\bar{z}(\lambda) \tag{3.3c}$$

where k is normally selected to make $Y = 100$ when the object is a perfect white, that is, an ideal, nonfluorescent isotropic diffuser with a reflectance equal to unity throughout the visible spectrum. The spectral profile of several standard sources is given later in Figure 3.47.

These results are used to calculate the *chromaticity coordinates* in the above diagram as follows:

$$x = \frac{X}{(X + Y + Z)} \tag{3.4a}$$

$$y = \frac{Y}{(X + Y + Z)} \tag{3.4b}$$

$$z = \frac{Z}{(X + Y + Z)} \tag{3.4c}$$

One of the most popular transformations is the CIE *L*a*b** version (called CIELAB for short) which is one of most widely accepted attempts to make distances in color space more uniform in a visual sensation sense [76,203]. Here

$$L^* = 116(Y/Y_n)^{1/3} - 16 \tag{3.5}$$

which represents the achromatic lightness variable, and

$$a^* = 500\left[(X/X_n)^{1/3} - (Y/Y_n)^{1/3}\right] \tag{3.6}$$

$$b^* = 500\left[(Y/Y_n)^{1/3} - (Z/Z_n)^{1/3}\right] \tag{3.7}$$

represent the chromatic information, where X_n, Y_n, Z_n are the X, Y, Z tristimulus value of the *reference white.* Color differences are given as

$$\Delta E_{ab}^* = [(\Delta L^*)^2 + (\Delta a^*)^2 + (\Delta b^*)^2]^{1/2}. \tag{3.8}$$

In practical terms, results where $\Delta E_{ab}^* = 1$ represent approximately one just noticeable visual difference (see Section 3.8). However, the residual nonlinearity of the CIELAB chromaticity diagram, the remarkable adaptability of the human eye to many other visual factors, and the effect of experience require situation-specific experiments. Only such experiments can determine rigorous tolerance limits and specifications. Color appearance models that account for many such dependencies and nonlinearities have been developed.[4,76] Attempts to standardize the methodology have been developed by CIE TC1-34 as CIECAM97s and proposed CIECAM02. (See Appendix A of Reference 4).

For readers not familiar with conventional graphic arts, printing and photographic analysis, densitometers are widely used to characterize those imaging systems. They measure transmission or reflection density, D,

$$D_f = \text{Log10}(1/R_f) \tag{3.9}$$

where $R_f = Y/Y_{ref}$ from the above equations. It is called the reflectance (or transmittance for films and filters) factor and may be expressed as a % or decimal. The subscript "f" indicates that there are many factors such as light source optical geometries and filters that need to be specified. The "ref" indicates the measurement of Y for a white reference, one of the many factors. Since many tests and test targets used to evaluate scanners and digital cameras are derived from these disciplines, it is useful to examine the relationship between density, L^*, and reflectance as in Figure 3.19. It is seen that to a first approximation (within $.05L^*$) a
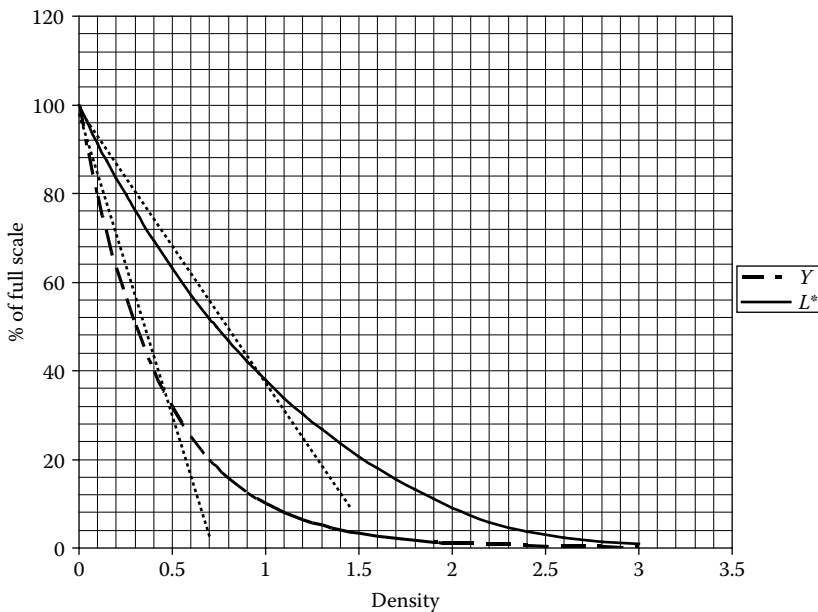
FIGURE 3.19
Reflectance (in the form of Y/Y_n as a %) and L^* as a function of density (See Table 3.8 in the appendix for values of all three) Dotted lines show linear approximations visually fit to each curve, anchored at the maximum value.

straight line approximates the density–*L** relationship to a density of about 1.2 which is a useful range for many image quality measurements, representing the eye's response better than the reflectance factor which is only approximately linear to density of 0.5. An abridged conversion table listing the corresponding values is given in the appendix as Table 3.8.

Another very important color description tool is the Munsell system in which painted paper chips of different colors have been arranged in a three-dimensional cylindrical coordinate system. The vertical axis represents *value* (akin to lightness), the radius represents *chroma,* and the angular position around the perimeter is called *hue.* These have been carefully standardized and are very popular as color references.[76]

3.3 PRACTICAL CONSIDERATIONS

Several overall systems design issues are of some practical concern, including the choice of scan frequency as well as motion errors and other nonuniformities. They will be addressed here in fairly general terms.

3.3.1 Scan Frequency Effects

As digital imaging evolved in the previous decade, it had generally been thought that the spatial frequency, in raster lines or pixels per inch, which is used either to create the output print or to capture the input document, is a major determinant of image quality. Today there is a huge range of scan frequencies emanating from a huge range of products and applications from low-end digital cameras and fax machines, through office scanners and copiers, to high-end graphic arts scanners, all used with a plethora of software and hardware image processing systems that enlarge and reduce and interpolate the originally captured pixel spacings to something else. Then, other systems with yet additional processing and imaging affects are employed to render the image prior to the human reacting to the quality. It is only at this point in the process, where all the signal and noise effects roll up that the underlying principles from other parts of this chapter can be used to quantify overall image quality. Needless to say, scan frequency or pixel density is only one of these effects, and to assert it is *the* dominant effect is questionable in all but the most restrictive of circumstances. Yet it is an important factor and many type A shortcut experiments have attempted to address the connection between the technology variable of pixel density and various dimensions of overall image quality.

If a scanned imaging system is designed so that the input scanning is not aliased and the output reconstruction faithfully prints all of the information presented to it, then the scan frequency tends to determine the blur, which largely controls the overall image quality in the system. This is frequently not the case, and, as a result, scan frequency is not a unique determinant of image quality. In general, however, real systems have a spread function or blur that is roughly equivalent to the sample spacing, meaning they are somewhat aliased and that blur correlates with spacing. However, it is possible to have a large spot and much smaller spaces (i.e., unaliased), or vice versa (very aliased). The careful optimization of the other factors at a given scan frequency may have a great deal more influence on the information capacity of any electronic imaging system and therefore on the image-quality performance than does scan frequency itself. To a certain extent, gray information can be readily exchanged for scan frequency. We shall subsequently explore this further when dealing with the subject of information content of an imaging system.

In the spirit of taking a snapshot of this huge and complex subject, Figure 3.20 summarizes three types of practical findings, two about major applications of scanned or digital images, namely digital photography and graphic arts–digital reprographics, and one simplification of human perception. The curves in the lower graph (solid dots) show results of two customer acceptability experiments with digital photography, varying camera resolution and printing on 8 bpp contone printers (Al from Reference 77, A2 from Reference 78). Experiments on digital reprographics are shown by the curves with the open symbols, which suggest acceptable enlargement factors for input documents scanned at various resolutions and printed at various output screen resolutions. Finally we can put this in perspective by noting, as triangles along the frequency axis, the resolution limitations of the HVS at normal and close inspection viewing distances using modest 6% and very sensitive 1% contrast detection thresholds. Returning to Figure 3.3, both applications are type A methods, while the HVS limits were inferred from visual algorithms.

A fairly general practice is to design aliased systems in order to achieve the least blur for a given scan frequency. Therefore, another major effect of scan frequency concerns the interaction between periodic structures in the input and the scanning frequency of the system that is recording the input information. These two interfere, producing beat patterns
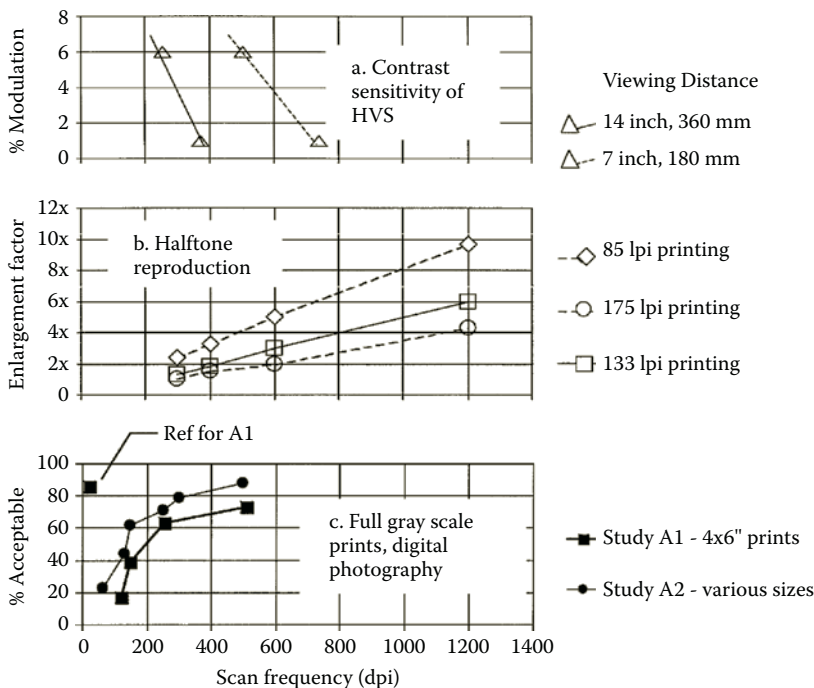
FIGURE 3.20
Summary of practical findings about sampling frequency in graphic arts (middle chart) for halftones (adapted from Cost, F. *Pocket Guide to Digital Printing;* Delmar Publishers: Albany, NY, 1997.) in digital photography (bottom chart, adapted from Miller, M.; Segur, R. Perceived IQ and acceptability of photographic prints originating from different resolution digital capture devices. *Proceedings of IS&T Image Processing, Image Quality, Image Capture Systems (PICS) Conference, Savannah, GA,* 1999; 131–137 and Daniels, C.M., Ptucha, R.W., Schaefer, L. The necessary resolution to zoom and crop hardcopy images, *Proceedings of IS & T Image Processing. Image Quality, Image Capture Systems (PICS) conference,* Savannah, Georgia, 143, 1999) and related HVS contrast sensitivity (top chart) reference values. (Adapted from Fairchild, M.D. *Color Appearance Models;* Addison-Wesley: Reading, MA, 1998.)

at sum and difference frequencies leading to the general subject of moiré phenomena. Hence, small changes in scan frequency can have a large effect on moiré.

One of the major considerations in selecting output scan frequency is the number of gray levels required from a given halftone screen. Recall the discussion of Figure 3.15 dot matrices from 4 × 4 to 12 × 12 are shown in Table 3.1 at a range of printing frequencies from 200 to 1200 raster lines per inch (7.87–47.2 raster lines/mm). For example, a 10 × 10 matrix of thresholds (on top row) can be used to generate a 51 gray level, 45° angle screen (two shifted 5 × 10 submatrices) as in Figure 3.15 but with different thresholds in each. Its screen frequencies are shown in the ninth column at the eight different printer scan frequencies and varies from 28 to 170 halftone dots/inch. Also indicated in the table between the bold lines is the approximate useful range for the visual system. The range starts at a lower limit of approx. 65 dots/in (2.56 dots/mm) halftone screen, formerly found in newspapers. This results in noticeably coarse halftones and has recently moved into the range of 85 dots/in (3.35 dots/mm) to 110 dots/in (4.33 dots/mm) in modern newspapers. The upper bound represents a

TABLE 3.1

Relationship among Halftone Matrix Size (Given in Pixels), Maximum Possible Number of Gray Levels in the Halftone, and Output Scan Frequency (in Pixels/Inch)

Matrix in pixels	4 × 4	3 × 3	4 × 4	6 × 6	5 × 5	8 × 8	6 × 6	10 × 10	12 × 12
Angle	45°	90°	90°	45°	90°	45°	90°	45°	45°
No of gray levels—type[a]: Conventional	9 - A	10 - B	17 - B	19 - A	26 - B	33 - A[§]	37 - B	51 - A	73 - A
Expanded[b] 2x	17 - C	19 - D	33 - D	37 - C	51 - D	65 - C	73 - D	101 - C	145 - C
Expanded[b] 4x	33	41 – E	65 – E	73	101 – E	129	145 – E	201	289
Scan freq. in pixels/in ↓ 1200	426	400	300	282	240	212	200	170	141
1000	352	333	250	236	200	176	167	142	118
800	284	267	200	188	160	142	133	114	94
600	212	200	150	141	120	106	100	85	71
500	176	166	125	118	100	88	84	71	59
400	142	133	100	94	80	71	67	57	47
300	106	100	75	71	60	53	50	43	36
200	71	67	50	47	40	35	33	28	24

(Annotations: Practical upper limit (~200 LPI); Typical hafftton (~100 LPI); Halftone dot, practice lower limit (~65 LPI))

Entries are given in halftone dots/inch measured along the primary angle (row 2) of the halftone pattern. Dot types are given as (See quadrants of Figure 3.15): (A) the conventional 45° halftone where quadrants Q1 = Q4, Q2 = Q3; (B) conventional 90° halftone where Q1 = Q2 = Q3 = Q4. Expansions of the number of halftone gray levels show three new types: (C) = Type A except Q3 and Q4 thresholds are set at halfway between those in Q1 and Q2 (45° double dot), (D) where Q1 = Q4, but Q2 and Q3 thresholds are set halfway between those in Q1 (90° double dot); (E) where Q1 through Q4 thresholds are each set to generate intermediate levels among each other (90° quad dot). The number of gray values includes one level for white.

[a] Type refers to specific halftone structures A–E (see caption) where appropriate.

[b] Number of gray levels is increased by 2× or 4× over conventional by gray pixels or multicentered dots.

[§] Example shown in Figure 3.15.

materials limit of around 175 dots/in (6.89 dots/mm), which is a practical limit for many lithographic processes. The number of gray levels is shown in the third, fourth, and fifth rows. Conventional dots are single centered like Figure 3.15. Increased number of gray levels for double dots is shown as the "expanded by 2×" row and quad dots (four centers) by the row labeled as "expanded by 4×" are indicated by the row. This table assumes that the pixels are binary in nature. If a partially gray or high addressability output imaging system is employed then the number of levels in the table must be multiplied by the number of gray levels or subpixels per pixel appropriate to the technology. The use of these techniques and supercells to expand the gray level resolution has increased in recent years as real time microprocessing has enabled reasonable speed and memory for such approaches.

3.3.2 Placement Errors or Motion Defects

Since the basic mode of operation for most scanning systems is to move or scan rapidly in one direction and slowly in the other, there is always the possibility of an error in motion or other effect that results in locating pixels in places other than those intended. Figure 3.21 shows several examples of periodic raster separation errors, including both a sinusoidal and a sawtooth distribution of the error. These are illustrated at 300 raster lines/in (11.8 lines/ mm) with ±10 through ±40 μm (±0.4 through ±1.6 mils) of spacing error, which refers to the local raster line spacing and not to the error in absolute placement accuracy. Error frequencies of 0.33 cycles/mm (8.4 cycles/in) and 0.1 cycles/mm (2.5 cycles/in) are illustrated.

For input scanners, which convert an analog signal to a digital one, the error takes the form of a change in the sampling of the analog document. Since sampling makes many mistakes,
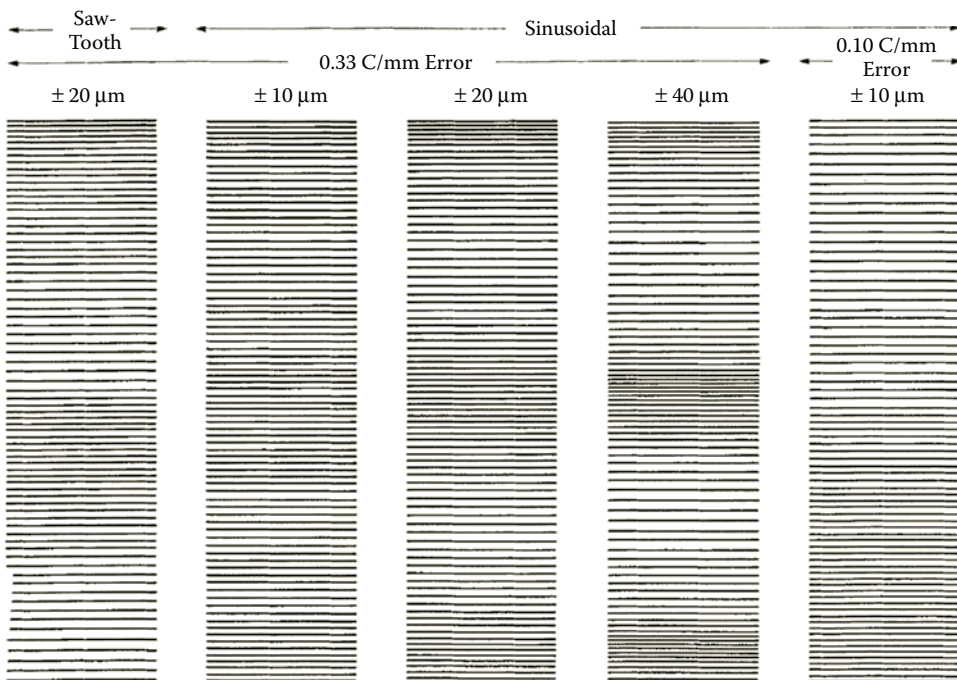
FIGURE 3.21
Enlarged examples of rasters with specified image motion variation at 300 dpi.

the sampling errors due to motion nonuniformity are most visible in situations where the intrinsic sampling error is made to appear repeatable or uniform, and the motion error, therefore, appears as an irregular change to an otherwise uniform pattern. Long angled lines that are parallel to each other provide such a condition because each line has a regular periodic phase error associated with it, and a motion error would appear as a change to this regular pattern. Halftones that produce moiré are another example, except that the moiré pattern is itself usually objectionable so that a change in it is not often significant.

In patterns with random phase errors such as text, the detection of motion errors is more difficult. Effects that are large enough to cause a two-pixel error would be perceived very easily; however, effects that produce less than one-pixel error on average would tend to increase the phase errors and noise in the image generally and would therefore be perceived on a statistical basis. Many identical patterns repeated throughout a document would provide the opportunity to see the smaller errors as being correlated along the length of the given raster line that has been erroneously displaced, and would therefore increase the probability of seeing the small errors.
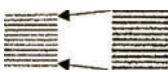
Motion errors in an output scanner that writes on some form of image-recording material can produce several kinds of defects. In Table 3.2, several attributes of the different types of raster distortion observables are shown. The first row in the matrix describes the general kind of error, that is, whether it is predominantly a pixel placement error or predominantly a developable exposure effect or some combination of the two. The second row is a brief word description or name of the effect that appears on the print. The third row describes the spatial frequency region in cycles/mm in which this type of error tends to occur. The next row indicates whether the effect is best described and modeled as one-dimensional or two-dimensional. Finally, a graphical representation of an image with the specific defect is shown in the top row, while the same image appears in the bottom row without the defect.

The first of the columns on the left is meant to show that if the frequency of the error is low enough then the effect is to change the local magnification. A pattern or some form of texture that should appear to have uniform spacings would appear to have nonuniform spacings and possibly the magnification of one part of the image would be different from that of another. The second column is the same type of effect except the frequency is much higher, being around 1 cycle/mm (25 cycle/in). This effect can then change the shape of a character, particularly one with angled lines in it, as demonstrated by the letter *Y*.

Moving to the three right-hand columns, which are labeled as developable exposure effects, we have three distinctly different frequency bands. The nature and severity of these effects depend in part on whether we are using a "write white" or "write black" recording system and on the contrast or gradient of the recording material. The first of these effects is labeled as structured background. When the separation between raster lines increases and decreases, the exposure in the region between the raster lines where the Gaussian profile writing beams overlap increases or decreases with the change. This gives an overall increase or decrease in exposure, with an extra large increase or decrease in the overlap region. Since many documents that are being created with a laser beam scanner have relatively uniform areas, this change in exposure in local areas gives rise to nonuniformities in the appearance in the output image.

In laser printers, for example, the text is generally presented against a uniform white background. In a positive "write white" electrophotographic process, such as is used in many large xerographic printers, this background is ideally composed of a distribution of uniformly spaced raster lines that expose the photoreceptor so that it discharges to a level where it is no longer developable. As the spacing between the raster lines increases, the exposure between them decreases to a point where it no longer adequately discharges the

TABLE 3.2

The Effects of Motion Irregularities, Defects, or Errors on the Appearance of Scanned Images

Error Category	Pixel Placement Error		Combination of Both			Developable Exposure Effects	
Effect on print	Spacing Nonuniformity	Character Distortion/Fast Scan Jitter	Halftone Nonuniformity	Line Darkness Nonuniformity	Structured Background	Ragged/Structured Edges	Unsharp Line-text-edge Images
Typical frequencies of motion error (c/mm)	<0.5	0.5–2	0.1–6	0.005–2	0.005–8	1–8	4–20+
Dimensions of effect	1-D	2-D	2-D (1-D)	1-D (2-D)	1-D	2-D	2-D
Scanner type	input/output	input/output	input/output	input/output	output	output	output
Example with little or no defect needs space in "orno"							
Comment	Shows local reduction, magnification is also possible		Illustrates 2 levels, bottom case has very small defect		Shows "write white" system	Shows strong low & high freq.—approx 3 × 3 mm sample	Examples: higher magnification than at left eye blurs structures shown

photoreceptor, thereby enabling some weak development fields to attract toner and produce faint lines on a page of output copy. For this reason among others, some laser printers use a reversal or negative "write black" form of electrophotography in which black (no light) output results in a white image. Therefore, white background does not show any variation due to exposure defects, but solid dark patches often do.

The allowable amplitude for these exposure variations can be derived from minimum visually perceivable modulation values and the gradient of the image-recording process.[79,80] In the spatial frequency region near 0.5 cycles/mm (13 cycles/in), where the eye has its peak response at normal viewing distance, an exposure modulation of 0.004–0.001 $\Delta E/E$ has been shown to be a reasonable goal for a color photographic system with tonal reproduction density gradients of 1–4.[81]

If the frequency of the perturbation is of the order 1–8 cycles/mm (25–200 cycles/in), and especially if the edges of the characters are slightly blurred, it is possible for the nonuniform raster pattern to change the exposure in the partially exposed blurred region around the characters. As a result, nonuniform development appears on the edge and the raggedness increases as shown by the jagged appearance of the wavy lines in column 7. The effects are noticeable because of the excursions produced by the changes in exposure from the separated raster lines at the edges of even a single isolated character. The effect is all the more noticeable in this case because the darkened raster lines growing from each side of the white space finally merge in a few places. The illustration here, of course, is a highly magnified version of just a few dozen raster lines and the image contained within them.

In the last column we see small high-frequency perturbations on the edge, which would make the edge appear less sharp. Notice that structured background is largely a one-dimensional problem, just dealing with the separation of the raster lines, while character distortion, ragged or structured edges, and unsharp images are two-dimensional effects showing up dramatically on angled lines and fine detail. In many cases the latter require two dimensions to describe the size of the effect and its visual appearance.

Visually apparent darkness for lines in alphanumeric character printing can be approximately described as the product of the maximum density of the lines in the character times their widths. It is a well-known fact in many high-contrast imaging situations that exposure changes lead to line width changes. If the separation between two raster lines is increased, the average exposure in that region decreases and the overall density in a write white system increases. Thus, two main effects operate to change the line darkness. First, the raster information carrying the description of the width of the line separates, writing an actually wider pattern. Secondly, the exposure level decreases, causing a further growth in the line width and to some extent causing greater development, that is, more density. The inverse is true in regions where the raster lines become closer together. Exposure increases and linewidth decreases.

If these effects occur between different strokes within a character or between nearby characters, the overall effect is a change in the local darkness of text. The eye is generally very sensitive to differences of line darkness within a few characters of each other and even within several inches of each other. This means that the spatial frequency range over which this combination of stretching and exposure effect can create visual differences is very large, hence the range of 0.005–2 cycles/mm (0.127–50 cycles/in). Frequencies listed in Table 3.2 cover a wide range of effects, also including some variation of viewing distance. They are not intended as hard boundaries but rather to indicate approximate ranges.

Halftone nonuniformity follows from the same general description given for line darkness nonuniformity except that we are now dealing with dots. The basic effect, however, must occur in such a way as to affect the overall appearance of darkness of the small region of an otherwise uniform image. A halftone works on the principle of changing a

certain fractional area coverage of the halftone cell. If the spatial frequency range of this nonuniformity is sufficiently low, then the cell size changes at the same rate that the width of the dark dot within the cell changes. Therefore the overall effect is to have no change in the percent area coverage and only a very small change in the spacing between the dots. Hence, the region of a few tenths to several cycles/mm (several to tens of cycles/in) is the domain for this artifact. It appears as stripes in the halftone image.

The allowable levels for the effects of pixel placement errors on spacing nonuniformity and character distortion depend to a large extent upon the application. In addition to application sensitivity, the effects that are developable or partially developable are highly dependent upon the shape of the profile of the writing spot and upon amplification or attenuation in the marking system that is responding to the effects. Marking systems also tend to blur out the effects and add noise, masking them to a certain extent.

3.3.3 Other Nonuniformities

There are several other important sources of nonuniformity in a raster scanning system. First, there is a pixel-to-pixel or raster-to-raster line nonuniformity of either response in the case of an input scanner or output exposure in the case of an output scanner. These generally appear as streaks in an image when the recording or display medium is sensitive to exposure variations. These, for example, would be light or dark streaks in a printed halftone or darker and lighter streaks in a gray recorded image from an input scanner looking at a uniform area of an input document. A common example of this problem in a rotating polygon output scanner is the effect of facet-to-facet reflectivity variations in the polygonal mirror itself. The exposure tolerances described for motion errors above also apply here.

Another form of nonuniformity is sometimes referred to as *jitter* and occurs when the raster synchronization from one raster line to another tends to fail. In these cases a line drawn parallel to the slow scan direction appears to oscillate or jump in the direction of the fast scan. These effects, if large, are extremely objectionable. They will manifest themselves as raggedness effects or as unusual structural effects in the image, depending upon the document, the application, and the magnitude and spatial frequency of the effect.

3.3.3.1 *Perception of Periodic Nonuniformities in Color Separation Images*

Research on the visibility of periodic variations in the lightness of 30% halftone tints of cyan, magenta, yellow, and black color image separations printed on paper substrates has been translated into a series of guidelines for a specification for a high-quality color print engine.[82] (Figure 3.22). They were chosen to be slightly above the onset of visibility. Specifically they are set at $\{[1/3] \times [(2 \times$ "visible but subtle threshold") + ("obvious threshold")]\}$ and adjusted for a wider range of viewing distances and angles than during the experiments, which were at 38–45 cm. These guidelines are given in terms of colorimetric lightness units on the output prints. Visibility specifications must ultimately be translated into engineering parameters. We have selected the traditional CIE $L^*a^*b^*$ metrics version for illustration. These also tend to shows the smallest, most demanding ΔEs. Guidelines developed in terms of ΔE for other color difference metrics (CMC 2:1 and CIE-94) have also been developed,[82] and show different visual magnitudes, by as much as a factor of 2.

To translate these into a guidelines for the approximate optical scanner exposure variation, the ΔE values in this graph must be divided by the slope of the system response curve, in terms of $\Delta E/\Delta exposure$, for the color separation of interest. Exposure, *H*, is the general variable of interest since it is the integrated effect of intensity and time variations, both
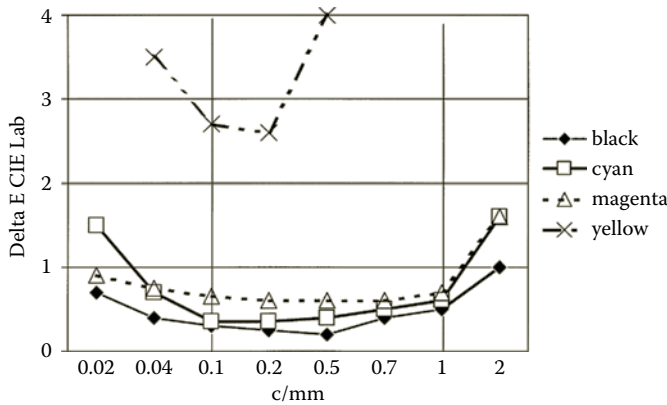
FIGURE 3.22

Guidelines for specification of periodic nonuniformities in black, cyan, magenta, and yellow color separations as indicated, in terms of ΔE derived from CIE *L*a*b*, plotted against the effective spatial frequency of the periodic disturbance. (Adapted from Goodman, N.B. Perception of spatial color variation caused by mass variations about single separations. *Proceedings of IS&T's NIP14: International Conference on Digital Printing Technologies, Toronto, Ontario, Canada,* 1998; 556–559.)

of which can result from the scanner errors discussed in the pages above. The system response would be approximated by the cascaded (multiplied) slopes of the responses of all the intermediate imaging systems between the scanner and the resulting imaging media assuming the small signal theory approximation to linearity of the cascaded systems. In the particular case where a system is positive working and linear such that, $\Delta R = \Delta H$, then taking the derivative of Equation 3.5 gives

$$\Delta E = \Delta L^*. \tag{3.10}$$

For $Y/Y_n = R = 0.70$. R and H are decimal output reflectance and normalized exposure values respectively (i.e., both full scale of 1 and a minimum of zero) R = 0.70 is the reflectance for a 30% halftone used above. For $\Delta E = 0.2$ this yields $\Delta R = 0.0041$ which also $= \Delta H$, indeed a very small exposure value yielding visible errors.

Specific relationships for exposure and reflectance, and for any of the other colorimetric units described in this research should be developed for each real system. The linear gain = 1.0 assumption shown here should not be taken for granted. The reader is also reminded that these results are for purely sinusoidal errors of a single frequency and a single color and that actual nonuniformities occur in many complex spatial and color forms.

3.4 CHARACTERIZATION OF INPUT SCANNERS THAT GENERATE MULTILEVEL GRAY SIGNALS (INCLUDING DIGITAL CAMERAS)

In this section we will discuss the elementary theory of performance measurements and various algorithms or metrics to characterize them, the scanner factors that govern each, some practical considerations in the measurements, and visual effects where possible. Generally speaking, this is the subject of analyzing and evaluating systems that acquire

sampled images. Originally this was explored as analog sampled images in television, most notably by Schade[83,84] in military and in early display technology.[85] As computer and digital electronics technology grew, this evolved into the general subject of evaluating digital sampled image acquisition systems, which include various cameras and input scanners. Modern scanners and cameras are different only in that a scanner moves the imaging element to create sampling in one direction while the camera imaging element is static, electronically sampling a two-dimensional array sensor in both dimensions. This topic can be divided into two areas. The first concerns scanners and cameras that generate output signals with a large number of levels (e.g., 256), where general imaging science using linear analysis applies.[13] The second deals with those systems that generate binary output, where the signal is either on or off (i.e., is extremely nonlinear) and more specialized methods apply.[86] These are discussed in Section 3.5.

In recent years, the advent of digital cameras and the plethora of office, home, and professional scanners have promoted wide interest in the subject of characterizing devices and systems that produce digital images. Also, several commercially available image analysis packages have been developed for general image analysis, many using scanners or digital cameras, often attached to microscopes or other optical image magnification systems. Components of these packages and the associated technical literature specifically address scanner analysis or calibration.[87–89] A variety of standards activities have evolved in this area.[90–93] Additional related information is suggested by the literature on evaluating microdensitometers.[94,95] These systems are a special form of scanners in which the sensor has a single aperture of variable shape. Much of this work relates to transmitted light scanners but reflection systems have also been studied.[96] Methods for evaluating digital cameras and commercially available scanners for specific applications have been described by many authors.[77,93,97]

3.4.1 Tone Reproduction and Large Area Systems Response

Unlike many other imaging systems, where logarithmic response (e.g., optical density) is commonly used, the tonal rendition characteristics of input scanners are most often described by the relationship between the output signal (gray) level and the input reflectance or brightness. This is because most electronic imaging systems respond linearly to intensity and therefore to reflectance. Three such relationships are shown in Figure 3.23. In general these curves can be described by two parameters, the offset, O, against the output gray level axis and the gain of the system Γ, which is defined in the equation in Figure 3.23. Here g is the output gray level, and R is the relative reflectance factor. If there is any offset, then the system is not truly linear despite the fact that the relationship between reflectance and gray level may follow a straight-line relationship. This line must go through the origin to make the system linear.

Often the maximum reflectance of a document will be far less than the 1.0 (100%) shown here. Furthermore, the lowest signal may be significantly higher than 1% or 2% and may frequently reach as much as 10% reflectance. In order to have the maximum number of gray levels available for each image, some scanners offer an option of performing a histogram analysis of the reflectances of the input document on a pixel-by-pixel or less frequently sampled basis. The distribution is then examined to find its upper and lower limits. Some appropriate safety factor is provided, and new offset and gain factors are computed. These are applied to stretch out the response to cover as many of the total (256 here) output levels as possible with the information contained between the maximum and minimum reflectances of the document.
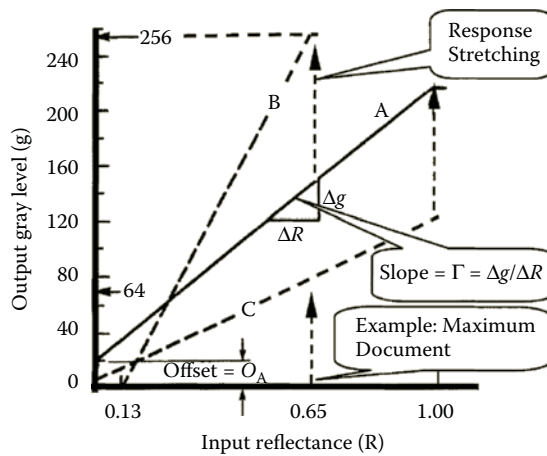
FIGURE 3.23
Typical types of scanner input responses, illustrating the definitions of "gain" (i.e., slope), "offset," and "response stretching."

Other scanners may have a full gray-scale capability from 4 to 12 bits (16–4096 levels). In Figure 3.23, curve C is linear, that is, no offset and a straight-line response up to a reflectance of 1.0 (100%), in this case yielding 128 gray levels. Curve A would represent a more typical general purpose gray response for a scanner while curve B represents a curve adjusted to handle a specific input document whose minimum reflectance was 0.13 and whose maximum reflectance was 0.65. Observe that neither of these curves is linear. This becomes very important for the subsequent forms of analysis in which the nonlinear response must be linearized before the other measurement methods can be applied properly. This is accomplished by converting the output units back to input units via the response function.

In a digital scanner the sensors themselves are fairly linear as can be seen in Figure 3.24 which plots exposure in linear units (lux-s) versus output in millivolts (mV). The response is strictly linear from 0 to 2.2 lux-s and then begins to roll over as it saturates. Notice the difference between the "linear saturation exposure" and the "saturation exposure" which is a graphical construct projecting the linear part of the curve to the maximum signal. It is often observed that digital sensors are linear but it can be seen from Figure 3.24 that this is only true for most but not all of the response curve. The scanner or camera designer is free to use as much or as little of the nonlinear high end of the curve as he desires. For digital cameras the indicated standard exposure differs by camera specifications but is usually in the linear region

It is also possible to arrange the electronics in the video processing circuit so that equal steps in exposure do not generate equal steps in electronic or digital response, but rather are appropriately spaced steps in some units that are more significant, either visually or in terms of materials properties. A logarithmic A/D converter is sometimes used to create a signal proportional to the logarithm of the reflectance or to the logarithm of the reciprocal reflectance (which is the same as "density"). Some scanners for graphic arts applications function in this manner. Another common conversion is making the signal proportional to L^*. Both of these require a larger number of levels to start with than what is output. These systems are highly nonlinear, but may work well with a limited number of gray levels, for example with 8 bits (256 levels) rather than the 10 or 12 bits as discussed earlier.

Many input scanners operate with a built-in calibration system that functions on a pixel-by-pixel basis. In such a system, for example, a particular sensor element that has greater
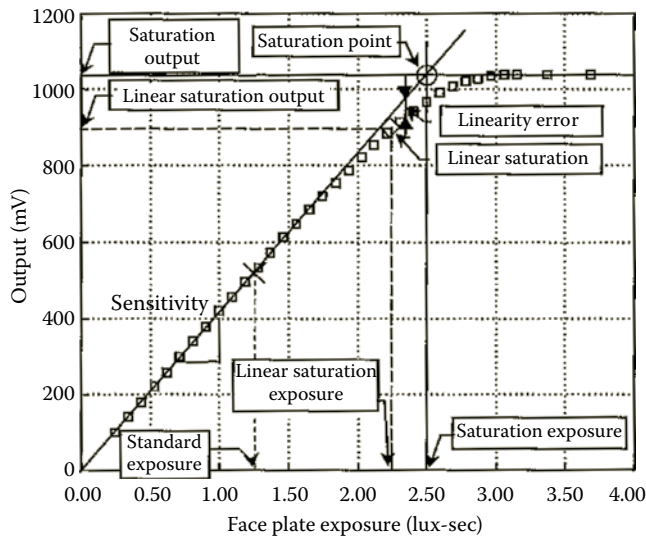
FIGURE 3.24

Fundamental electronic response to light of a sensor used in scanners and cameras showing the linear and nonlinear regions. (Reproduced with permission of the publisher from Nakamura, *J. Image Sensors and Signal Processing for Digital Still Cameras;* Taylor and Francis: Boca Raton, FL, 2006; Mizoguchi, T. Ch 6: Evaluation of image sensors, 179–203.

responsivity than others may be attenuated or amplified by adjusting either the gain or the offset of the system or both. This would ensure that all photosites (individual sensor elements) respond equally to some particular calibrated input, often, as is common with most light measuring devices such as photometers and densitometers, using both a light and dark reflectance reference (e.g., a white and black strip of paint).

It is possible in many systems for the sensor to be significantly lower or higher in responsivity in one place than another. As an example, a maximum responsivity sensor may perform as shown in curve A while a less sensitive photosite may have the response shown in curve C. If curve C was captured with the same A/D converter at the same settings (as is often the case in high-speed integrated circuits), the maximum signal range it contains has only 120 gray levels. A digital multiplier can operate upon this to effectively double each gray level, thereby increasing the magnitude of the scale to 220 or 240, depending upon how it handles the offset. Note that if some of the elements of a one-dimensional sensor responded as curve C, others as A, with the rest in between, then this system would exhibit a kind of one-dimensional granularity or nonuniformity, whose pattern depends upon the frequency of occurrence of each sensor type. This introduces a quantization error varying spatially in one-pixel-wide strips, and ranging, for this example, from strips with only 120 steps to others with 240 steps, yet covering the same distribution of output tones.

An ideal method for measuring tone reproduction is to scan an original whose reflectance varies smoothly and continuously from near 0% to near 100%, or at least to the lightest "white" that one expects the system to encounter. The reflectance is evaluated as a function of position, and the gray value from the scanner is measured at every position where it changes. Then the output of the system can be paired with the input reflectance at every location and a map drawn to relate each gray response value to its associated input reflectance. A curve like Figure 3.23 can then be drawn for each photosite and for various statistical distributions across many photosites.

The classic concepts of quality in tone reproduction generally extend to processes and devices beyond the capture device. Hence the idea of quality for a scanner involves how well it integrates into an overall system that would include a printer or display. This integration is facilitated by image processing, both hardwired in the scanner and through off-line software systems. The graphical construction of a multi quadrant "Jones Plot" has often been used in photography to characterize how a film integrates with camera/optics, film processing, an enlarger and printing paper and even the visual system.[60,61] Similar systems plots can be constructed for the digital system starting either with the camera or the scanner. One such example, using representative system data is shown in Figure 3.25.

Starting at the axis labeled "original density" one creates four quadrants in a clockwise progression starting with Quadrant 1 (lower right) as a plot of digital output level (DOL) versus input Density (or equivalent Log Exposure) for the scanner or camera in question. This is a type of OECF (Optoelectronic Conversion Function) Curve.[65] In this illustration Density of the original target is plotted increasing to the left (Log exposure would increase to the right) and DOL (some call this value digital count or gray value) increases toward
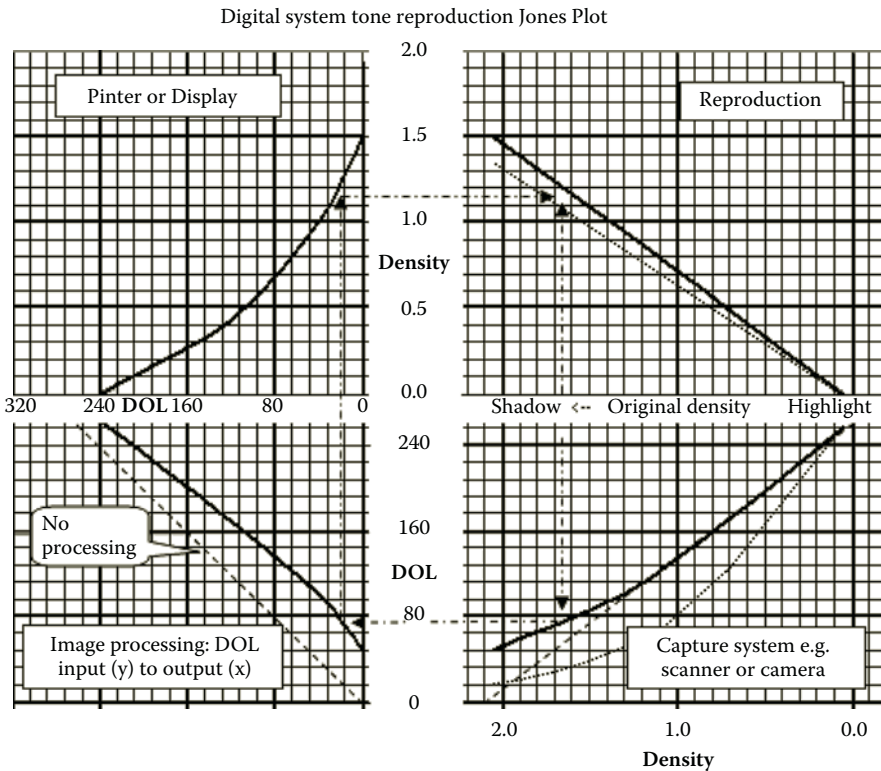
FIGURE 3.25

Jones plot for tonal response in a digital imaging system showing the cascade of components using four quadrants, Q1–Q4: Q1 (lower right) is for digital capture system (scanner or camera) showing density of the original test object (x-axis) mapped to digital output level (i.e., DOLs on y-axis), Q2 (lower left) is the image processing which maps the same DOLs on y-axis to image processing digital output levels (DOLs on x-axis) The latter are also digital input levels in Q3 (upper left) for either a printer or display. Here printer digital input levels map to output printed density (y-axis). In Q4 (upper right) the resulting solid curve (follow the dashed arrow through all four quadrants to see the cascading) gives printed density (y-axis) compared to original test object's density (x-axis). This is the scanning (or photographic) system's overall tone reproduction. See text for dotted/dashed curves.

the top. A dashed line indicates a linear response that follows the actual curve down to the dark region where it begins to "tail up", due largely to flare light. The fact that the log values of density in the bolder solid curve agree so well with the linear values of electronic output (DOLs) suggests that the on-board image processing in this scanner is creating a nonlinear response (for the linear sensors as noted above) in order to better fit some output needs of printing or viewing. This would be typical of some digital cameras as well as some scanners where off-line image processing was expected. The lighter dotted line represents the output of a typical scanner integrated with the printer shown in Quadrant 3, a so-called all-in-one system or a digital copier.

In many such evaluations two of the other three quadrants are specified and the goal is to derive the missing curve. Consider that the rendering device (Quadrant 3 clockwise) is a printer with a fixed density response to a given array of input DOLs. Assume that the it is desired that reproduction (Quadrant 4) be a linear relationship between density of the original and that of the print, even though the maximum densities do not match. This leaves the image processing (Quadrant 2) to be determined. A linear, one for one, image processing between input from the first scanner and output DOLs (dashed curve) would result in a very light print with a somewhat curved density reproduction relationship. The solid curve in Quadrant 2 (Image Processing) results in the desired linear density relationship in Q4.

The second scanner curve (dotted) is less linear but includes on-board image processing which predistorts the output to compensate for the highly curved printer density response curve. This scanner response directly provides another linear final tone reproduction in Quadrant 4, although with slightly lower maximum density. In the Jones Plot this result uses the dashed "no image processing" curve in Quadrant 2 since off-line image processing is not possible in an all-in-one (copier) system. This scanner curve is the same one used in Figure 3.26.

Most scanners operate with sufficiently small detector sites or sensor areas that they respond to input granularity. Thus, a single pixel or single photosite measurement will not
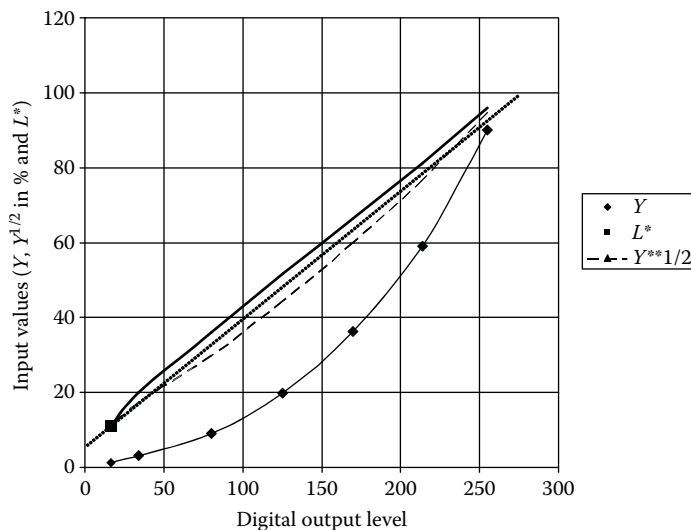
FIGURE 3.26
Scanner output digital levels (x-axis) as predicted by the input test target reflectance values or CIE Y (diamonds), L^* (large squares), $Y^{1/2}$ (small triangles), and a straight dotted line visually fit to the last two. The ordinate is the input value plotted on a relative scale of 0–100. Therefore Y (which ≈ reflectance) and $Y^{1/2}$ are given in %.

suffice to get a solid area response to a so-called uniform input. Some degree of averaging across pixels is required, depending upon the granularity and noise levels of the input test document and the electronic system.

The use of a conventional step tablet or a collection of gray patches, where there are several discrete density levels, provides an approximation to this analysis but does not allow the study of every one of the discrete output gray levels. For a typical step tablet with approximately 20 steps 0.10 reflection density, half of the gray values are measured by only 4 steps, 0, 0.1, 0.2 and 0.3 density (or 50% reflectance). Thus a smoothly varying density wedge is more appropriate for the technical evaluation of an electronic input scanner.

However, suitable wedges are difficult to fabricate repeatably and the use of uniform patches of several discrete densities is common in many operations. See, for example, Reference 97 and the ISO standard targets in Figures 3.48b, 3.49 or the IEEE target in Figure 3.50b (see middle of pattern). Nonetheless, wedges are available (see for example the top of Figure 3.50b), and are essential to accurately evaluate binary scanning (see Section 3.5.2).

Returning to the large area tonal response of the scanner itself, it is tempting to describe it as the linear equation for the sensor itself but the fact is that most scanners today have some built-in image processing associated with them and it is more practical to use a curve. To compensate for some common printer and display response, scanner's tone response neglecting flare can often be mapped as

$$DOL = Hr^{1/\gamma} \tag{3.11}$$

where Hr is the relative exposure from the input and γ (gamma) is a constant designed to compensate for the exponential-shaped curves often found in output printers or displays. Values of 1.8, 2.2 are examples for Mac and PC monitors and 3 to emulate L^* but a general purpose scanner may desire to satisfy all these conditions with some hybrid and a few other terms. Results for a recent desktop scanner are shown in Figure 3.26—an x versus y inverted type of OECF[21] curve—using the resulting digital output levels as the x-axis and various characterizations of the input as the y-axis to deduce the vendors image processing. The system is not linear in reflectance but is approximately linear in either L^* or $\gamma = 2$. Note $\gamma = 2$ is halfway between the Mac and PC standards.

Setting the maximum point equal to 100% input reflectance is often a waste of gray levels since there are no documents whose real reflectance is 100%. A value somewhere between 70% and 90% would be more representative of the upper end of the range of real documents. Some systems adjust automatically to the input target and are therefore difficult to evaluate. They are highly nonlinear in a way that is difficult to compensate. See Gonzalez and Wintz[98] for an early discussion of automatic threshold or gray scale adjustment and Hubel[93] for more recent comments on this subject as it relates to color image quality in digital cameras. Most amateur and some professional digital cameras fall into this automatic domain[93] as do many scanners. A system that finds this point automatically is optimized for each input differently and is therefore difficult to evaluate in a general sense.

An offset in the positive direction can be caused either by an electronic shift or by stray optical energy in the system (as shown in Q1 of Figure 3.25). If the electronic offset has been set equal to zero with all light blocked from the sensor, then any offset measured from an image can be attributed to optical energy. Typical values for flare light, the stray light coming through the lens, would range from just under 1% to 5% or more of full scale.[96] While offset from uniform stray light can be adjusted out electronically, signals from flare light are document dependent, showing up as errors in a dark region only when it is surrounded by a large field of white on the document. Therefore, correction for this

measured effect in the particular case of an analytical measurement with a gray wedge or a step tablet surrounded by a white field may produce a negative offset for black regions of the document that are surrounded by grays or dark colors. If, however, the source of stray light is from the illumination system, the optical cavity, or some other means that does not involve the document, then electronic correction is more appropriate. Methods for measuring the document-dependent contribution of flare have been suggested in the literature.[96,97,99] Some involve procedures that vary the surround field from black to white while measuring targets of different widths;[96] others use white surround with different density patches.[97]

A major point of confusion can occur in the testing of input scanners and many other optical systems that operate with a relatively confined space for the illumination system, document platen, and recording lens. This can be thought of as a type of integrating cavity effect. In this situation, the document itself becomes an integral part of the illumination system, redirecting light back into the lamp, reflectors, and other pieces of that system. The document's contribution to the energy in the illumination depends on its relative reflectance and on optical geometry effects relating to lamp placement, document scattering properties, and lens size and location. In effect the document acts like a position-dependent and nonlinear amplifier affecting the overall response of the system. One is likely to get different results if the size of the step tablet or gray wedge used to measure it changes or if the surround of the step tablet or gray wedge changes between two different measurements. It is best, therefore, to make a variety of measurements to find the range of responses for a given system. These effects can be anywhere from a few percent to perhaps as much as 20%, and the extent of the interacting distances on the document can be anywhere from a few millimeters to a few centimeters (fraction of an inch to somewhat over one inch). Relatively little has been published on this effect because it is so design specific, but it is a recognized practical matter for measurement and performance of input scanners. An electronic correction method exists.[100,101]

3.4.2 MTF and Related Blur Metrics

We will now return to the subject of blur. Generally speaking, the factors that affect blur for any type of scanner include (Table 3.3): the blur from optical design of the system, motion

TABLE 3.3

Factors Affecting Input Scanner Blur and Pointers to Useful MTF Curves That Describe Selected Cases

---

Solid-state scanners

- Lens aberrations as functions of wavelength (see Figure 3.53 if diffraction limited, for example, some microscope optics), field position, orientation, focus distance (see n =3 or 4 of Figure 3.54 for useful equation to fit system with various lens performance)
- Sensor: Aperture dimensions (see Figure 3.51), charge transfer efficiency (CCD), charge diffusion, leaks in aperture mask
- Motion of sensor during reading (Figure 3.51)
- Electronics rise time (measured frequency response)

Flying spot laser beam scanner

- Spot shape and size at document (Gaussian case see Figure 3.52)
- Lens aberrations (as above)
- Polygon aperture or equivalent
- Motion during reading (Figure 3.51)
- Sensor or detector circuit rise time (measured frequency response)

---

of the scanning element during one reading, electronic effects associated with the rise time of the circuit, the effective scanning aperture (sensor photo site) size, and various electro-optical effects in the detection or reading out of the signal. The circuits that handle both the analog and the digital signals, including the A/D converter, may have some restrictive rise times and other frequency response effects that produce a one-dimensional blur.

To explore the analysis of these effects, refer back to Figure 3.5. A primary concept begins with a practical definition of an ideally narrow line object and the image of it. Imagine that the narrow line object profile shown at the top right of Figure 3.5a was steadily reduced in width until the only further change seen in the resulting image Figure 3.5b is that the height of the image peak changes but not the width of its spreading. This is a practical definition of an ideally narrow line source. Under these conditions we would say that the peak of the image on the right of Figure 3.5b was a profile of the *line spread function* for the imaging system. [It is also seen at higher sampling resolution in Figure 3.7b.]

To be completely rigorous about this definition of the line spread function, we would actually use a narrow white line rather than a black line. If the input represented a very fine point in two-dimensional space we would refer to its full two-dimensional image as *a point spread function.* This spreading is a direct representation of the blur in any point in the image and can be convolved with the matrix of all the pixels in the sampled image to create a representation of the blurred image. The line spread function is a one-dimensional form of the spreading and is usually more practical from a measurement perspective. In the case illustrated, the line spread function after quantization would be shown in Figure 3.5c as the corresponding distribution of gray pixels.

There are several observations to be made about this illustration, which underscore some of the practical problems encountered in typical measurements. First, the quantized image in Figure 3.5c is highly asymmetric while the profile of the line shown in Figures 3.5a and b appears to be more symmetric. This results from sampling phase and requires that a measurement of the line spread function must be made, adjusting sampling phase in some manner (Figure 3.7b). This is especially important in the practical situation of evaluating a fixed sampling frequency scanner. Secondly, note the limited amount of information in any one phase. It can be seen that the smooth curve representing the narrow object in Figure 3.5b is only represented by three points in the sampled and quantized image.

The averaging of several phases would improve on this measurement, increasing both the intensity resolution and the spatial resolution of the measurement. One of the easiest ways to do this is to use a long narrow line and tip it slightly relative to the sampling grid so that different portions along its length represent different sampling phases. One can then collect a number of uniformly spaced sampling phases, each being on a different scan line, while being sure to cover an integer number of complete cycles of sampling phase. One cycle is equivalent to a shift of one complete pixel. The results are then combined in an interleaved fashion, and a better estimate of the line spread function is obtained. (This is tantamount to increasing the sampling resolution, taking advantage of the one-dimensional nature of the test pattern.) This is done by plotting the recorded intensity for each pixel located at its properly shifted absolute position relative to the location of the line. To visualize this consider the two-phase sampling shown in Figure 3.7b and c. There the resulting pixels from phase A could be interleaved with those from phase B to create a composite of twice the spatial resolution. Additional phases would further increase effective resolution.

In the absence of nonlinearities and nonuniformities, the individual line spread functions associated with each of the effects in Table 3.3 can be mathematically convolved with each other to come up with an overall system line spread function.

3.4.2.1 MTF Approaches

For engineering analysis, use of convolutions and measurements of spread functions are often found to be difficult and cumbersome. The use of an *optical transfer function* (OTF) is considered to have many practical advantages from both the testing and theoretical points of view. The OTF is the Fourier transform of the line spread function. This function consists of a modulus to describe normalized signal contrast attenuation (or amplification), and a phase to describe shift effects in location, both given as a function of spatial frequency. The signal is characterized as the modulation of the sinusoidal component at the indicated frequency. Therefore the contrast altering function is described as a *modulation transfer function* (MTF). The value of OTF analysis is that all of the components in a linear system can be described by their OTFs, and these are multiplied together to obtain the overall system response. The method and theory of this type of analysis has been covered in many journal articles and reference books.[13,47,78,102]

Certain basic effects can be described in analytic form as MTFs and a few of these are indicated in Table 3.3 and illustrated in Section 3.9 in Figures 3.51–3.53, plotted in logarithmic form to facilitate graphical manipulation. Several photographic MTF curves are plotted in Figure 3.55 to provide a reference both as a range of input signals for film scanners or a range of output filters that transform optical signals to permanently readable form. One may also consider using these with an enlargement factor for understanding input of photographic prints to a desktop or graphic arts scanner. (For example, a spatial frequency of 10 cycles/mm on an 8× enlarged print is derived from the 80 cycles/mm pattern on the film. Therefore the film MTF at 80 cycles/mm is an upper limit input signal for an 8×10 in enlargement of a 35-mm film.) Other output MTFs would involve display devices such as monitors, projection systems, analog response ink systems, and xerographic systems. Obtaining the transform of the line spread function has many of the practical problems associated with measuring the line spread function itself plus the uncertainty of obtaining an accurate digital Fourier transform using a highly quantized input.

There are several commonly used methods for measuring the OTF. These include:

1. Measuring images of narrow lines using appropriate compensation for finite widths.
2. Directly measuring images of sinusoidal distributions of radiation.[103,104]
3. Harmonic analysis of square-wave patterns.[97,103,104]
4. Taking the derivative of the edge profile in the image of a very sharp input edge. This generates the line spread function, and then the Fourier transform is taken, taking care to normalize the results properly.[13,92] (Table 3.9, ISO TC42, WG18 and Figures 3.48a and b)
5. Spectral analysis of random input (e.g., noise) targets with nearly flat spatial frequency spectrum.

It should also be mentioned at this point that for most characterizations of imaging systems the modulus, that is, the MTF, is more significant than the phase. The phase transfer function, however, may be important in some cases and can be tracked either by careful analysis of the relative location of target and image in a frequency-by-frequency method or by direct computation from the line spread function.

In general, these methods involve the use of input targets that are not perfect. They must have spatial frequency content that is very high. The frequency composition of the input target is characterized in terms of the modulus of the Fourier transform, $M_{in}(f)$, of

its spatial radiance profile. The frequency decomposition of the output image is similarly characterized, yielding $M_{out}(f)$. Dividing the output modulation by the input modulation yields the MTF as

$$\text{MTF}(f) \;=\; \frac{M_{out}(f)}{M_{in}(f)} \tag{3.12}$$

The success of this depends upon the ability to characterize both the input and the output accurately.

A straightforward method to perform this input and output analysis involves imaging a target of periodic intensity variations and measuring the modulation on a frequency-by-frequency basis. If the target is a set of pure sine waves of reflectance or transmittance, that is, each has no measurable harmonic content, and the input scanner is linear, then the frequency-by-frequency analysis is straightforward. Modulation of a sinusoidal distribution is defined as the difference between the maximum and minimum divided by their sum. The modulation is obtained directly, measuring the maximum and minimum output gray values g', and the corresponding input reflectance (or transmittance or intensity) values, R, of Equation 3.17 for each frequency pattern. Expanding the numerator and denominator for Equation 3.16 and the case of sinusoidal patterns and linear systems yields

$$MTF(f) = \frac{-[g'_{max}(f) - g'_{min}(f)]/[g'_{max}(f) + g'_{min}(f)]}{[R_{max}(f) - R_{min}(f)]/[R_{max}(f) + R_{min}(f)]} \tag{3.13}$$

where the prime is used to denote gray response that has been corrected for any non-linearity as described below.

Figures 3.27 and 3.28 show an example of this process. In Figure 3.27 we see the layout of a representative periodic square-wave test target (aR) and a sinusoidal test target (aL) which exhibits features of well-known patterns[105] available today in a variety of forms (e.g., from Applied Image Reference 106). The periodic distributions of intensity (reflectance) are located in different blocks in the center of the pattern. Uniform reflectance patterns of various levels are placed in the top and bottom rows of the sinusoid to enable characterizing the tone response. A similar arrangement of uniform blocks is used with the square waves but not shown here. This enables correcting for its nonlinearities should there be any. Parts (b) and (c) show enlargements of parts of the square-wave pattern selecting a lower and a higher frequency. Parts (f) and (g) are enlargements of a gray image display of the electronically captured image of the same parts of the test target. Parts (d), (e), (h), and (i) show profiles of the patterns immediately above them.

To calculate a MTF, the modulation of each pattern is measured. For sinusoidal input patterns, one can use Equation 3.17 directly, finding the average maximum and minimum for many scan lines for each separate frequency. These modulation ratios, plotted on a frequency-by-frequency basis, describe the MTF. For square-wave input, the input and output signals must be Fourier transformed into their spatial frequency representations and only the amplitudes of the fundamental frequencies used in Equation 3.17. Schade[103] offers a method to compute the MTF by measuring the modulations of images of each square wave directly (i.e., the square wave response) and then unfolding for assumed perfect input square-waves without taking the transforms.

From a practical standpoint it is important to tip the periodic patterns slightly as seen in parts (d) and (e) to cover the phase distributions as described above under the spread
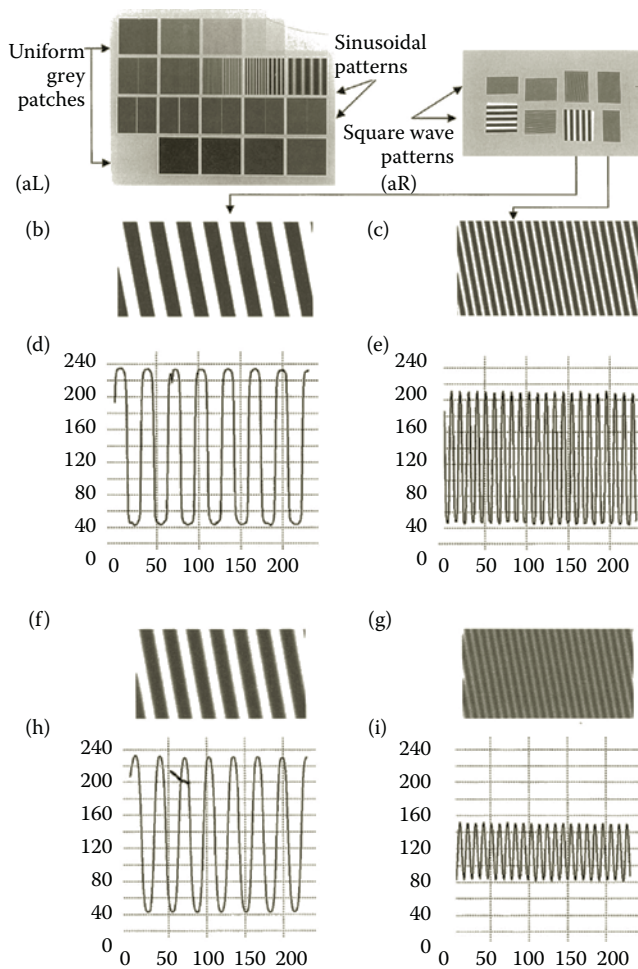
FIGURE 3.27

Example of images and profiles used in MTF analysis. Part (aL) shows a full pattern of gray patches and sinusoidal reflectance distributions at various frequencies. Part (aR) shows the frequency components of a square-wave test chart. Note that the bars are slanted slightly to facilitate measuring at different sampling phases. The figures on the left, (b), (d), (f), and (h), come from a low-frequency square-wave pattern as indicated by the arrow. The figures on the right, (c), (e), (g), and (i) are from a higher frequency square wave. Enlargements of the test patterns in (aR) are shown in parts (b) and (c). Slightly blurred images after scanning (as might be seen on a display of the scanner output) are shown in parts (f) and (g). Profiles of each of these images are displayed beneath them in parts (d), (e), (h), and (i), respectively. Because these are square-wave test patterns, special analysis of these patterns is required to compensate for effects of harmonics as described in the text. The reader should ignore small moiré effects caused by the reproduction process used to print this illustration.

function discussion. A new higher resolution image can be calculated by interleaving data points from the individual scan lines, each of which is phase shifted with respect to the sine or square wave.

Figure 3.28 shows several examples of linear and nonlinear response curves. It describes correcting the output of an MTF analysis (i.e., using g' in Equation 3.17) for the case of nonlinearity with offset. Here the maximum and minimum values for the sine waves are unfolded through the response curve to arrive at minimum and maximum input
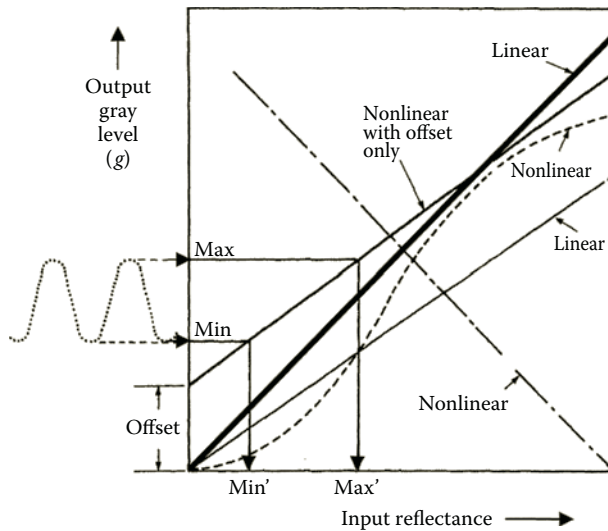
FIGURE 3.28
Examples of linear and nonlinear large-area response curves with an illustration of output modulation correction for offset using effective gray response at max' and min'.

reflectances, that is, the linear variables. For a full profile analysis, as needed for a Fourier transform method, and to obtain corrected modulation, each output gray level must be modified by such an operation.

If the response curve for the system was one of those indicated as linear in Figure 3.28, then no correction is required. It is important to remember that while the scanner system response may obey a straight-line relationship between output gray level and the reflectance, transmittance, or intensity of the input pattern, it may be offset due to either optical or electronic biases (e.g., flare light, electronic offset, etc.). This also represents a nonlinearity and must be compensated.

As the frequency of interest begins to approach the sampling frequency in an aliased input scanning system, the presence of sampling moiré becomes a problem. This produces interference effects between the sampling frequency and the frequency of the test pattern. If the pattern is a square wave, this may be from the higher harmonics (e.g., 3×, 5× the fundamental). When modulation is computed from sampled image data using maxima and minima in Equation 3.17, errors may arise. There are no harmonics for the sinusoidal type of patterns, a distinct advantage of this approach.

See Figure 3.29 for an example of these phase effects on a representative MTF curve. It shows errors for test sine waves whose period is a submultiple of the sampling interval. Consider the case where the sinusoidal test pattern frequency is exactly one-half the sampling frequency, that is, the Nyquist frequency. In this case, when the sampling grid lines up exactly with the successive peaks and valleys of the sine wave, we get a strong signal indicating the maximum modulation of the sine wave (point A). When the sampling grid lines up at the midpoint between each peak and valley of the sinusoidal image (phase shifted 90° relative to the first position), each data point will be the same, and no modulation whatever results (point B). There is no right or wrong answer to the question of which phase represents the true sine-wave response, but the analog or highest value is often considered as the true MTF. Each phase may be considered as having its own sine-wave
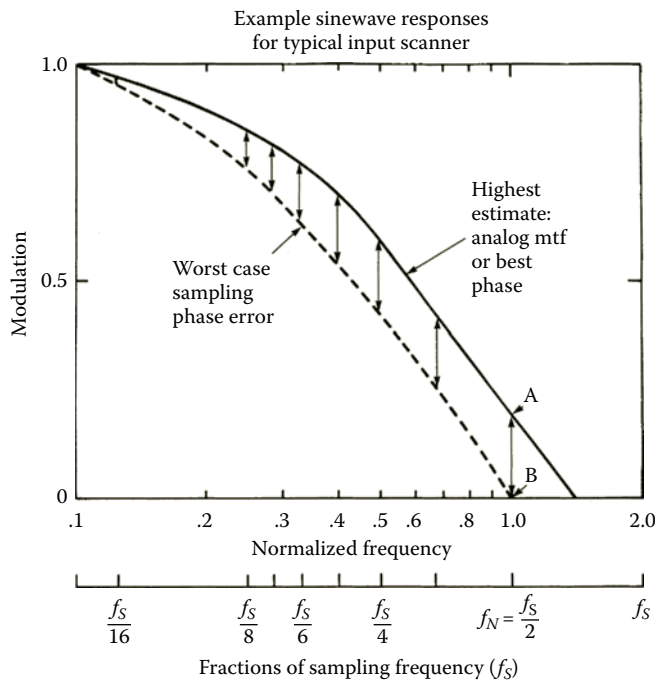
FIGURE 3.29
Example of the possible range of measured sine wave response values of an input scanner, showing the uncertainty resulting from possible phase variations in sampling.

response. Reporting the maximum and minimum frequency response and reporting some statistical average are both legitimate approaches, depending upon the intended use of the measurement. It is common practice to represent the average or maximum and the error range for the reported value.

The analog MTF, on the other hand, is only given as the maximum curve, representing the optical function before sampling. Therefore, the description of upper and lower phase boundaries for sine wave response shows the range of errors in the measurement of the MTF which one might get for a single measurement. This strongly suggests the need to use several phases to reduce error if the analog MTF is to be measured. Mathematically, phase errors may be thought of as a form of microscopic nonstationarity complicating the meaning of MTF for sampled images at a single phase. The use of information from several phases reduces this complication by enabling one to approximate the correct analog MTF that obeys the principle of stationarity.

In the case of a highly quantized system, meaning one having a relatively small number of gray levels, quantization effects become an important consideration in the design and testing of the input scanner. The graph in Figure 3.30 shows the limitation that quantization step size, E_q, imposes on the measurement of the MTF using sine waves. The number of gray levels used in an MTF calculation can be maximized by increasing the contrast of the sinusoidal signal that is on the input test pattern. It can also be increased by repeated measurements in which some analog shifts in signal level are introduced to cause the quantization levels to appear in steps between the previous discrete digital levels and therefore at different points on the sinusoidal distributions. The latter could be accomplished by changing the light level or electronic gain.
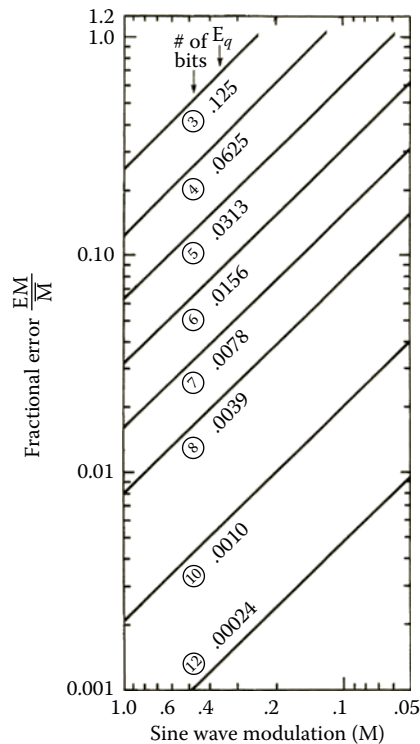
FIGURE 3.30

Errors in MTF measurements, showing the effects of modulation at various quantization errors. EM is in zero to peak units. $\overline{M}$ is average modulation. The numbers in the circles on each line indicate the system quantization in bits. E_q is the size of quantization step, where a full-scale signal = 1.0.

It is also important to note that because the actual MTF can vary over the field of view, a given measurement may only apply to a small local region over which the MTF is constant. (This is sometimes called an isoplanatic patch or stationary region within the image.) To further improve the accuracy of this approach, one can numerically fit sinusoidal distributions to the data points collected from a measurement, using the amplitude of the resultant sine wave to determine the average modulation. Taking the Fourier transform of the data in the video profile may be thought of as performing this fit automatically. The properly normalized amplitude of the Fourier transform at the spatial frequency of interest would, in fact, be the average modulation of the sine wave that fits the video data best.

The approach involving the application of square-wave test patterns (as opposed to sinusoidal ones, which have intrinsic simplicity as an advantage) has been shown by Newell and Triplett[104] to have significant practical advantages. They also show square-wave analysis has excellent accuracy when all-important details are carefully considered, especially the sampling nature of the analysis and the noise and phase effects. Square-wave test patterns are commonly found in resolving power test targets and are much easier to fabricate than sine waves, because the pattern exists as two states, foreground (e.g., black bars) and background (e.g., white bars). Two levels of gray bars may also be used depending on desired contrast. Fourier transform analysis and paying attention to the higher harmonics have been particularly effective.[97,104] It has been shown that the general Discrete Fourier

Transform (DFT) algorithms where the length of the input can be altered is much better suited to MTF analysis than use of the Fast Fourier Transform (FFT) where the required power of two sampling points are a limitation.

One successful practice[104] included tipping the bar patterns to create a one pixel phase shift over eight scan lines and averaging over approximately 30 scan lines to reduce noise. The DFT was used and tuned to the precise frequency of the given bar pattern by changing the number of cycles of the square wave being sampled and using approximately 1000 data points. An improvement in MTF accuracy of several percent was demonstrated using the DCT over the more common FFT.

It is generally advisable to measure the target's actual harmonic content rather than to assume that it will display the theoretical harmonics of a perfect mathematical square wave. Likewise, other patterns of known spectral content can be calibrated and used. Edge analysis techniques are also popular.[91,92,107] Using similar care such as a 5° slanted edge, a standardized algorithm, and specification on the edge quality, these achieve good accuracy too.

3.4.2.2 The Human Visual System's Spatial Frequency Response

As a matter of practical interest, several spatial frequency response measurements of the HVS are shown in Figure 3.31. These provide a reference to compare to system MTFs. The work of several authors is included.[108–114] The curves shown have all been normalized to 100% at their respective peaks to provide a clearer comparison. Except for the various normalizing factors, the ordinates are analogous to a modulation transfer factor of the type described by Equation 3.17. However, MTFs are applicable only to linear systems, which the human eye is not. The visual system is in fact thought to be composed of many independent, frequency-selective channels,[115,116] which, under certain circumstances, combine to give an overall response as shown in these curves. It will be noted that the response of the visual system has a peak (i.e., modulation amplification relative to lower frequencies) in the neighborhood of 6 cycles/degree (0.34 cycles/milliradian) or 1 cycle/mm (25 cycles/in) at a standard viewing distance of 340 mm (13.4 inches). The variations among these curves reflect the experimental difficulties inherent in the measurement task and may also illustrate the fact that a nonlinear system such as human vision cannot be characterized by a unique MTF.[117] For this reason such curves are called contrast sensitivity functions (CSFs) and not MTFs. For readers desiring a single curve, the luminance CSF reported by Fairchild[4] is given in Section 3.9, Figure 3.56. While similar in shape to many curves in Figure 3.31, it displays a greater range of responses and also shows the red–green and blue–yellow chromatic CSFs.

3.4.2.3 Electronic Enhancement of MTFs: Sharpness Improvement

These visual frequency response curves suggest that the performance of an imaging system could be improved if its frequency response could be increased at certain frequencies. It is not possible with most passive imaging systems to create amplification at selected frequencies. The use of electronic enhancement, however, can impart such an amplified response to the output of an electronic scanner. Amplification here is meant to imply a high-frequency response that is greater than the very low-frequency response or greater than unity (which is the most common response at the lowest frequencies). This can be done by convolving the digital image with a finite-impulse response (FIR) electronic filter that has negative sidelobes on opposite sides of a strong central peak. The details of FIR
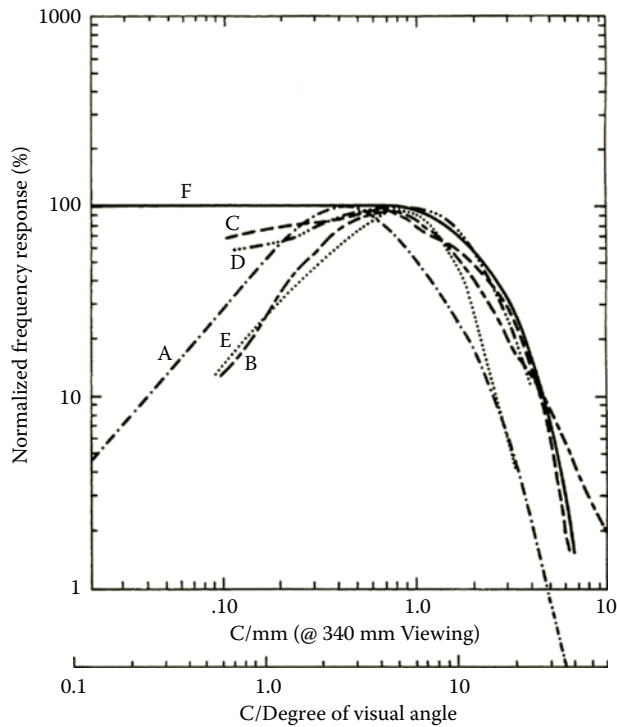
FIGURE 3.31
Measured spatial frequency response of the human visual system, showing the effects of experimental conditions on the range of possible results. Findings are presented according to (A) Campbell,[108] (B) Patterson[114] (Glenn et al.[112]), (C) Watanabe et al.,[111] (D) Hufnagel (after Bryngdahl),[117] (E) Gorog et al.,[109] (F) Dooley and Shaw.[113] All measurements are normalized for 100% at peak and for 340-mm viewing distance. Note the universal visual angle scale at the bottom. See Figure 3.56 (from Fairchild[4] for a seventh and more recent curve showing a larger response range and the two chromatic channels.

filter design are beyond the scope of this chapter, but the effects of two typical FIR filters on the system MTF are shown in Figure 3.32.

3.4.3 Noise Metrics

Noise in an input scanner, whether the scanner is binary or multilevel gray, comes in many forms (see Section 3.2.2). A brief outline of these can be found in Table 3.4. Various specialized methods are required in order to discriminate and optimize the measurement of each.

In this table we see that there are both fixed and time-varying types of noise. They may occur in either the fast or the slow scan direction and may either be additive noise sources or multiplicative noise sources. They may be either totally random or they may be structured. In terms of the spatial frequency content, the noise may be flat (white), that is, constant at all frequencies to a limit, or it may contain dominant frequencies, in which case the noise is said to be colored. These noise sources may be either random or deterministic; in the latter case, there may be some structure imparted to the noise.

The sources of the noise can be in many different components of the overall system, depending upon the design of the scanner. Instances of these may include the sensor of
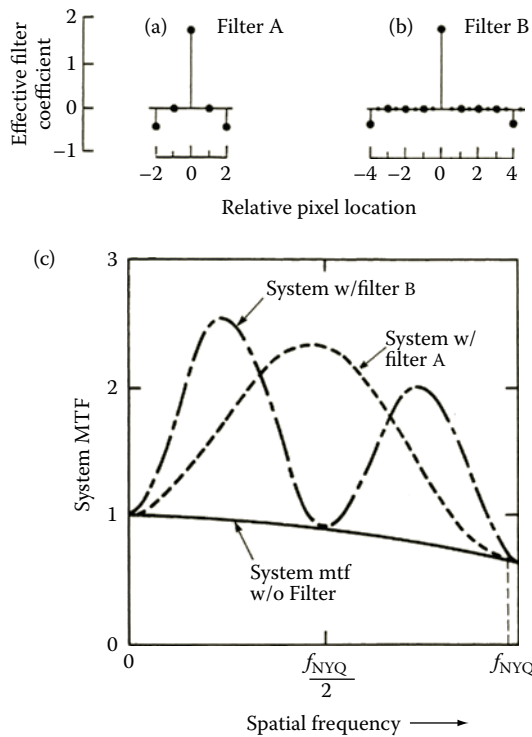
FIGURE 3.32
Two examples of enhancement of a scanning system MTF, using electronic finite-impulse response (FIR) filters in conjunction with an input scanner: (a) the one-dimensional line spread function for "filter A"; (b) line spread function for "filter B", which is the same shown as in (a) but four pixels wider; (c) shows the effect of these filters on the MTF of an aliased high-quality scanner.

TABLE 3.4

Types and Sources of Noise in Input Scanners

Category	Type
Distribution	Fixed with platen, time varying
Type of operation	Multiplicative, additive
Spatial frequency	Flat (white), colored
Statistical distribution	Random, structured, image-dependent
Orientation	Fast scan, slow scan, none (two-dimensional)
Sources	Sensor, electronic system, motion error Calibration error, photons, lamp, laser Controls, optics

the radiation, or the electronics, which amplify and alter the electrical signal, including, for example, the A/D converter. Other noise sources may be motion errors, photon noise in low-light-level scanners, or noise from the illuminating lamp or laser. Sometimes the optical system, as in the case of a laser beam input scanner, may have instabilities that add noise. In many scanners there is a compensation mechanism to attempt to correct for fixed noise. This typically utilizes a uniformly reflecting or transmitting strip of one or more different densities parallel to the fast scan direction and located close to the input position

for the document. It is scanned, its reflectance(s) is memorized by the system, and it is then used to correct or calibrate either the amplifier gain or its offset or both.

Such a calibration system is, of course, subject to many forms of instabilities, quantization errors, and other kinds of noise. Since most scanners deal with a digital signal in one form or another, quantization noise must also be considered.

In order to characterize noise in a gray output scanner, one needs to record the signal from a uniform input target. The most challenging task is finding a uniform target with noise so low that the output signal does not contain a large component due to the input or document noise. In many of these scanners, the system is acting much like a microdensitometer, which reacts to such input noise as the paper fibers or granularity in photographic, lithographic, or other apparently uniform samples.

The basic measurement of noise involves understanding the distribution of the signal variation. This involves collecting several thousand pixels of data and examining the histogram of their variation or the spatial frequency content of that variation. Under the simplifying assumptions that we are dealing with noise sources that are linear, random, additive, and flat (white), a typical noise measurement procedure would be to evaluate the following expression:

$$\sigma_s^2 = \sigma_t^2 - \sigma_o^2 - \sigma_m^2 - \sigma_q^2 \tag{3.14}$$

where σ_s = the standard deviation of the noise for the scanner system (s); σ_t = the total (t) standard deviation recorded during the analysis; σo = the standard deviation of the noise in the input object (o) measured with an aperture that is identical to the pixel size; σ_m = the standard deviation of the noise due to measurement (m) error; and σ_q = the standard deviation of the noise associated with the quantization (q) error for those systems that digitize the signal. This equation assumes that all of the noise sources are independent. Removing quantization noise is an issue of whether one wants to characterize the scanner with or without the quantization effects, since they may in fact be an important characteristic of a given scanner design. The fundamental quantization error[118] is

$$\sigma_q^2 = \frac{2^{-2b}}{12} \tag{3.15}$$

where b = the number of bits to which the signal has been quantized.

The second and third terms in Equation 3.18 give the performance of the analog portion of the measurement. They would include the properties of the sensor amplification circuit and the A/D converter as well as any other component of the system that leads to the noise noted in the table above. The term σ_q^2 characterizes the digital nature of the scanner and, of course, would be omitted for an analog scanning system.

Equation 3.18 is useful when the noise in the system is relatively flat with respect to spatial frequency or when the shape of the spatial frequency properties of all of the subsystems is similar. If, however, one or more of the subsystems involved in the scanner is contributing noise that is highly colored, that is, has a strong signature with respect to spatial frequency, then the analysis needs to be extended into frequency space. This approach uses Wiener or power spectral analysis.[47,119] Systems with filters of the type shown in Figure 3.32 would exhibit colored (spatial frequency dependent) noise. A detailed development of Wiener spectra is beyond the scope of this chapter. However, it is important here to realize its basic

form. It is a particular normalization of the spatial frequency distribution of the *square* of the signal fluctuations. The signal is often in optical density (*D*), but may be in volts, current, reflectance, and so on. The normalization involves the area of the detection aperture responsible for recording the fluctuations. Hence units of the Wiener spectrum are often [μm *D*]² and can be [μm *R*]². (See Reference 119; the latter units are more appropriate for scanners because they respond linearly to reflectance, or, more generally, to irradiance.)

3.5 EVALUATING BINARY, THRESHOLDED, SCANNED IMAGING SYSTEMS

3.5.1 Importance of Evaluating Binary Scanning

Many output scanners accept only binary signals, that is, on or off signals for each pixel or subpixel. This translates to only black or white pixels on rendering. A binary thresholded image may be generated directly by the scanner, or reduced to this state through image processing just prior to delivery to an output scanner. It may also be the degenerate state of inappropriate gray or dithered image processing in which signals are overamplified in a variety of ways, to look like thresholded images. Irrespective of how they are generated, binary thresholded renderings remain an important class of images today and often produce image characteristics that are surprising to the uninitiated. Understanding and quantifying this type of imaging become an important part of the evaluation of the overall input scanner to output scanner-printer system.

As noted earlier, there are two types of binary digital images, either thresholded or dithered signals. To a first order, dithered systems (halftoned or error diffused) can be evaluated in a way similar to that used to evaluate full gray systems, with one simplifying assumption. The underlying concept is that, within the effective dither region over which a halftone dot is clustered or the error is diffused, the viewer does not notice the dither pattern. As a result, these systems are primarily evaluated using instruments and methods whose resolution is equal to or larger than the effective dither region and hence are confined largely to tone rendition and some forms of image noise. Extensive discussions of these measurement approaches are beyond the scope of this chapter. Many of the basic underlying principles are discussed in Section 3.5 of Chapter 3 of the earlier edition of this book and in 2.2.3 under halftone system response and detail rendition. Limited discussion follows here.

To understand and evaluate binary images, a few new concepts are explored and appropriate analytic methods developed.[120]

3.5.1.1 Angled Lines and Line Arrays

To adequately describe performance over a range of sampling phases, it is important that the image structures must be measured at a large number of sampling phases. In other words the evaluation is repeated several times with respect to the input pattern at positions predetermined to create images at different sampling phases. These may be produced by shifting the components by various fractions of a pixel. Tilting lines or rectilinear patterns by a few degrees generates a continuum of phases along the edge of the designated structure. Without tilting for example, a fine line may be imaged in one test as two-pixel wide,

and on another random test and therefore at another sampling phase, it may be imaged as one-pixel wide.

3.5.2 General Principles of Threshold Imaging Tone Reproduction and Use of Gray Wedges

For binary system response, that is, binary tone reproduction, testing can best be accomplished by having smooth calibrated structures that allow finding the on–off binary transitions to a small fraction, say 1 part in 200, of an input characteristic like reflectance.

A calibrated gray wedge is useful. This device resembles a photographic step tablet except that it varies smoothly from a very low density to a very high density without steps. Ideally it would vary linearly in reflectance or transmittance as a function of distance, but the physical means for creating wedges often make them somewhat logarithmic. Accurate measurement of transmittance or reflectance versus distance from some reference mark on the wedge is used to calibrate the pattern, as shown in the graph in the top of Figure 3.33. Note that the picture at the bottom shows the image of a negative working system—that is, the maximum transmittance gives black and the minimum gives white.

The distance at which the wedge turns from black to white (or is 50% black and 50% white pixels for a noisy image) is measured for a given *gray threshold* and converted to a *transmittance or reflectance) threshold.*

3.5.2.1 Underlying Characteristic Curve and Noise

If one is trying to determine the underlying characteristic curve of the scanner, a series of specified transmittances or reflectances can be determined along the wedge. The digital
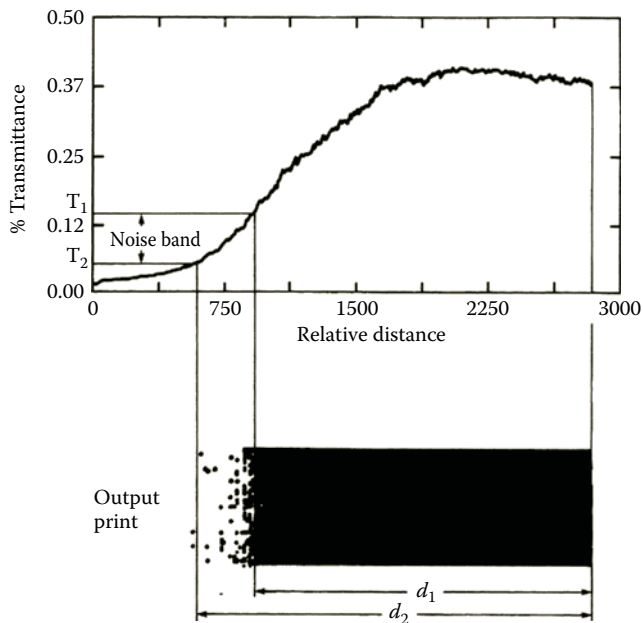
FIGURE 3.33
Transmittance profile of a gray wedge and a corresponding output print (binary image), both as function of distance in arbitrary units. Smallest dots at left are individual pixels.

output gray level of the threshold setting that creates the black to white transition at each specified transmittance or reflectance is then plotted. This describes the underlying characteristic curve of the binary system in output threshold gray level versus input transmittance or reflectance and is a type of OECF (optoelectronic conversion function) curve (Reference 65).

Because of noise in the typical system, including noise on the input document, the location along the length of the wedge where the image changes from white to black will not be a sharp straight line. Rather it will be a region of noise as shown in Figure 3.33. Typically the middle of this transition region $(T_2 + T_1)/2$ is identified as the transmittance or reflectance at which an average threshold is set. A unit called Gray Wedge Noise = GWN = $(T_1 - T_2)$ can be set here where T_2 and T_1 the transmittances at a fixed probability of finding a minimal or maximum response (such as 95% white or black).

3.5.3 Binary Imaging Metrics Relating to MTF and Blur

Given the on–off nature of a binary thresholded image, a linear approach such as MTF analysis does not work. To deal with this nonlinearity we can pose three specific types of questions about imaging performance: (1) *Detectability:* what is the smallest isolated detail that the system can detect? (2) *Discriminability of fine detail:* what is the finest, most complex small structure or fine texture that the system can handle (legibility, resolving power)? (3) *Fidelity of reproduction:* for the larger details and structures, how do the images compare to the original input such as some reasonable width line? To create a specific metric in each of these categories, one defines a specific test object or test pattern that relates to the imaging application. One then defines a set of rules or criteria by which to judge performance against that pattern. These include rules for determining threshold variation/selection and phase probabilities for decision criteria. These are more completely described in the earlier edition (Reference 53) including line width detectability and fidelity.

3.5.3.1 Resolving Power (A Measure for Discrimination of Fine Detail)

Resolving power is a commonly used descriptor of image quality for nearly every kind of imaging system. Its application to binary electronic imaging and scanning systems is therefore appealing. However, because of its extreme sensitivity to threshold and test pattern design, it must be applied with great care to prevent misleading results. Its primary value is in understanding performance for fine structures. The metrics noted above apply to isolated detail, while resolving power tends to emphasize the ability to distinguish many closely spaced details. In general, it can often be considered as an attempt to measure the cutoff frequency, that is, the maximum frequency for the MTF of a system. Binary systems are so nonlinear that even an approximate frequency-by-frequency MTF analysis cannot be considered.

The basic concept of a resolving power measurement is to attempt (through somewhat subjective visual evaluations) to detect a pattern in the thresholded video, which, to some level of confidence, resembles the pattern presented in the test target. For example, one may establish a criterion of 75% confidence that the image represents five black bars and four white spaces at the appropriate spacing. Values of 50%, 95%, 100% or any other confidence could also be used. As in all the metrics above, each judgment must be measured over a wide range of sampling phases for the bar pattern, resulting in an appropriate average confidence over all phases. Tipping the bar target so the length of the bars intercepts an integer number of sampling phases is again convenient. More detail on this test is given

in the earlier edition (Reference 53) including unusual and non-intuitive sampling patterns that may arise, pseudo- or spurious resolution (Reference 121) and gray in white spaces between bars in well-resolved image that is caused by light scattering in test target substrates[71–73] and may change expected thresholding results.

Binary imaging systems have extremely powerful contrast enhancement properties under the right circumstances. Selecting exactly the right threshold, one between the light and dark part of a resolving power image, amplifies a 1% or 2% modulation of the optical or gray electronic image to an on–off pattern (i.e., 100% modulation) that can be easily resolved in the video bit map.

Because it is possible to detect these low contrast patterns, it is also common to detect the situation known as pseudo- or spurious resolution. Here the blurring due to the input scanner is in a particular form that causes the light bars of the pattern to turn dark, and the dark bars to turn light.[121] There are many strategies available for using resolving power. These would include:

1. Varying the threshold and noting the pattern that is resolved (see Figure 3.34).
2. Fixing the bar spacing of interest, and looking for the threshold at which it is detected.

It should be noted in any digital system (binary or gray) that there is a strong dependence on angular orientation. Unlike resolving power in a conventional optical system, a nonzero or non-90° orientation may in fact perform better because of the independent MTFs in the x- and y-directions and the rectangular sampling grid.

Resolving power test targets come in many forms and substrates and these make a significant difference in the results, as noted earlier. Some of these are illustrated in Figure 3.35.
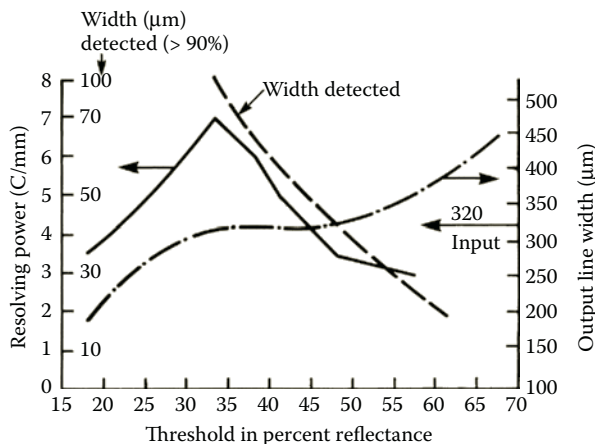
FIGURE 3.34

Plots of line width detectability, fidelity, and resolving power as functions of threshold setting in a binary imaging system. (Adapted from Lehmbeck, D.R. *Imaging Performance Measurement Methods for Scanners that Genereate Binary Output.* 43rd Annual Conference of SPSE, Rochester, NY, 1990; 202–203.). Arrows on each curve indicate which axis represents the ordinate for that curve. "Output line width" in μm is for images of 320 μm "input line width" as noted by arrow at right. The "Width Detected" curve refers to the left inside axis and is given in μm of the input line width, which is detected at the designated threshold for >90% of the sampling phases.
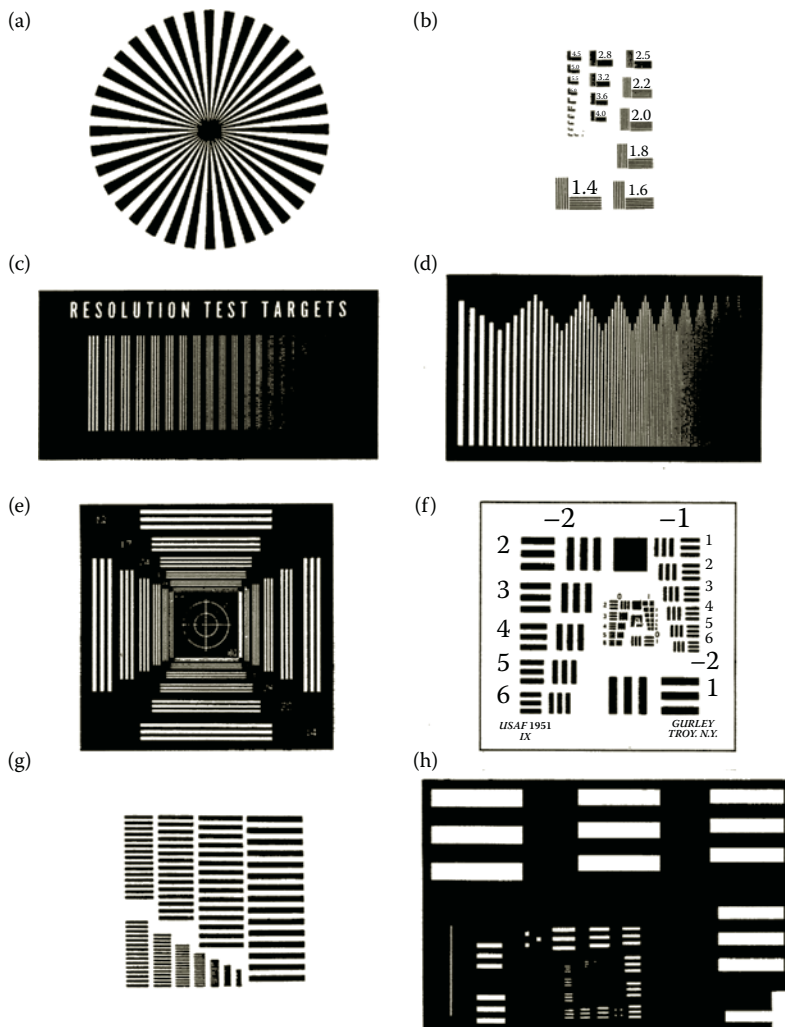
FIGURE 3.35

Images made with various bar pattern test targets in common use for measurement of resolving power and related metrics. See text for identification and description of each type: (a) radial graded frequency chart, (b) NBS Microcopy test chart, (c) a machine readable chart, (d) Sayce chart – a linear graded frequency pattern, (e) NBS lens testing chart, (f) USAF test chart, (g) Ealing test pattern, and (h) portion of ANSI resolving power chart.

Only the coarser patterns are imaged in this illustration and no attempt should be made by the reader to use these images for testing. They are illustrations only. Two general forms exist: those with discrete changes in the bar spacing and those with continuous variation in the bar spacing. In the former category there are several fairly commonly encountered types, designed for visual testing, namely the NBS lens testing type (e), NBS microcopy test chart type (b), the US Air Force type (f), the Cobb chart type (2 bars, not shown) and finally, the ANSI Resolving Power test patterns (h). This form also includes the extended square-wave types as represented by either Ronchi rulings (not shown) or ladder charts (not shown), which are simply larger arrays of the Ealing test pattern (g) which shows 15 bars of each square wave. Machine-readable forms are also useful where the modules are arranged in a pattern that can be scanned in a single straight line, as in (c). The

differences between these can be seen in the aspect ratio of the bars in the various patterns, the number of bars per frequency, the layout of the pattern itself, whether it is a spiral or a rectangle displaying progression of spacing, and the actual numerical progression of different frequency patterns within the target. In many cases low contrast versions or reverse polarity (white and black parts are switched) are also available.

The second major class, the continuously varying frequency pattern, is exemplified by the Sayce chart (d) and the radial graded frequency chart (a). The Sayce patterns are particularly useful for automated readout, provided the appropriate phase information is obtained (coordinates of each black bar) to prevent the pseudoresolution phenomena described above.

3.5.3.2 Line Imaging Interactions

A strategy for evaluating line and text imaging against all of the above metrics is to establish a fixed threshold that optimizes system performance for one of the major categories, such as line width fidelity, and then to report performance for the other variables, such as detectability and resolving power, at that threshold. One may also choose to plot detectability, fidelity, and resolving power as a function of threshold on a single plot in order to observe the relationships among the three and find an optimum threshold, trading off one against the other. This is illustrated in Figure 3.34 for a particular scanner. Such a plot provides several useful perspectives relating to the effects of blur on a binary system. It is clear in this example that the maximum fidelity occurs between 35% and 45% threshold while the maximum fine-line detectability keeps growing as the threshold drops below 30%. The resolving power has a distinct maximum at about 33%. Such a plot is different for each system and is governed by the shape of the underlying MTF curves and the various nonlinear interactions produced by image processing and the electronics.

3.5.4 Binary Metrics Relating to Noise Characteristics

Conventional approaches to measuring the amplitude of the noise fluctuations using various statistical measures of the distribution are not appropriate for binary systems. In these systems the noise shows up as pixels that are of the wrong polarity; that is, a black pixel that should have been white or a white pixel that should have been black. In general it is the distribution and location of these errors that need to be characterized. The practical approach to this problem is to examine the noise in a context equivalent to the main applications of interest for the binary imaging system. The resulting metrics include:

1. The range of uncertainty associated with determining the threshold using a gray wedge as described above in Section 3.5.2, which has led to the *gray wedge metric for noise.*

2. The noise seen on edges of lines and characters, which has led to the *line edge range metric for noise.*

3. The characterization of noise in a halftone image, that is, *halftone granularity.*

These are all described below.

3.5.4.1 Gray Wedge Noise

Figure 3.33 shows the transmittance profile of a gray wedge as a function of distance. The thresholded image of that gray wedge is shown below this profile with the x-axis lined

up to correspond to the position in the profile plot. Here, as noted earlier The GWN (Gray Wedge Noise) = $T_2 - T_1$.

To make a statistically satisfactory measure of noise, a probability distribution is used with the criteria for determining the positions d_1 and d_2. As illustrated, these are the point where the signal is 95% black and the point where the signal is 95% white. Under the assumption of normally distributed noise this would represent plus or minus approximately two standard deviation limits on the noise distribution.

To fully characterize a binary system with this metric, one plots the width of the noise band in effective transmittance as a function of the independent variable, threshold (converted to transmittance).

3.5.4.2 Line Edge Noise Range Metric

The noise associated with the edges of lines is an important type of noise to be directly evaluated for many practical reasons. The image of every line has a microscopic gray region associated with it where the intensity falls off gradually from the white surround field into the black line. For the image of the edge of a line oriented at a very small angle to the sampling matrix there is a distribution of gray varying along the edge of the line. It gradually increases from white to black. In a binary system this scan line appears white until it reaches a fractional coverage along the tilted edge required by the threshold, and then changes to black. Much like the case associated with the wedge in the previous metric as the edge approaches the transition point where the threshold causes a change in the binary signal, the probability for an error resulting from noise increases. Thus the binary signal along the length of this slightly tipped line acts much like the signal for the wedge in the previous example, oscillating between black and white. This provides the basis for a second metric, which we refer to as the line edge noise range metric.

In Figure 3.36, a slightly tilted input line is shown relative to several scan lines. The binary video bit map for this line is shown in the lower part of Figure 3.36. Vertical lines mark the location at which the edge of the line makes a transition from the center of one raster line to the next. In the video bit map this transition is noisy and the two ranges in which this uncertainty of the black to white transition exists are indicated as N_1 and N_2. The centers
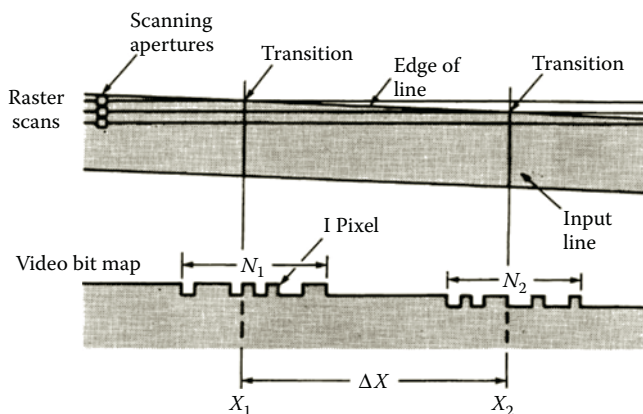
FIGURE 3.36
Scanning of a slightly tilted line, with the corresponding binary video bit map image, showing noise effects, which define edge noise range (ENR).
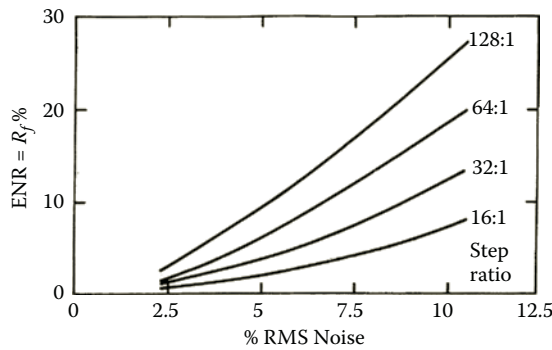
FIGURE 3.37
Relationship between LENR and RMS noise for various step ratios.

of these noisy transition regions are marked by the transition lines at X_1 and X_2 and are separated by the distance ΔX. The metric can be applied to edges of lines, or edges of solids, or any straight edge and is therefore generically referred to as "edge noise range" or ENR and simply defined as

$$\text{ENR} = \frac{\sum_{i=1}^{n} N_i}{\sum_{i=1}^{n-1} \Delta X_i}. \tag{3.16}$$

The numerator and the denominator are averages over a large number of transitions along one or more constantly sloping straight lines. ΔX is the number of pixels per "step." The range N is determined by subtracting the pixel number of the first white pixel in the black region from that of the last black pixel in the white region along the length of the line in each transition region.

Figure 3.37 shows the relationships among ENR, the step ratio ΔX, and the percentage of RMS noise in the imaging system, assuming additive white Gaussian noise distributions (private communication, J.C. Dainty, 1984) These are not intuitive relationships. For example, it should be noted from Figure 3.37 that an increase of a factor of 2 in RMS noise for a given angle line produces a line edge noise range increase of anywhere from $2\frac{1}{2}$ to 4× depending upon the slope of the line and the exact noise level. It is noted that the noise is highly dependent on the angle of the line. Gradually sloping lines not only produce a larger absolute range but also a larger fractional range.

Here the MTF of the imaging system was considered to be perfect. The effect of blur, that is, decreases in MTF, is to increase the magnitude of ENR above those values shown.

It must also be noted that document noise will create extra fluctuations along the edge of the line and also increase the length of the range.

3.5.4.3 Noise in Halftoned or Screened Digital Images

This is the binary situation where a gray signal is created by a gray scanner and then converted to a binary halftone signal via processing in order to print on a binary rendering device (i.e., it is not a characteristic of a binary scanner per se). Scanning a typical

photograph and applying a halftone screen of the type described earlier (2.2.3) results in a bit map in which some of the arrangements of pixels in the halftone cells do not follow the prescribed growth pattern for the screen (see Figure 3.38). The noise in the scanning system itself can produce pixel-by-pixel changes in the effect of threshold at each one of the sites within the halftone cell. Some of these errors are introduced by the partial dotting mechanism, described earlier, when the granularity of the otherwise uniform input document, which was scanned to create the image, is of sufficient contrast to change the structure inside areas formed by individual halftone cell's threshold matrix. See Section 3.2.2.3 and Figures 3.14 and 3.15 for a review of these mechanisms.

One way of evaluating this type of noise is to create images of a series of perfectly uniform patches of differing density and process them through the halftoning method of choice. One then measures the RMS fluctuation in the percent area coverage for the resulting halftone cells, one patch at a time. To the extent that the output system is insensitive to the orientation of the bit map inside the halftone cell, this fluctuation becomes a reasonable measure of the granularity of the digital halftone pattern. For the electronic image, it can be calculated with a simple computer program that searches out each halftone cell and calculates its area coverage, collecting the statistics over a large number of halftone cells.

There are many image analysis packages on the market that will find particles in a digital image and evaluate their statistical distribution. They are found in biology, medicine, or metallurgy software applications as well as in image analysis packages.[87–89] In this case a "particle" is a halftone dot whose area corresponds to the number of pixels.
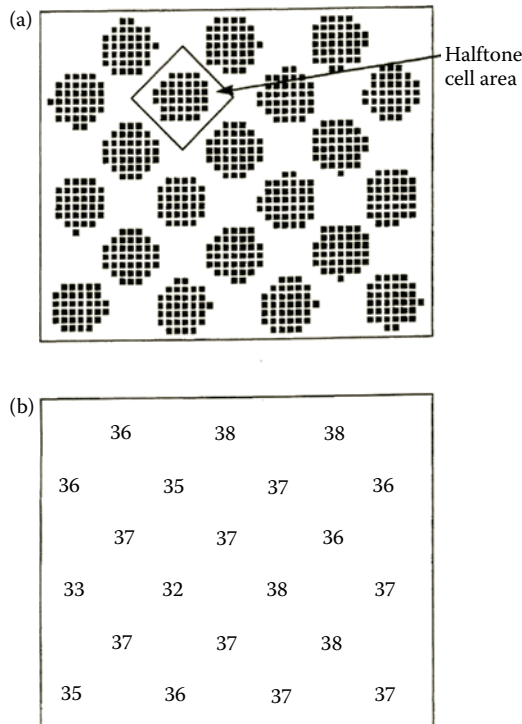
FIGURE 3.38
Noise (granularity) in a binary halftone image, part (a) Bitmap of the halftone rendering of a scanned image of a uniform area on an original. Part (b) Number of black pixels in each cell where the average number of black pixels per cell is 36.4 and the estimate of the standard deviation is 1.56 pixels.

If the image has been printed, a microdensitometer can be set up with an aperture that exactly covers one halftone cell, then scanned along rows of halftone dots. The RMS fluctuations or the low-frequency components of the noise spectra can then be evaluated. This is sometimes referred to as *aperture filtered granularity* measurement.

3.6 SUMMARY MEASURES OF IMAGING PERFORMANCE

Many attempts have been made to take the general information on image quality measurement and reduce it to a single measure of imaging performance. These often take the form of shortcut "D" in Figure 3.3. While none of the resulting measures provides a single universal figure of merit for overall subjective image quality, each brings additional insight to the design and analysis of particular imaging systems and its quality. Each has achieved some level of success in a limited range of applications. Preferred image quality, however, is a psychological reaction to a complex set of trade-offs and visual stimuli. There is a very subjective, application-oriented aspect to this reaction that does not readily lend itself to analytical description.

Instead, in an attempt to help the engineer control or design his systems, we shall describe a number of metrics.

The metrics are described here in their general form, many were developed for analog imaging systems such as cameras and film, others for digital. To the extent that the scanning systems in question are unaliased and have a large number of gray levels associated with them, the direct application of analog metrics is valid. In general, it should be remembered that digital imaging systems are not symmetric in slow and fast scan orientations in either noise or spatial frequency response (MTF). Therefore, what is given below in one-dimensional units must be applied in both dimensions for successful analysis of a digital input scanner. These concepts can be extended to an entire imaging system with little modification if the subsequent imaging modules, such as a laser beam scanner, provide gray output writing capability and generate no significant sampling or image conversion defects of their own (i.e., they are fairly linear). Since full gray scale input scanners are usually linear, most of these concepts can be applied to them, with the qualification that some display or analysis technique is required to convert the otherwise invisible electronic image to a visual or numerical form.

Most of these summary measures can be described by curves like the illustrative ones shown in Figure 3.39. Here we show some common measures, generically symbolized by F for both signal and noise, such as intensity, modulation or (modulation)^2 plotted as a function of spatial frequency f. A signal $S(f)$ is shown generally decreasing from its value at 0 spatial frequency to the frequency f_{max}. A limiting or noise function $N(f)$ is plotted on the same graph starting at a point below the signal; it too varies in some fashion as spatial frequency increases. The various unifying constructs (metrics) involve very carefully considered approaches to the relationship between S and N, to their respective definitions, and to the frequency range over which the relationship is to be considered, along with the frequency weighting of that relationship

3.6.1 Basic Signal-to-Noise Ratio

The simplest of all signal-to-noise measures is the ratio of the mean signal level $S(0)$ to the standard deviation $N(0)$ of the fluctuations at that mean. If the system is linear and the
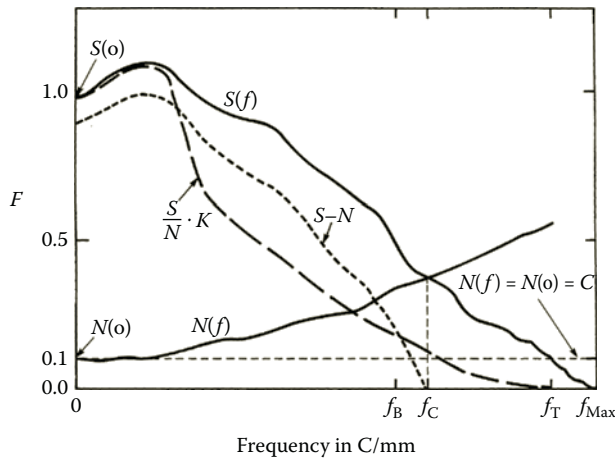
FIGURE 3.39
Signal, $S(f)$, noise, $N(f)$, and various measures of the relationships between them, plotted as functions of spatial frequency. Various critical frequencies are noted as points on the frequency axis.
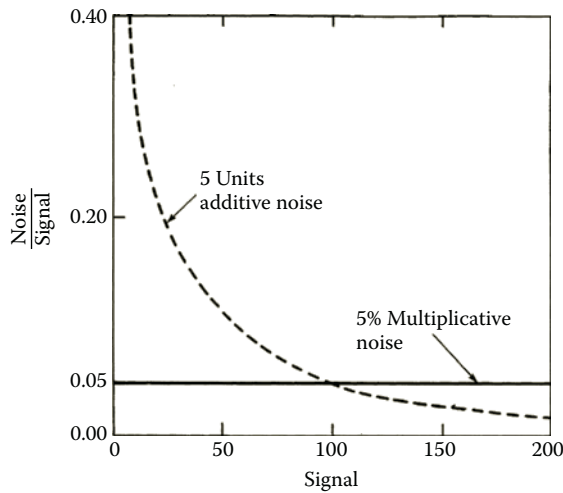
FIGURE 3.40
Relative noise (noise/signal) as a function of signal level for additive and multiplicative noise.

noise is multiplicative, this is a useful single number metric. If the noise varies with signal level, then this ratio is plotted as a function of the mean signal level to get a clearer picture of performance. Hypothetical elementary examples of this are shown in Figure 3.40, in which are plotted both the multiplicative type of noise at 5% of mean signal level (here represented as 100) and additive noise of 5% with respect to the mean signal level. It can be seen why such a distinction is important in evaluating a real system. It should be noted that in some cases, multiplicative or additive noise might vary as a function of signal level for some important design reason.

In comparing signals to noise, one must also be careful, to ensure that the detector area over which the fluctuations are collected is appropriate for the application to which the signal-to-noise calculation pertains. This could be the size of the input or output pixel, of

the halftone cell, or of the projected human visual spread function. The data must also be collected in the orientation of interest. In general, for scanned imaging systems there will be a different signal-to-noise ratio in the fast scanning direction than in the slow scanning direction.

3.6.2 Detective Quantum Efficiency and Noise Equivalent Quanta

When low light levels or highly noise limited situations occur, it is desirable to apply the concepts of detective quantum efficiency (DQE) and noise equivalent quanta (NEQ). These fundamental measurements have been extensively discussed in the literature.[47]

If we set gain of a system response to a constant r in arbitrary output units, and assume the distribution of fluctuations obeys normal statistics, then we can write

$$DQE = \frac{r^2 q}{\sigma_0^2} \qquad (3.17)$$

$$NEQ = \frac{r^2 q^2}{\sigma_0^2} \qquad (3.18)$$

where σ_0 represents an estimate of the standard deviation of the distribution of the output fluctuations. Here for DQE the average signal level q is divided by the square of the standard deviation (i.e., the variance) and contains a modifier that is related to the characteristic amplification factors associated with a particular detection system, namely r, which also enters as a square. It should also be pointed out that DQE is an absolute measure of performance, since q is an absolute number of exposure events, that is, number of photons or quanta.

For illustration purposes, Figure 3.39 shows all of the above constructs.

3.6.3 Application-Specific Context

The above descriptions are frequently derived from the fundamental physical characteristics of various imaging systems, but the search for the summary measure of image quality usually includes an attempt to arrive at some application-oriented subjective evaluation, correlating subjective with objective descriptions. Applications that have been investigated extensively include two major categories: those involving detection and recognition of specific types of detail and those involved in presenting aesthetically pleasing renderings of a wide variety of subject matter. These have centered on a number of imaging constraints, which can usually be grouped into the categories of display technologies and hard-copy generation. Many studies of MTF have been applied to each.[85,97,122–126] All of these studies are of some interest here. Note that modern laser beam scanning tends to focus on the generation of hard copy where the raster density is hundreds of lines per inch and thousands of lines per image compared with the hundreds of lines per image for early CRT technology used in the classical studies of soft display quality.

3.6.4 Modulation Requirement Measures

One general approach characterizes $N(f)$ in Figure 3.39 as a "demand function" of one of several different kinds. Such a function is defined as the amount of modulation or signal required for a given imaging and viewing situation and a given target type. In one class

of applications, the curve $N(f)$ is called the threshold detectability curve and is obtained experimentally. Targets of a given format but varying in spatial frequency and modulation are imaged by the system under test. The images are evaluated visually under conditions and criteria required by the application. Results are stated as the input target modulation required (i.e., "demanded") for being "just resolved" or "just detected" at each frequency. It is assumed that the viewing conditions for the experiment are optimum and that the threshold for detection of any target in the image is a function of the target image modulation, the noise in the observer's visual system, and the noise in the imaging system preceding the observer. At low spatial frequencies this curve is limited mostly by the HVS, while at higher frequencies imaging system noise as well as blur may determine the limit.

One such type of experiment involves measuring the object modulation required to resolve a three-bar resolving power target. For purposes of electronic imaging, it must be recalled that the output video of an input scanner cannot be viewed directly, and therefore any application of this method must be in the systems context, including some form of output writing or display. This would introduce additional noise restrictions. The output could be a CRT display of some type, such as a video monitor with gray-scale (analog) response. Another likely output would be a laser beam scanner writing on xerography or on silver halide film or paper. The details for measuring and using the demand function can be found in work by Scott[127] for the example of photographic film and in Biberman,[85] especially Chapter 3 for application to soft displays.

3.6.5 Area under the MTF Cure (MTFA) and Square Root Integral (SQRI)

Modulation detectability, while useful for characterizing systems in task-oriented applications, is not always useful in predicting overall image quality performance for a broad range of imaging tasks and subject matters. It has been extended to a more general form through the concepts of the threshold quality factor[128] and area under the MTF curve (MTFA).[129,130] These concepts were originally developed for conventional photographic systems used in military photo-interpretation tasks.[128] They have been generalized to electro-optical systems applications for various forms of recognition and image-quality evaluation tasks, mostly involving soft displays.[85] The concept is quite simple in terms of Figure 3.39. It is the integrated area between the curves $S(f)$ and $N(f)$ or, equivalently, the area under the curve labeled $S-N$. In two dimensions, this is

$$\text{MTFA} = \int_0^{f_{cx}} \int_0^{f_{cy}} [S(f_x, f_y) - N(f_x, f_y)] \, df_x \, df_y \tag{3.19}$$

where S is the MTF of the system and N is the modulation detectability or demand function as defined above, and f_{cx} and f_{cy} are the two-dimensional "crossover" frequencies equivalent to f_c shown in Figure 3.39.

This metric attempts to include the cumulative effects of various stages of the scanner, films, development, the observation process, the noise introduced into the perceived image by the imaging system, and the limitations imposed by psychological and physiological aspects of the observer by building all these effects into the demand function $N(f)$. Extensive psychophysical evaluation and correlation has confirmed the usefulness of this approach[130] for recognition of military reconnaissance targets, pictorial recognition in general, and for some alphanumeric recognition.

Related approaches using a visual MTF weighting have been successfully applied to a number of display evaluation tasks, showing good correlation with subjective quality.[123]

Many studies examine differences in quality where noise factors are relatively constant. One of these is the square root integral (SQRI) model of Bartend.[124,131] Here, the demand function is specified by a general contrast sensitivity of the HVS and the comparison of the quality of two images of interest is specified in JND units (see Section 3.8 for a definition of JND, a just noticeable difference).

$$J = \frac{1}{\ln 2} \int_0^{W_{\max}} \sqrt{\frac{M(w)\,d(w)}{M_t(w)w}} \tag{3.20}$$

where $M(w)$ is the cascaded MTF of the image components, including that of the display and $M_t(w)$ is threshold MTF of the HVS, both in units of angular spatial frequency w. Results are to be interpreted with the understanding that 1 JND is "practically insignificant." It is equal to a 75% correct response in a paired comparison experiment. Note that 3 JND is "significant," and 10 JND is "substantial."[124,132] The $M_t(w)$ term describes the HVS as the threshold contrast for detecting a grating of angular frequency w as follows:

$$1/M_t(w) = aw \exp(-bw)\sqrt{1 + c\exp(bw)} \tag{3.21}$$

where

$$a = \frac{540(1 + 0.7/L)^{-0.2}}{1 + 12/[s_w(1 + w/3)^2]} \tag{3.22a}$$

$$b = 0.3(1 + 100/L)^{0.15} \tag{3.22b}$$

$$c = 0.6 \tag{3.22c}$$

L is the display luminance in cd/m² and s_w is the display size or width in degrees. These equations have been shown to have high correlation with perceived quality over a wide range of display experiments,[124] one of which is shown in Figure 3.41. Here the resolution, size, and subject matter of projected slides were varied and the equation was fit to the data.

It is noted by Barten that noise should be taken into account in the modulation threshold function, which is done by using a root sum of squares method of a weighted noise modulation factor to $M_t(w)$.[121] Other authors have expanded on these concepts, extending them to include more fundamentals of visual mechanisms.[125,126,133]

3.6.6 Measures of Subjective Quality

Several authors have explored the broader connection between objective measures of image quality and overall aesthetic pictorial quality for a variety of subject matters encountered in amateur and professional photography.[134–140] The experiments to support these studies are difficult to perform, requiring extremely large numbers of observers to obtain good statistical measures of subjective quality. The task of assessing overall quality is less well defined than the task of recognizing a particular pattern correctly, as evaluated in most of the studies cited above. It would appear that no single measurement criterion has become
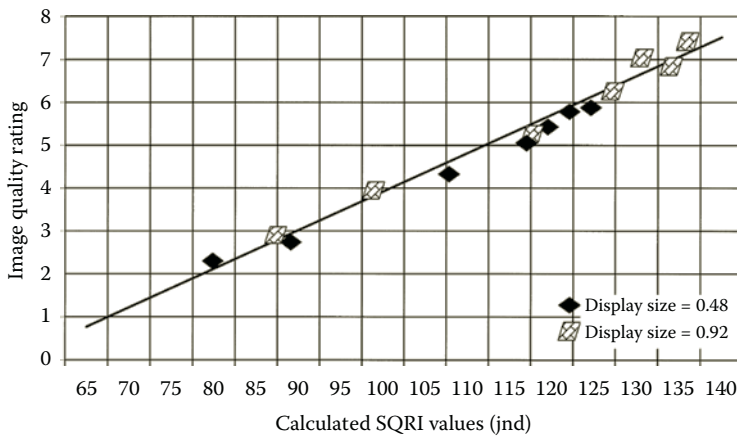
FIGURE 3.41
Linear regression between measured subjective quality and calculated SQRI values for projected slides of two different sizes, as indicated, illustrating the good fit. (Adapted from Barten, P.G.J. The Square Root Integral (SQRI): A new metric to describe the effect of various display parameters on perceived image quality. *Proceedings of SPIE conference on Human Vision, Visual Processing, and Digital Display, Los Angeles, CA,* 1989; Vol. 1077, 73–82.)

universally accepted by individuals or organizations working in this area. Below we shall discuss a few of the key descriptors, but we do not attempt to list them all.

Many of the earlier studies tended to focus on the signal or MTF-related variable only. In one such series of studies,[134,135] S in Figure 3.39 is defined as the modulation of reflectance on the output print (for square waves) divided by the modulation on the input document (approximately 0.6 for these experiments). The quality metric is defined as the spatial frequency at which this ratio falls to 0.5. This is indicated in Figure 3.39 by the frequency f_b for the curve S as drawn. In these studies, a landscape without foreground was rated good if this characteristic or critical frequency was 4–5 cycles/mm (100–125 cycles/in), but for a portrait 2–3 cycles/mm (50–75 cycles/in) proved adequate. Viewing distance was not a controlled variable. By using modulation on the print and not simply MTF, the study has included the effects of tone reproduction as well as MTF. Granularity was also shown to have an effect, but was not explicitly taken into account in the determination of critical frequency.

Several studies have shown that the visual response curve discussed earlier can be connected with a measure of $S(f)$ to arrive at an overall quality factor. See, for example, system modulation transfer acutance (SMT acutance) by Crane[139] and an improvement by Gendron[140] known as cascaded area modulation transfer (CMT) acutance. One metric, known as the subjective quality factor (SQF),[136] defines an equivalent passband based on the visual MTF having a lower (initial) cutoff frequency at f_i and an upper (limiting) frequency of f_l. Here, f_i is chosen to be just below the peak of the visual MTF, and f_l is chosen to be four times f_i (two octaves above it). For prints that are to be viewed at normal viewing distance [i.e., about 340 mm (13.4 in)], this range is usually chosen to be approximately 0.5–2.0 cycles/mm (13–50 cycles/in).

The MTF of the system is integrated as follows.

$$\text{SQF} = \int_{f_{x=0.5}}^{2}\int_{f_{y=0.5}}^{2} S\big(f_x,\,f_y\big)\mathrm{d}\big(\log_{10}f_x\big)\mathrm{d}\big(\log_{10}f_y\big). \tag{3.23}$$

This function has been shown to have a high degree of correlation with pictorial image quality over a wide range of picture types and MTFs. It is possible that a demand function similar to that described in the MTF concepts above could be applied to further improve the performance. The SQF metric is applied to the final print as it is to be viewed and may be scaled to the imaging system, when reduction or enlargement is involved, by applying the appropriate scaling factor to the spatial frequency axis.

It should be noted that there is a significant difference between the upper band limit of this metric at 2 cycles/mm (50 cycles/in) and the critical frequency described above in Biedermann's work for landscapes, which is in the 4–5 cycles/mm (100–125 cycles/in) region. But there is good agreement for the portrait conclusions of the earlier work, which cites an upper critical frequency of 2–3 cycles/mm (50–75 cycles/in). Authors of both metrics acknowledge the importance of granularity or noise without directly incorporating granularity into their algorithms. Granger[141] discusses some effects of granularity and digital structure in the context of the SQF model, but calls for more extensive study of these topics before incorporating them into the model.

It is clear that when the gray content and resolution of the digital system are high enough to be indistinguishable from an analog imaging system, then these techniques, which are general in nature, should be applicable. The quantization levels at which this equivalence occurs vary broadly. Usually 32 to 512 levels of gray suffice, depending on noise (higher noise requires fewer levels), while resolution values typically range from 100 to 1000 pixels/in (4–40 pixels/mm), depending on noise, subject matter, and viewing distance.

Another fairly typical approach to quantifying overall subjective image quality involves measuring the important attributes of a set of images made under a range of technology variables of interest and then surveying a large number of observers, usually customers for the products using the technology. They are asked for their overall subjective reaction to each image. A statistical regression is then performed between the measured attributes and the average subjective score for each image. This is the "type D shortcut" illustrated in Figure 3.3. An equation describing quality is derived using only the most important terms in the regression, that is, those that describe most of the variance. The "measures" may also be visual perceptions, that is, the "nesses," in which case the result is Engeldrum's "image quality models,"[12,41] but must include all the factors that could have any reasonable bearing on quality. Sometimes the technology variables themselves are used (type A shortcut, Figure 3.3). This makes the resulting equation less general in its applicability but gives immediate answers to product questions.

Below is an example of an *image quality model*[142] selecting five visual perception attributes from a list of 10 general image quality attributes[143] to describe a series of 48 printed color images from lithography, electrophotography, inkjet, silver halide, and dye diffusion, under a wide variety of conditions. A linear regression against overall preferences of 61 observers yielded the following equation.

$$\begin{aligned} \text{Avg. Preference} = {} & 8.8 \text{ Color Rendition} + 5.5 \text{ MicroUniformity} \\ & + 4.4 \text{ Effective Resolution} + 3.5 \text{ MacroUniformity} \\ & + 1.9 \text{ Gloss Uniformity}. \end{aligned} \tag{3.24}$$

A plot showing the Preference versus a fitted three-dimensional surface for the top two correlates is given in Figure 3.42.
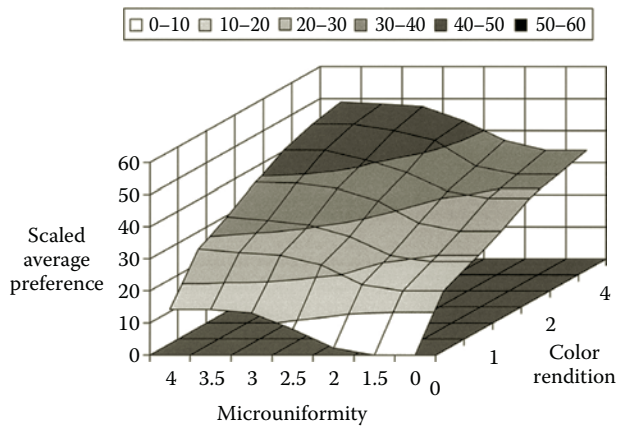
FIGURE 3.42

Illustration of the multivariate nature of a typical image quality model showing relationship between scaled image quality preference and two of several variables: color rendition and microuniformity. (Adapted from Natale-Hoffman, K.; Dalai, E.; Rasmussen, R.; Sato, M. *Proceedings of IS&T Image Processing, Image Quality, Image Capture Systems (PICS) Conference, Savannah, GA,* 1999; 266–273.)

Other regression equations between physical or measured image parameters and customer preference have been developed in many different imaging environments some tackling huge lists of variables.[144]

$$\text{Portrait IQ rating} = 0.393 \times 10^{-8} \left(C^* \text{ of red } 100\% \right)^{5.2}$$
$$+ 69.51 \times \exp(-0.125 \times \text{graininess of cyan } 60\%)$$
$$- 0.000173 \times \left(H^0 \text{of } 1.0 \text{ Neutral Solid } - 305.0 \right)^2$$
$$- 0.409 \times \left(C^* \text{ of blue } 10\% - 4.90 \right)^2$$
$$+ 47.7 \times \exp(-0.0766 \times \text{graininess of skin color})$$
$$- 0.0197 \times \left(C^* \text{ of blue } 40\% - 23.5 \right)^2$$
$$- 0.0452 \times \left(C^* \text{ of Cyan } 70\% - 36.8 \right)^2$$
$$- 15.22$$

A study comparing perceived image quality and acceptability of photographic prints of images from different resolution digital capture devices[77] directly compared perceived quality by a range of individuals with varying experience in photography and computers. Photographic prints (4 × 6 in) showing optimum tone and color rendering were used as output for viewing. Their results and others were given earlier in Figure 3.20.

3.6.7 Information Content and Information Capacity

There are numerous articles in the imaging science literature that analyze imaging systems in terms of information capacities and describe their images as having various information contents.

Using basic statistics of noise and spread function concepts from Section 3.2.1.3 a simple description of image information is given by Equation 3.33.[46,50,51] It defines image

information, *H*, as

$$H = a^{-1} \log_2 \left[\frac{\text{probability of "density message" being correct}}{\text{probability of a specific density as input}} \right] \tag{3.25}$$

where *a* is the area of the smallest resolvable unit in the image (i.e., 2 × 2 pixels based on unaliased sampling from the sampling theorem) and the log factor is from the classic definition of information in any message,[145] here being messages about density (any other signal units can be used if done so consistently and they constitute a meaningful message in some context). To convert this into more useful terms let

$$H = a^{-1} \log_2 \left[\frac{p}{1 - M} \right] \underset{p \to 1}{\cong} \left[\frac{\log_2 M}{a} \right] \tag{3.26}$$

where the numerator is set equal to *p*, the probability that a detected level within a set of levels is actually the correct one (i.e., the reliability), and *M* is the number of equally probable distinguishable levels (i.e., the quantization) from Equation 3.1 in Section 3.2.1.3. Assuming a high reliability such that *p* approaches unity, the simplification on the right results. The standard deviation of density in Equation 3.1 must be measured with a measuring tool whose aperture area is equal to *a*.

An approximation useful in comparing different photographic materials uses the standard deviation of density σ_a at a mean density of approximately 1 to 1.5 and Equation 3.30 results:

$$H = a^{-1} \log_2 \left[\frac{L}{6 s_a} \right] \tag{3.27}$$

where *k* was set to ± 3 (= 6), leading to *p* = 0.997 (~1); *L* is the density range of the imaging material.

Since the standard deviation of density is strongly dependent on the mean density level, it is more accurate and also common practice to measure the standard deviation at several average densities and segment the density scale into adjacent, empirically determined, unequal distinguishable density levels. These levels are separated by *k* standard deviations of density as measured for each specific level.[46,50,51] If the input scanner itself is very noisy, then the σ_a term must represent the combined effects of both input noise and scanner noise. This was covered in Section 3.4.3 (see References 46, 50, and 51 for further information).

Another approach uses all of the spatial frequency based concepts developed above for the MTF and the Wiener spectrum and can incorporate the HVS as well. It produces results in bits/area that are directly related to the task of moving electronic image data from an input scanner to an output scanner or other display. Much of the research in this area began on photographic processes, but has also been applied to electronic scanned imaging. Both are addressed here. The basic equation for the spatial frequency based information content of an image is given[146] by

$$H_i = \frac{1}{2} \int_{-\infty}^{\infty} \log_2 \left[1 + \frac{\Phi_S(f)}{\Phi_N(f)} \right] df \tag{3.28}$$

where H_i is the information content of the image, Φ_S is the Wiener spectrum of the signal, Φ_N is the Wiener spectrum of the system noise, and f is the spatial frequency, usually given in cycles per millimeter. This equation is in one-dimensional form for simplicity, in order to develop the basic concepts. For images, these concepts must, as usual, be extended to two dimensions. Unlike the work dealing with the photographic image, the assumption of uniform isotropic performance cannot be used to simplify the notation to radial units. For digital images the separation of the orthogonal x- and y-dimensions of the image must be preserved.

Alternative methods for calculating information capacity do not include explicit spatial frequency dependence but do explicitly handle probabilities.[46,50,51,147] They served as the basis for our discussion of quantization and Equation 3.29, is rewritten as

$$H_i = N \log_2(pM) \tag{3.29}$$

where N is the number of independent information storage cells per unit area. It may be set equal to the reciprocal of the smallest effective cell area of the image, for example, a number of pixels or the spread function. Here, p is the reliability with which one can distinguish the separate messages within an information cell, and M is the number of messages per cell. M is determined by the number of statistically different gray levels that can be distinguished in the presence of system noise at the reliability p, using noise measurements made with the above cell area.

Generalizing Equations 3.29 and 3.1 to the "generic" units of Figure 3.39, L is set equal to S_0 and σ_a is set equal to σ_s for the maximum signal and its standard deviation, respectively. We select a spread function for an *unaliased* system equal to 2×2 pixels and translate this to frequency space using the reciprocal of the sampling frequencies f_{sx} and f_{sy} in the x- and y-directions. This gives a generalized, sampling-oriented version of Equations 3.29 and 3.32 as

$$H_i = \frac{f_{sx} f_{sy}}{4} \log_2 \left[\frac{S_0}{k \boldsymbol{s}_s} \right] \tag{3.30}$$

where S and σ_s are measured in the same units. k can be set to determine the reliability for a given application. Values from 2^{45} to 20^{43} have been proposed for k for different applications; 6 is suggested here, making $p = 0.997$. This assumes that σ_s is a constant (i.e., additive noise) at all signal levels. If not, then the specific functional dependence of σ_s on S must be accounted for in determining the quantity in the brackets, measuring the desired number of standard deviations of the signal at each signal level over the entire range.[46,51] While this approach predicts text quality and resolving power[147] and deals with the statistical nature of information, it does not (as noted above) permit the strong influence of spatial frequency to be handled explicitly.

Equation 3.31 may be expanded to illustrate the impact of the MTF on information content, giving

$$H_i = \frac{1}{2} \int_{-\infty}^{\infty} \log_2 \left[1 + \frac{K^2 \Phi_i(f) |\mathrm{MTF}(f)^2|}{\Phi_N(f)} \right] df \tag{3.31}$$

where $\Phi_i(f)$ is the Wiener spectrum of the input scene or document and MTF(f) is the MTF of the imaging system (assumed linear) with all its components cascaded. At this point we need to begin making some assumptions in order to carry the argument further. The constant K in the equation is actually the gain of the imaging system. It converts the units of the input spectrum into the same units as the spectral content of the noise in the denominator. For example, a reflectance spectrum for a document may be converted into gray levels by a K factor of 256 when a reflectance of unity (white level) corresponds to the 256th level of the digitized (8-bit) signal from a particular scanner; the noise spectrum is in units of gray levels squared.

Various authors have gained further insight into the use of these general equations. Some of those investigating photographic applications have extended their analysis to allow for the effect of the visual system;[137] others attempted to apply some rigor to the terms in the equation that are appropriate for digital imaging.[148,149] Others have worked on image quality metrics for digitally derived images,[138] but some have tended to focus on the relationship to photointerpreter performance.[150]

Several of these authors have suggested that properly executed digital imagery does not appear to be greatly different from standard analog imagery in terms of subjective quality or interpretability. One almost always sees these images using some analog reconstruction process to which many analog metrics apply. It therefore seems reasonable to combine some of this work into a single equation for image information and to hypothesize that it has some direct connection with overall image quality when applied to a scanner whose output is viewed or printed by an approximately linear display system. It must also be assumed that the display system noise and MTF are not significant factors or can be incorporated into the MTF and noise spectra by a single cascading process. A generic form of such an equation is given below as Equation 3.35 without the explicit functional dependencies on frequency in order to show and explain the principles that follow (expanding on the analysis in Reference 137).

$$H_i = \frac{1}{2}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \log_2\left[1 + \frac{K^2\Phi_i\,|\mathrm{MTF}|^2\,R_1^2}{\left\{1 + 12\left(f_x^2 + f_y^2\right)\right\}\left\{\left[\Phi_a + \Phi_n + \Phi_q\right]R_2^2 + \Phi_E\right\}}\right]\mathrm{d}f_x\mathrm{d}f_y. \qquad (3.32)$$

Let us begin by examining the numerator. Several authors have attempted to multiply the MTF of the imaging system by a spatial frequency response function for the HVS to arrive at an appropriate weighting for the signal part of Equation 3.34. Kriss and his coworkers[137] observed that a substantial increase in the enhancement beyond the eye's peak response produced larger improvements in overall picture quality than did equivalent increases in enhancement at the peak of the eye's response. The pictures with large enhancement at the eye's peak response were "sharper," but were also judged to be too harsh. These results indicate that the HVS does not act as a passive filter and that it may weigh the spatial frequencies beyond the peak in the eye's response function more than those at the peak.

Lacking a good model for the visual system's adaptation to higher frequencies as described above, Kriss et al. proposed the use of the reciprocal of the eye frequency response curve as a weighting function, $R_1(f)$, that could be applied to the numerator. The conventional eye response $R_2(f)$ should be applied in the denominator to account for the perception of the noise, since the eye is not assumed to enhance noise but merely to filter it. The noise term, ΦN, in the earlier equation has been replaced by the expression in the square brackets and multiplied by $R_2^2(f)$. The reciprocal response, R_1, is set equal to 0.0 at 8 cycles/mm (200 cycles/in) in order to limit this function.

Next let us examine the noise effects themselves. A major observation is that noise in the visual system, within one octave of the signal's frequency, tends to affect that signal. It can be shown that the sum in the first curly brackets in the denominator of Equation 3.35 provides a weighting of noise frequencies appropriate to this one-octave frequency-selective model for the visual system.[137] Several authors[116,126,133,151] describe frequency-selective models of the visual system. The present construct for noise perception was first described by Stromeyer and Julesz.[152] A term for the Wiener spectrum of the noise in the visual process, Φ_E, has been added to the second factor in the denominator to account for yet one more source of noise. It is not multiplied by the frequency response of the eye, since it is generated after the frequency-dependent stage of the visual process.

The factor in square brackets in the denominator contains three terms unique to the digital imaging system.[149] These are Φ_a, the Wiener spectrum of the aliased information in the passband of interest; Φ_n, the Wiener spectrum of the noise in the electronic system, nominally considered to be fluctuations in the fast scan direction; and Φ_q the quantization noise determined by the number of bits used in the scanning process. We have thus combined in Equation 3.35 important information from photographic image quality studies, including vision models and psychophysical evaluation, with scanning parameters pertinent to electronic imaging.

The study of information capacity, information content, and related measures as a perceptual correlate to image quality for digital images is an ongoing activity. By necessity it is focused on specific types of imaging applications and observer types. For example, an excellent database of images and related experiments on quality metrics was built for aerial photography as used by photointerpreters.[150]

Experiments correlating subjective quality scores with the logarithm of the basic information capacity, taking the log of H_i as defined in Equations 3.1 and 3.29 showed correlation of 0.87 and greater for subjective quality of pictorial images.[153] Specific MTF and quantization errors were studied. The results were normalized by the information content of the original.

By use of various new combinations of the same factors discussed above, it was possible to obtain even higher correlations. A digital quality factor was defined[153] as

$$\text{DQF} = \left[\frac{\int_0^{f_n} \text{MTF}_s(f) \text{MTF}_v(f) d(\log f)}{\int_0^{10} \text{MTF}_v(f) \, d(\log f)} \right] \times \log_2 \left[\frac{L}{L/M + 2s} \right] \tag{3.33}$$

where we retain the one-dimensional frequency description used for simplicity by the original authors and the subscripts "s" and "v" refer to the system under test and the visual process, respectively. L is the density range of the output imaging process, f_n is the Nyquist frequency, M is the number of quantization levels, and σ is the RMS granularity of the digital image using a $10 \times 1000 \ \mu m$ microdensitometer slit. The first factor is related to the SQF described in Equation 3.27, and the second factor is related to the fundamental definition of image information capacity in Equation 3.29. A correlation coefficient of 0.971 was obtained for these experiments, using student observers and pictures showing a portrait together with various test patterns. It must be noted, however, that information capacity or any of these information-related metrics cannot be accepted, without psychophysical verification, as a general measure of image quality when different imaging systems or circumstances are to be compared.[15,154] Since systems models are used to determine MTFs and information capacities and hence arrive at useful descriptions of technology variables,

these are good examples of the type A shortcut regression models described in Figure 3.3, but are restricted to the limitations of such regression shortcuts.

In conclusion, this brief overview of specific quality metrics should give the reader some perspectives on which ones may be best suited to his or her needs The variety of these metrics, and the considerable differences among them, are evidence of the inherent diversity of imaging applications and requirements. Given this diversity, together with the large and rapidly expanding range of imaging technologies, it is hardly surprising that no single universal measure of quality has been found.

3.7 SPECIALIZED IMAGE PROCESSING

Most scanned images either begin or end in a digital form that needs to be efficiently managed in the larger context of a computer system, often in a network with other devices. This brings other dimensions to scanned image quality, namely the need to control the size of the files and the quality of the scanned images beyond the devices themselves. Controlling the file size is the subject of image compression.[11,155–157] Compression is an image quality issue because several methods do so at the expense of image quality, with lossy compression being one example and reduced sampling versus increased gray resolution, that is, resolution enhancement,[55] being another. Finally there is the color management challenge: finding a method to ensure that a color scanned image created by any of a number of scanners will look well when printed on any of a number of differently designed or maintained color printing devices.[6,14]

3.7.1 Lossy Compression

Image compression is a technology of finding efficient representation[155,156] for digital images to:

1. Reduce the size and cost of on-board or off-line computer memory and disk drive space required for their storage;
2. Reduce the bandwidth and or time needed to manipulate, send, or receive images in a communication channel; and
3. Improve effective access time when reading from storage systems.

The need to improve storage is easily seen in the graphic arts business, where an 8.5 × 11 in, 600 × 600 in, 32 bit color image is approximately 10^9 (or a billion) bits/image. Even the good quality portable amateur still cameras require 6+ megabytes (1 byte = 8 bits) per color image. Needless to say, transmitting such large files or accessing them takes tremendous amounts of time or bandwidth. Many standards groups are actively trying to create order out of the plethora of possible compression methods in order to reduce the number and types of tools needed to work in our highly interactive world of communications and networks.

There are generally two types of compression: lossless and lossy. Lossless takes advantage of better ways to encode highly redundant spatial or spectral information in the image, such as many contiguous white pixels in a text document. Results vary from compression ratios well over 100:1 on some text to 1.5:1 or less on many pictures. Group 3 and

Group 4 facsimile ("fax") standards, established by the CCITT (Consultative Committee of the International Telephone and Telegraph), now ITU-T, the Telecommunication Standardization sector of the International Telecommunication Union) are perhaps the best known and apply only to binary images.[157] Other standards include JBIG,[11,158] which is especially important for black and white halftones where it achieves about 8:1 compression while best CCITT methods actually expand file size by almost 20% over the uncompressed version.[11,159] See Table 3.9 for links to these standards groups and the latest upgrades.

All compression involves several different operations from transformation of the data to allow for efficient coding (e.g., discrete cosine transform (DCT)) to the actual symbol-encoding step where many technologies have developed. The latter include Huffman,[160] LZ,[161–163] and LZW[164] encoding, which are often cited as important parts of complete compression schemes.

Lossy compression is important from an image quality perspective since it removes information contained in the original image and therefore potentially causes a reduction in image quality to gain a compression advantage. Sometimes lossy compressions are said to be "visually lossless" in that they only give up information about the original that they claim cannot be detected by the HVS (recall the limits discussed in Section 3.2.1.3). Simply invoking binary imaging, for example, is an excellent method of compression, which is visually lossless when scanning ordinary black text on a white substrate at high resolution. It reduces a gray image from 8 bits to one and preserves all the edge information if it is high enough in resolution while throwing away all the useless gray levels in between. It does not work well on a photograph, where the primary information is in the tones that are all lost! Most lossy compression methods are very complex, involving advanced signal processing and information theory[6,17,157] beyond the scope of this chapter.

The best-known lossy compression technique is called JPEG (after the Joint Photographic Experts Group formed under the joint direction of ISO-IEC/JTC1/SC2/WG10—see Table 3.9—and CCITT SGVIII NIC in 1986).[11] It is aimed at still-frame, continuous tone, monochrome, and color images. In the case of JPEG, the underlying algorithm is a DCT of the image one 8 × 8 pixel cell at a time. It then makes use of the frequency-dependent quantization sensitivity of the eye (Figure 3.11) to alter the quantization of the signal on a frequency-by-frequency basis within each cell.

Many lossy compression methods are adjustable depending on the users' needs, so that the amount of compression is proportional to the amount of loss. They can be adjusted to a visually lossless state or to some acceptable state of degradation for a given user or design intent. The JPEG technique is adjustable by programming a table of coefficients in frequency space, called a Q table, which specifies the quantization at each of several spatial frequency bands. It can also be adjusted using a scaling factor applied to the Q table. Psychometric experiments (see Section 3.8) should be employed to determine acceptable performance in making such changes, using the exact scanning and marking methods and objects of interest. There are many other features of the JPEG approach that cannot be covered here. It has routinely been able to show an order of magnitude better compression over raw continuous tone pictures[11] with very little to no apparent visual loss of quality.

As noted earlier, compression is often aimed at improving communication of data and as such it is closely linked with file formats. In recent years a heavy focus on both the Internet and fax[159,165] has led to significant progress. JPEG and GIF have become widely used in the Internet,[165] where, in a greatly simplified view, it is seen that the former is lossy in spatial terms while the latter is lossy in color terms.

Color fax standards[159,166] have recently been developed in which the color, gray, and bitonal information is encoded into multiple layers for efficient transmission and compression.

These are formally known as TIFF-FX formats and generally fall into the broad category of mixed raster content or MRC.[159]

A new standard, JPEG 2000[167] has been developed, which, in addition to several other improvements, utilizes wavelets as an underlying technology and includes several optional file formats called the JP family of file formats. One, the JPM file format[168] with extension .jpm is aimed at compression of compound images, those having multiple regions each with differing requirements for spatial resolution and tonality. It employs this multiple layer approach. MRC formats allow the optimization of image quality, color quality via good color management, and best compression, all in one package. The base mode of MRC decomposes a mixed content image into three layers: a bitonal (binary) Mask layer, and color Foreground, and Background layers. The wavelet approach in JPEG 2000 causes less objectionable artifacts than the DCT-based baseline JPEG.[168]

3.7.2 Nonlinear Enhancement and Restoration of Digital Images

The characteristics of a scanned image may be altered in nonlinear ways to enable its portability between output devices of different resolutions while maintaining image quality and consistency of appearance. This may also be done to improve quality by reducing sampling effects or otherwise enhancing image appearance when compared to a straightforward display or print of the bit map. These are the general goals of digital image enhancement and restoration, topics that have been covered extensively in the literature and pursued by many imaging and printer corporations. They have been summarized by Loce and Dougherty.[55] Many of the techniques fall in the domain of morphological image processing,[3,169] which treats images as collections of well-defined shapes and operates on them with other well-defined shapes. It is most often used with binary images where template matching, that is, finding an image shape that matches the filter shape and then changing some aspect of the image shape, is a good general example. Two particular examples illustrate some of the underlying concepts.

"Anti-aliasing" is a class of operations in which "jaggies" or staircases (i.e., sampling artifacts or "aliased" digital images of tilted lines) in binary images are reduced to a less objectionable visual form. In Figure 3.43 the staircased image of a narrow line is analyzed by a filter programmed to find the jaggies (template matching) and then operated on, pixel by pixel, to replace certain all-white or all-black edge pixels with new pixels, each at an appropriate level of gray, in this case one of three levels. The gray pixels may be printed using
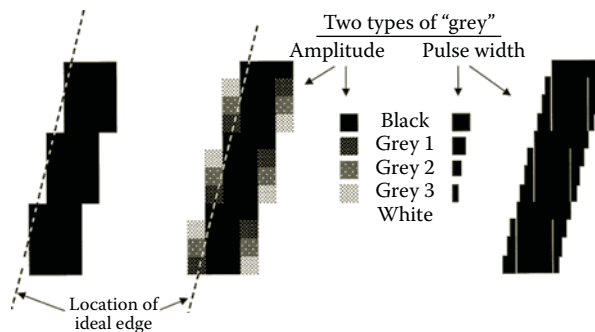
FIGURE 3.43
Anti-aliasing by the amplitude and pulse-width methods. Pixels narrowed by pulse-width changes are shown separated from the full pixels by a narrow white line only to illustrate where each is located.

conventional means of gray writing such as varying exposure on a continuous tone printing medium at the output stage. This may be thought of as amplitude modulation. A similar but often more satisfactory effect, producing sharper edges, can be achieved by using high addressability or pulse-width modulation in conjunction with printing processes having an inherently sharp exposure threshold rather than continuous tone response.

Methods to evaluate the prints to determine the reduction of the appearance of the jaggies involve scanning along the edge of a line containing the effects of interest with a long microdensitometer slit whose length covers the space from the middle of the black line to the clear white surround. The resulting reflection profile is proportional to the excursions of the edge. It indicates the additional effects of the printing and measurement processes on decreasing or increasing the jaggies and can be analyzed for its visually significant components against an appropriate CSF.[170] Some of these components are random based on the marking process, others are periodic based on the angle of the line and the resulting frequency of the staircase effect.

Figure 3.44 shows an example of several practical effects of such enhancement and restoration on an italic letter "b." The upper figure is a representation of a conventional bit map of the original computer generated letter. Note the jaggies or staircase on the straight but tilted stroke at the left and a variety of undesirable effects throughout the character. Using the observation window employed by Hewlett Packard's RET (Resolution Enhancement Technology)[55,171,172] as shown on the left, roughly 200 pixel-based templates are compared to the surrounding pixels for each individual pixel in the original "b," a part of which is shown here. A decision is made regarding how large a mark, if any, should replace that pixel, based on a series of rules developed for a particular enhancement scheme, in this case the RET algorithm. The mark in this case is created by modifying the width of the pulse in the horizontal dimension as illustrated. The resulting map of full and width-modulated pixels is shown in the lower part of the letter "b." Note that some of the narrow pixels can be positioned left or right. This is called pulse-width position modulation, PWPM.

When the individual pulses are blurred and developed by the marking process, they will tend to merge into the body of the letter, both physically and visually, to produce even
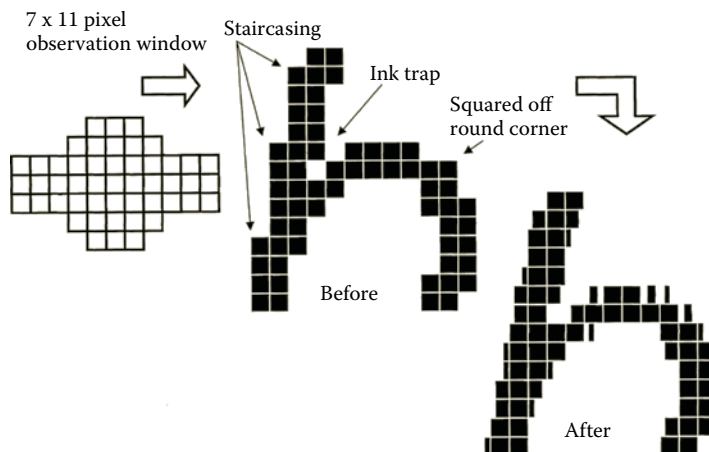
FIGURE 3.44
An example of resolution enhancement on a portion of an italic letter "b." (Adapted from Tong, C., Resolution enhancement in laser printers, *Proceedings of SPIE conferences on color Imaging: Device-Independent color,* Color Hardcopy and Graphic Arts II, San Jose, CA, 1997.

smoother edges than shown as a bit map image here. There are many similar techniques patented prior to and following the above and sold by other companies such as Xerox,[173] IBM,[174] Destiny,[175] and DP-Tek,[178] now owned by Hewlett Packard, to name only a few, each with its own special features. They all create the effect on the HVS equivalent to that provided by a higher resolution print and many enhance the images in other ways as well, such as removing ink traps or sharpening the ends of tapered serifs.[55]

3.7.3 Color Management

Color measurement systems, as discussed earlier in Section 3.2.2.5, are the key to managing color reproduction in any situation. The advances in scanned color imaging systems that separated input and output scanning devices and inserted networks, electronic image archives, monitors and preprinting (prepress) software in between have made it desirable to automate the management of accurate or pleasing (not the same) color reproduction. This in turn has meant automating or at least standardizing and carefully controlling the objective measurement of the color performance of the many input, output, and image manipulation devices and a variety of methods for insuring consistency.[6,14]

The basic concept is to encode, transmit, store, and manipulate images in a device-independent form, carrying along additional information to enable decoding the files at the step just before rendering to an output device, that is, just before making it device specific. CIELAB (i.e., *L*a*b**) based reference color space, above, is commonly used to relate characteristics of both of these types of devices to an objective standard. Standardized operating system software, operating with standardized tools and files accomplished this, but is beyond the scope of this chapter. See References 29 and 30 as well as 6-ch 4, and the papers cited in them for more details and Reference 23 for a practical guide to using the tools that are available at this time.

Today, a common ANSI standard target known as IT-8.7 (see Figure 3.49) is manufactured by Kodak (shown), Fuji, and Agfa, each using their own photographic dyes. It is scanned by the scanner of interest into a file of red, green, and blue (RGB) pixels. It is also measured with a spectrophotometer to determine the CIE *L*a*b** values for all 264 patches. Color management software then compares both results and constructs a *source profile* of the scanner color performance.

A well-known example of a color management system is the approach organized by the International Color Consortium (ICC). See Table 3.9. It has created a standard attempting to serve as a cross-platform device profile format to be used to characterize color devices. It enables the device-independent encoding and decoding primarily developed for the printing and prepress industries, but allows for many solutions providers.

This profile, often called an "ICC profile" if it follows the Consortium's proforma, is a lookup table that is carried with all RGB files made with that scanner. It is useful for correction as long as nothing changes in the scanner performance or setup. Similarly, a destination profile is created, typically for a printer or a monitor. Here, known computer-generated patterns of color patches are displayed or printed, and measured with a spectrophotometer in *L*a*b**. Again, a comparison between the known input and the output is performed by the color management software, which creates a lookup table as a *destination profile*.

The color management architecture incorporates two parts. The first part is the *profiles* as described above. They contain signal processing transforms plus other material and data concerning the transforms and the device. Profiles provide the information necessary to convert device color values to and from values expressed in a color space known as a *profile connection space* (*L*a*b** in the ICC example). The second basic part is the *color management module* (CMM), which does the signal processing of the image data using the profiles.

Progress in color management and the ICC in particular have pulled together an important set of structures and guidelines.[6,14] These enable an open color management architecture that has made major improvements. Of course, gamut differences like those in Figure 3.18, are not a problem that color management, *per se*, can ever solve. It is also important to note that drift in the device characteristics between profile calibrations cannot be removed. It is reported[177] that (averaging over a wide range of colors) rotogravure images in a long run show $\Delta E_{ab}^* = 3.0$ and for offset $\Delta E_{ab}^* = 5.5$, (i.e., the range for 90% of images) while they report for *input scanners* $\Delta E_{ab}^* = 0.4$. They also report that the use of color management and ICC profiles improved system results from $\Delta E_{ab}^* = 9$ down to 5, and suggest in general, with good processes, that this is inherently as good as one can achieve. Similarly, Chung and Kuo[182] found they could achieve an $\Delta E_{ab}^* = 6.5$ as the average for the best scenario in color matching experiments using ICC profiles for a graphic arts application. Control over specific limited sets of system components, colors or small color ranges as well as newer measurement technologies can show much tighter tolerances than these. There is still a great deal of analysis and work that must be carried out to make color management more universal, easier, and more successful.[176,179–181]

3.8 PSYCHOMETRIC MEASUREMENT METHODS USED TO EVALUATE IMAGE QUALITY

3.8.1 Relationships between Psychophysics, Customer Research, and Psychometric Scaling

As one attempts to develop a scanned imaging system, there are usually some image quality questions that cannot be answered by previous experience or by reference to the literature. Often this reduces to a question of determining quantitatively how "something" new *looks visually* for "some task." It is a problem because no one else has ever evaluated the "something" or never used it for "some task" or both. We give the reader at least some pointers to the basic visual scaling discipline and tools to attack his own specialized problems.

As the Image Quality Circle[12,40] and the full framework in Figure 3.3 indicates, there are many places where one needs to quantify the human visual responses. Sometimes this is in the short-cut paths connecting technology variables (the "something") directly to customer quality preferences for "some tasks" through *customer research*. Sometimes, it is in creating a more thorough understanding by developing visual algorithms, which connect the physical image parameters, that is, attributes (other types of "something"), with the fundamental human perceptions of these attributes. The science of developing these latter connections is referred to as *psychophysics*. The underlying discipline for doing both engineering-oriented customer research and psychophysics is psychometric *psychometric scaling*. Hundreds of good technical papers, chapters, and whole books have been written on these subjects, but are often overlooked in imaging science and engineering for a variety of reasons. Many of the papers cited in this chapter draw on the rich resources of psychometric scaling disciplines in certain large corporations, government agencies, and universities to develop their algorithms. Engeldrum[12] has recently distilled many of the basic disciplines and compiled many of the classic references into a useful book and software toolkit for imaging systems development.

3.8.2 Psychometric Methods

There are many classes of psychometric evaluation methods, the selection of which depends on the nature of the imaging variable and the purpose of the evaluation. We can only describe them at a high level in this section. Figure 3.45 describes a framework for considering psychometric experiments, starting with two fundamental purposes, at the left, each of which breaks down into three basic approaches and six types of data.

The way in which the sample preparation is done, observer (called "respondent" in market research) quantity and selection methods, and the numbers of images shown can all be very different, depending on the purpose. In general the customer-user experiments require significantly more care in all areas, are restricted to user-like displays of relatively few images, and require several dozen to hundreds of respondents. They tend to focus on quantifying the "Customer Quality Preference" block in Figure 3.3.

Visual sciences experiments on psychophysics and perception are useful for developing the image quality models and especially the visual algorithms of Figure 3.3 and the comparisons between the HVS and measurements indicated in Figure 3.2. Here smaller numbers of observers, from a few to several dozen, are often deemed adequate. These observers are often experts or technical personnel and can be told to overlook certain defects in samples and concentrate on the visual characteristic of interest. Such observers can be asked to try more fatiguing experiments. These are often broken into several visits to the laboratory, something not possible with customer research.
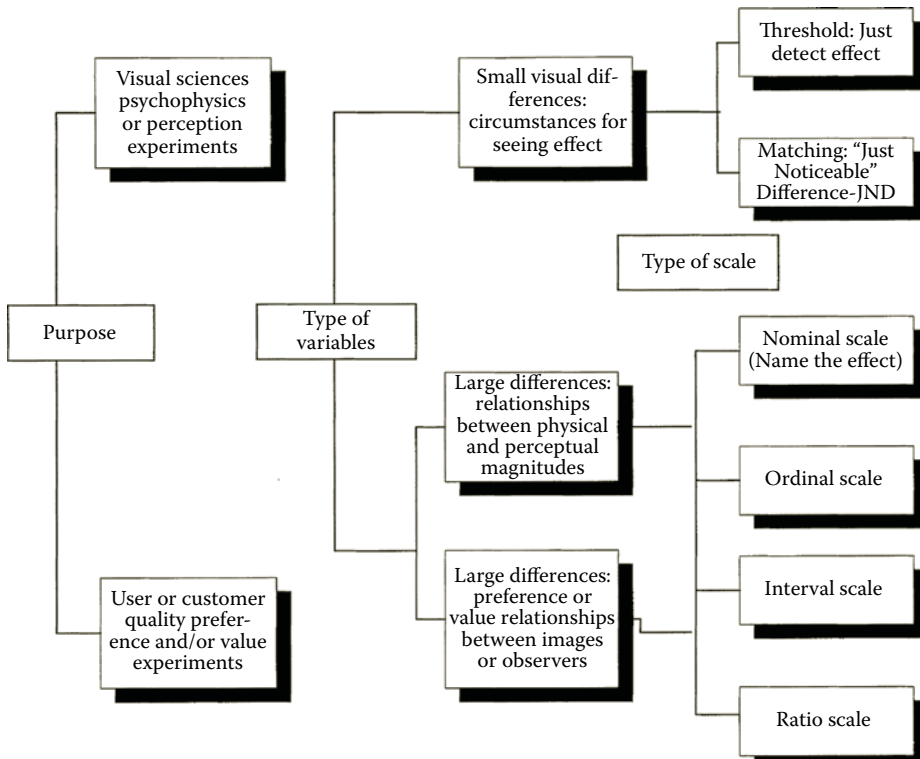
FIGURE 3.45
Psychometric experiments for diverse purposes, grouped in two classes here, can be further classified into three types of variables, which in turn lead to a few basic but significantly different types of scales.

In general, experiments with good statistical design should be used, in which a targeted confidence level is established. It is common practice in many customer and general experiments to seek 95% confidence intervals (any basic statistics book[183] will provide equations and tables to enable this, provided the scaling method is properly classified as shown below). This requires estimating the size of the standard deviation between observers and using it along with the confidence interval equation to determine the number of independent observations that translate to the number of observers. The experimenters in visual sciences can use fewer observers than a customer researcher because the visual sciences use variables and trained observers that have much better agreement, that is, smaller standard deviations. Also, these experimenters may require less statistical confidence because they are often more willing to use other technical judgment factors such as models and inferences from other work.

For either purpose, the decisions regarding basic approach must be determined by looking at the types of variables and the types of scales to be built. If the goal is to determine when some small signal or defect (such as a faint streak) is just visually detected, *threshold* scales are developed. They show the probability of detection compared to the physical attribute(s) of the image samples or the observation variables. One may wish to compare readily visible signals, such as images of well-resolved lines, trying to distinguish when one is just visually darker than another. This involves determining the probability, in *matching* experiments, of what levels of a variable(s) cause two images to be seen as *just noticeably different (JND)*, that is, just do not match each other. Dvorak and Hamerly[184] and Hamerly[185] give examples of JND scaling for text and solid area image qualities.

These experiments often explore fundamental mechanisms of vision and can draw on a relatively small number of observers in well-controlled experimental situations using electronic displays with side-by-side image comparisons. The temptation to substitute an easily controlled electronic display experiment for one in which the imaging media is identical to the actual images of interest (e.g., photographic transparencies viewed by projection, or xerographic prints viewed in office light) must be carefully weighed in each situation. Various forms of image noise, display factors affecting human vision (especially adaptation), as well as visual and psychological reference cues picked up from the surround, are often important enough to outweigh the ease of electronic display methods.

When the magnitude of visible variables is large and the goal is to compare quality attributes over a large range, as in the bottom two "variables" boxes in Figure 3.45, then a decision about the mathematical nature of the desired scale and the general nature and difficulty of the experimental procedure becomes important. The four basic types of scales[186] shown here were developed by Stevens[187–189] and are shown in increasing order of "mathematical power" in Figure 3.45 and with very short descriptions in Table 3.5.

There is an abundance of literature on the theory and application of scaling methods,[12] some of which are indicated in the table as column headings. Additional general references include References 190–199. Below is a very brief summary of the methods to assist the reader in beginning to sort through these choices. Here we assume the samples are "images," but they could just as well be patches of colored chips, displays on a monitor, pages of text, or any other sensory stimulus.

3.8.3 Scaling Techniques

The methods by which the various types of scales (Table 3.5) may be constructed are listed with very brief descriptions and the type of scale they may be used to construct in ( ). Where only complicated procedures enable a type of scale to be derived it is unbolded.

TABLE 3.5

Types of Scales

Type of Scale	Description and Analysis Operations
Nominal	Names of categories/classes
Ordinal	Ordered along variables, determines "greater than" or "less than," (gives arbitrary/unknown distances on the variable scale)
Interval	Ordinal scale + magnitude of differences are quantified. $y = ax + b$ (Equality of intervals may be determined and any linear transformation is OK. Mean, standard deviation, coefficient of correlation are valid. Basis for much image quality analysis)
Ratio	Interval scale where "none" of an attribute is assigned 0 response. $y = ax$ (i.e., $b = 0$) (Interval scale operations are OK and coefficient of variation and equality of ratios are valid—many lightness scales are an example with an absolute zero)

Engeldrum[12] gives a good discussion of all these methods. References using these methods are also listed for each.

3.8.3.1 Identification (Nominal)

In this simple scaling method, observers group images by identifying names for some attributes and collecting images with those attributes. The resulting nominal scales are useful in organizing collections of images into manageable categories.

3.8.3.2 Rank Order (Ordinal)

Observers arrange a set of images according to decreasing or increasing amount of the perceived attribute.[12,188,190,194] A median score for the group is frequently used to select the rank for each sample. Agreement between observers can be tested to understand the nature of the data by calculating the coefficient of concordance or the rank order coefficient.[197]

3.8.3.3 Category (Nominal, Ordinal, Interval)

Observers simply separate the images into various categories of the attribute of interest, often by sorting into labeled piles. This is useful for a large number of images, many of which are fairly close in attributes, so that there are some differences of opinion over observers or over time as to which category is selected. Interval scales can be obtained if the samples can be assumed to be normally distributed on the perceived attributed.[199]

3.8.3.4 Graphical Rating (Interval)

The observers score the magnitude of the image attribute of interest by placing an indicator on a short line scale that has defined endpoints for that attribute. The mean of the positions on the scale for all observers is used to get a score for each image.[12]

3.8.3.5 Paired Comparison (Ordinal, Interval, Ratio)

All images are presented to all observers in all possible pairwise combinations, usually one pair at a time, sometimes with a reference. The observer selects one of the pair as having

more of the attribute of interest. If there are N different images then there are $N(N - 1)/2$ pairs The proportion of observers for which each particular image is selected over each other image is arrayed in a matrix. The average score for each image (i.e., any column in the matrix) is then computed to determine an ordinal scale.[12,191,197,206] If it is assumed that the perceived attributes are normally distributed, then, as with the category method, an interval scale can be determined. This is done using Thurstone's Law of Comparative Judgement[199] in which six types of conditions for standard deviations describing the datasets are used to construct tables of Z-Deviates,[12,197] from which interval scales are directly obtained.

3.8.3.6 Partition Scaling (Interval)

The observer is given two samples, say S1 and S9, and asked to pick a third sample from the set, whose magnitude of the appearance variable under test is halfway between the two samples; call it S5 in this case. Next he finds a sample halfway between S1 and S5, call it S3, then he finds one between S5 and S9, calling it S7, and so on, until he has built a complete interval scale using as many samples and as fine a scale as desired.[4]

3.8.3.7 Magnitude Estimation (Interval, Ratio)

The observer is asked to directly score each sample for the magnitude of the attribute of interest.[12,187,189,194,195] Often, the observer is given a reference image at the beginning of his scoring process, called an anchor, whose attribute of interest is identified with a moderately high, easy to remember score, such as 100. His scores are based on the reference and he is coached in various ways to use values that reflect ratios. This process implies that a zero attribute gets a zero response and hence generates a ratio scale. However actual observations sometimes are more in line with an interval scale, and this needs to be checked after the test.

3.8.3.8 Ratio Estimation (Ratio)

This test may be done by selecting samples that bear specific ratios to a reference image. The experimenter does *not* assign a value to the reference. Alternatively, the observers may be shown two or more specific images at a time and asked to state the apparent ratios between them for the attribute of interest.[4,195]

3.8.3.9 Semantic Differential (Ordinal, Interval)

Typically used for customer research.[190] The image attributes of interest are selected and a set of bipolar adjectives is developed for the attributes. For example, if the attribute class were tone reproduction, the adjectives could be such pairs as darker–lighter, high contrast–low contrast, good shadow detail–poor shadow detail. Each image in the experiment is then rated on a several point scale between each of the pairs. Each scale is treated as an interval scale and the respondents' scores for each image and each adjective pair are averaged. A profile is then displayed.

3.8.3.10 Likert Method (Ordinal)

Typically used for customer research and attitude surveys. A series of statements about the image quality attributes of a set of images is provided (e.g., "The overall tones are perfect in

this images," "the details in the dark parts of this image are very clear"). The respondents are then asked to rate each statement on the basis of the strength of their personal feelings about it: strongly agree (+2); agree (+1); indifferent (0); disagree (–1); strongly disagree (–2). Note signs on numbers reverse for negative statements. The statements used in the survey are often selected from a larger list of customer statements. A previous set of judges maybe was used to determine those statements that produce the greatest agreement in terms of scores assigned to this set of images.

3.8.3.11 Hybrids (Ordinal, Interval, Ratio)

There are many approaches that combine the better features of these different methods to enable handling different experimental constraints and obtaining more accurate or more precise results. A few are noted here:

1. *Paired Comparison for Ratio*: Paired comparisons reduced to an interval scale that is fairly precise and transformed to accommodate a separate ratio technique (accurate but less precise) to set a zero. This gives a highly precise ratio scale.[197]

2. *Paired Comparison Plus Category*: The quality of each paired comparison is evaluated by the observer, using something like a Likert scale below. A seven-level scale from strongly prefer "left" (e.g., +3) to strongly prefer "right" (e.g., –3) is used.[206]

3. *Paired Comparison Plus Distance* using distance (e.g., linear scale on a piece of paper) to rate the magnitude of the difference between each pair, giving the same information as the graphical rating methods discussed earlier, but with the added precision of paired comparison.

4. *Likert and Special Categories*: A variety of nine-point symmetrical (about a center point) word scales can provide categories of preferences that are thought to be of equal intervals. One scale attributed to Bartleson[194] goes from: Least imaginable "…ness" → very little "…ness" → mild "…ness" → moderate "…ness" → average "…ness" → moderate high "…ness" → high "…ness" → very high "…ness" → highest imaginable "…ness." Another similar scale is 1 = Bad, 2 = Poor, 3 = Fair, 4 = Good, 5 = Excellent. Many other such scales are found in the literature.

3.8.4 Practical Experimental Matters Including Statistics

Each of these techniques has been used in many imaging studies, each with special mathematical and procedural variations well beyond the scope of this chapter. A short list of common procedural concerns is given in Table 3.6 (from literature[4,10,14,29] plus a few from the authors' experience).

The * items represent a dozen practical factors that must *always* be considered in designing nearly any major experiment on image quality or attributes of images.

The statistical significance of the results are often overlooked but cannot be stressed enough. For an interval scaling experiment that samples a continuous variable like darkness, standard deviations and means and subsequent confidence intervals on the responses can be calculated in straightforward ways to determine if the appearances of two samples are statistically different or to determine the quality of a curve fit. (See any statistics book on the confidence interval for two means given an estimate of the standard deviations for each sample's score, or to determine the confidence for a regression.)

TABLE 3.6

Factors that should be Considered in Designing Nearly Any Major Experiment on Image Quality or on Attributes of Images

Most important	Important
Complexity of observer task[a]	State of adaptation
Duration of observation sessions[a]	Background conditions
Illumination level[a]	Cognitive factors (many)
Image content[a]	Context
Instructions[a]	Control and history of eye movements
Not leading the observer in preference experiments[a]	Controls
Number of images[a]	Feedback (positive and negative effects)
Number of observers[a]	Illumination color
Observer experience[a]	Illumination geometry
Rewards[a]	Number of observation sessions
Sample mounting/presentation/identification methods[a]	Observer acuity
Statistical significance of results[a]	Observer age
	Observer motivation
	Range effects
	Regression effects
	Repetition rate
	Screening for color vision deficiencies
	Surround conditions
	Unwanted learning during the experiment

[a] See text.

In detection experiments it is often desired to know if two scanned images, which gave two different percentages of observers who saw a defect or an attribute, are significantly different from each other (market researchers call such experiments sampling for attributes). This involves computing confidence intervals for proportions and therefore estimating standard errors for proportions, a procedure less commonly encountered in engineering. If p = the fraction of observers detecting an attribute, q = the fraction not detecting an attribute (note $p + q = 1.0$) and n = number of observers, assuming n is a very small fraction of the population being sampled, then the standard error for proportions is

$$S_p = \left[(p \times q)/n \right]^{0.5} \tag{3.34}$$

and the confidence interval around p is

$$CI = Z \times S_p \tag{3.35}$$

where, for example, $Z = 1.96$ for 95% confidence and 1.28 for 80%. A few cases are illustrated in Table 3.7 to give the reader perspective on the precision of such experiments and the number of observers required. The first column shows the value of p, the fraction of observers finding the attribute of interest. The second column gives the confidence desired in % where 95 is common in many experiments, and 80 is about the lowest confidence cited

TABLE 3.7

Confidence Intervals Around *p* for Attribute Data from Statistics of Proportions

P	% Confidence	*n* = 4	*n* = 8	*n* = 20	*n* = 100	*n* = 500
0.99	*95*	*0.10*	*0.07*	*0.04*	**0.02**	**0.01**
	80	*0.06*	*0.05*	*0.03*	**0.01**	**0.005**
0.95	*95*	*0.21*	*0.15*	**0.10**	**0.04**	**0.02**
	80	*0.13*	*0.09*	**0.06**	**0.03**	**0.01**
0.90	*95*	*0.29*	**0.21**	**0.13**	**0.06**	**0.03**
	80	*0.17*	**0.12**	**0.08**	**0.04**	**0.02**
0.80	*95*	**0.39**	**0.28**	**0.18**	**0.08**	**0.03**
	80	**0.23**	**0.17**	**0.11**	**0.05**	**0.02**
0.60	*95*	**0.48**	**0.34**	**0.22**	**0.10**	**0.04**
	80	**0.31**	**0.21**	**0.13**	**0.06**	**0.02**
0.50	*95*	**0.49**	**0.35**	**0.22**	**0.10**	**0.04**
	80	**0.32**	**0.23**	**0.14**	**0.06**	**0.03**

Statistical uncertainties in experimental results for proportion data (e.g., percentages of "yes" or "no" answers). Table entries gives 80% and 95% confidence as one-sided confidence intervals, that is, positive or negative deviation from the *p* value in column 1, at a few percentages of positive responses "*p*" (row headings) and a few numbers of respondents "*n*" (i.e., sizes of groups interviewed) as column headings. Italic unbolded entries are for *p* values that cannot be realized or closely approximated with the associated *n* values.

in many texts and statistical tables. The numbers reported in the table are the deviations about the fraction in column one that constitute the confidence interval for the population of all observers that would detect the attribute. The unbolded italic numbers correspond to values of *p* that cannot be realized by straightforward means for an observer population as small as indicated (e.g., a "*p*" value of 0.99 could not be observed with only four people—it would take 100!). As an example, for a sample with an attribute that was seen 90% of the time by a sample of 20 observers, one can be 80% confident that 82% to 98% (0.90 ± 0.08) of all observers would see this attribute. One would also be 95% confident that between 77% and 100% (numerically 103%, which here is equivalent to 100%) would see it.

3.9 REFERENCE DATA AND CHARTS

The following pages are a collection of charts, graphs, nomograms, and reference tables, which, along with several earlier ones, the authors find useful in applying first-order analyses to many image quality engineering problems. Needless to say, a small library of computer tools covering the same material would provide a useful package. In addition to those in this section, there are a few graphs, charts, and tables of value to engineering projects included in the text where their tutorial value was considered more important.

These include Figure 3.17 on CIE standard observer color matching function, Figure 3.20 on scan frequency effects, Figure 3.22 on nonuniformity guidelines, Figures 3.30 and 3.31 on MTF, and, finally, Figure 3.37 on edge noise calculations. The tables include Table 3.1 on halftone calculations, Table 3.3, which serves as a directory to Figures 3.50 to 3.53 in this section, and Table 3.7, giving confidence intervals for proportions.

In this section additional graphs on basic colorimetry are provided as Figure 3.46 for a more precise *x*, *y* chromaticity diagram with a dominant wavelength example and some

standard light sources, and Figure 3.47 for spectral characteristic of four standard light sources. Table 3.8 gives useful conversions between imaging oriented variables of density or reflectance and colorimetry units of L* (see also Figure 3.19).

Figures 3.48–3.50 show, through annotations, the important structures in useful industry standard test patterns, two each for monochrome in Figure 3.48 and 3.50 and one for color in Figure 3.49 (reproduced here in black and white). Figure 3.48a is the monochrome ISO 12233 reflection test chart for testing digital still cameras which is also well suited to testing flat platen scanners for detail rendition and resolving power. Figure 3.48b is a monochrome test pattern suited for measuring pictorial monochrome scanners using methods referred to in ISO 16067 (see Table 3.9). Unlike Figure 3.48a which is two-level pattern, it contains a range of gray information to allow mapping tonal response and MTF analysis. Notice that both have many tilted edges and lines to show a wide range of sampling phases as well as untilted ones. Bar pattern and edge patterns are suited to software analysis of spatial frequency response using gray response while resolving power targets can be measured directly and visually with a monitor under magnification.

Next are some useful MTF equations and their corresponding graphs (plotted in log–log form for easy graphical cascading).
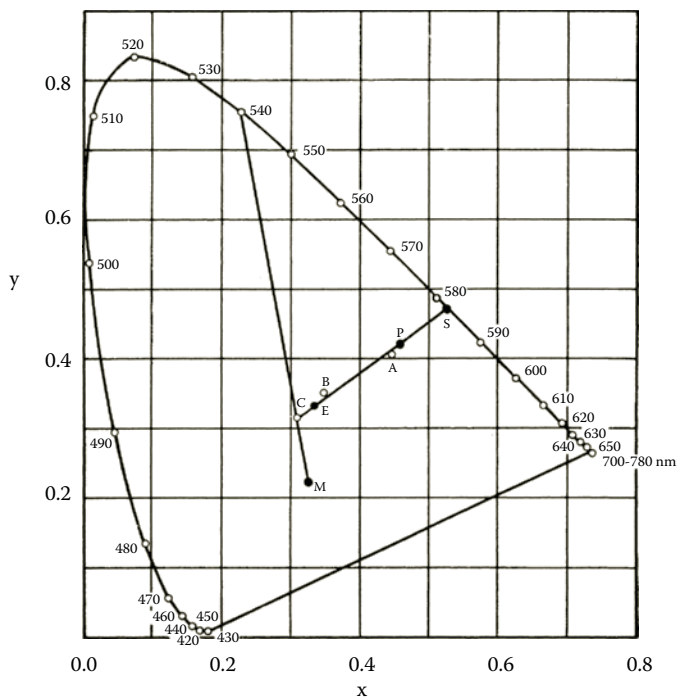
FIGURE 3.46
Dominant wavelength and purity plotted on the CIE *x, y* chromaticity diagram. The dominant wavelength for point P under illuminant C is found by drawing a straight line from the illuminant C point through P to the spectrum locus, where it intersects at 582 nm, the dominant wavelength. Excitation purity is the percentage defined by CP/CS, the percentage the distance from illuminant C to P is of the total distance from illuminant C to spectrum locus. Standard illuminants A, B, and E are also shown. See Figure 3.47 for the relative spectral power distributions of A, B, and C. E has equal amounts of radiation in equal intervals of wavelength throughout the spectrum. (From Hunter, R.S.; Harold, R.W. *The Measurement of Appearance,* 2nd Ed.; John Wiley and Sons: New York, 1987; 191; reproduced with permission of John Wiley & Sons, Inc.).
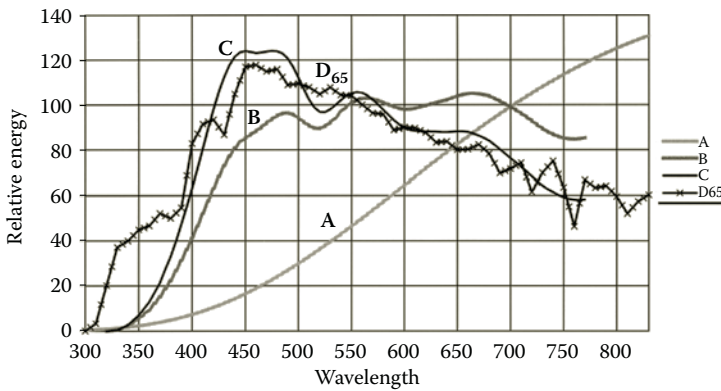
FIGURE 3.47
Standard illuminants A, B, C, and D_{65} showing relative spectral energy distribution. Wavelength is in nm.

TABLE 3.8

An Abbreviated Conversion Table for Density, % Reflectance (%Y/ $Y_m \times$ 100) and L^* Covering Densities to 4

Density	%Y	L^*	Density	Y	L^*
0	100	100	1.4	3.98	23.61
0.05	89.1	95.62	1.5	3.16	20.67
0.1	79.4	91.41	1.6	2.51	17.96
0.15	70.8	87.39	1.7	2	15.49
0.2	63.1	83.49	1.8	1.58	13.11
0.25	56.2	79.73	1.9	1.26	10.99
0.3	50.1	76.13	2	1	8.99
0.4	39.8	69.33	2.2	0.631	5.7
0.5	31.6	63.01	2.4	0.398	3.59
0.6	25.1	57.17	2.6	0.251	2.27
0.7	20	51.84	2.8	0.158	1.43
0.735	18.4	50	3	0.1	0.9
0.8	15.8	46.71	3.2	0.063	0.57
0.9	12.6	42.15	3.4	0.04	0.36
1	10	37.84	3.6	0.025	0.23
1.1	7.94	33.86	3.8	0.016	0.14
1.2	6.31	30.18	4	0.01	0.09
1.3	5.01	26.76			

Figure 3.51 is the MTF of two uniform, sharply bounded spread functions. The MTF of a uniform disc point spread function is defined as

$$T(N) = \frac{2J_1(Z)}{Z} \tag{3.36}$$

where $Z = \pi DN$, N = cycles/mm, and D = diameter of disk in mm.

(a)

Tilted edge patterns   Non-tilted bar patterns   Continuously varying resolving power bar patterns   Tilted bar patterns

(b)

Non-slanted square waves

High contrast resolving power: Landolt C

RIT alphanumeric

Slanted square waves

Neutral tone scale patches

Slanted edges (for MTF)

Near 45° edges

Slanted lines

Bars of 0.1 to 1.0 mm width
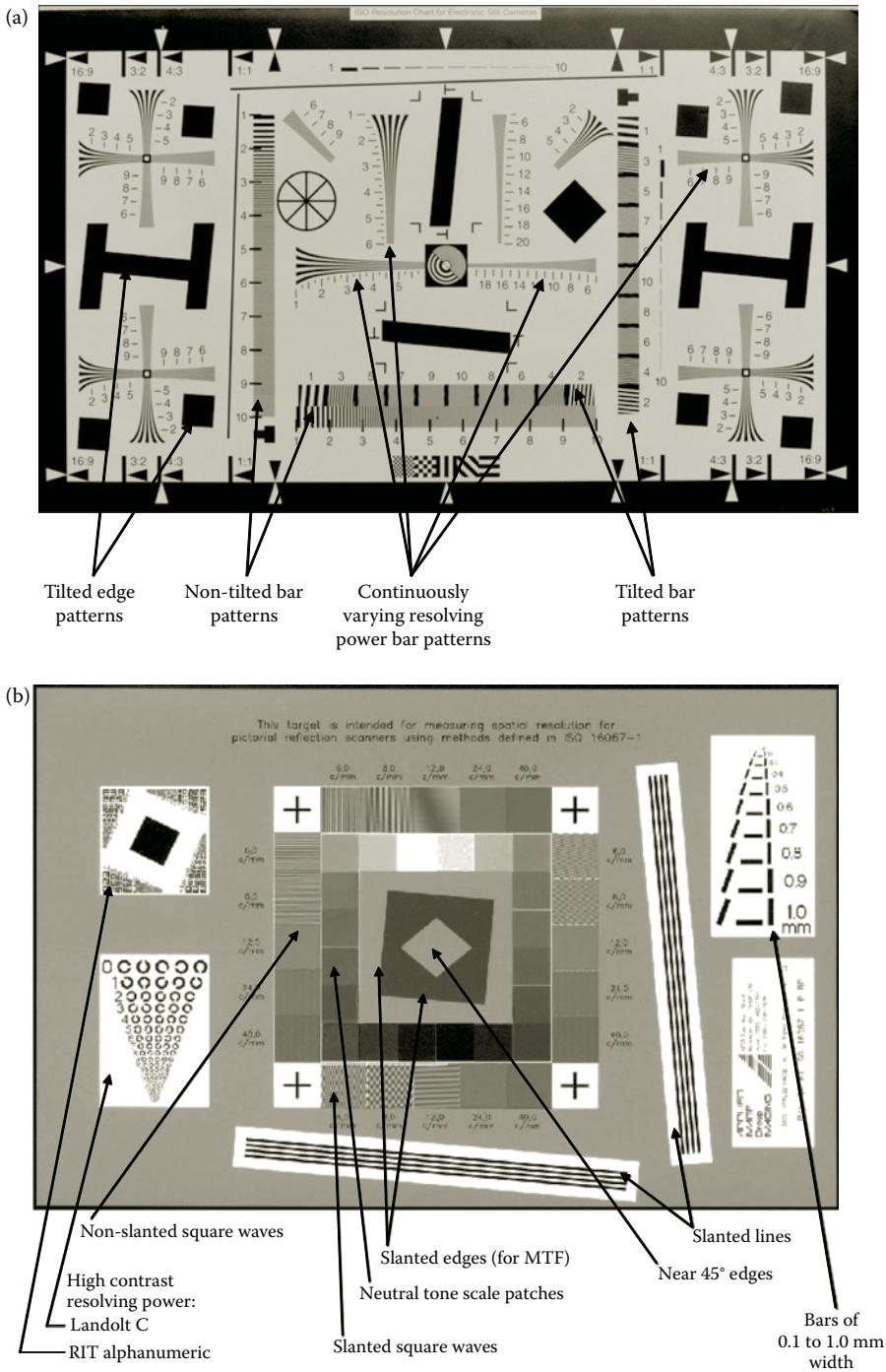
FIGURE 3.48
(a) Digital resolution target specified for digital cameras by ISO12233 which is also useful for testing scanners. (b) Applied Image[106] version of the ISO target for measuring spatial resolution for pictorial reflection scanners using methods defined in ISO16067-1. See Table 3.9 for pointers to standards. (Do not use this printed reproduction for testing, it is considerably degraded.)
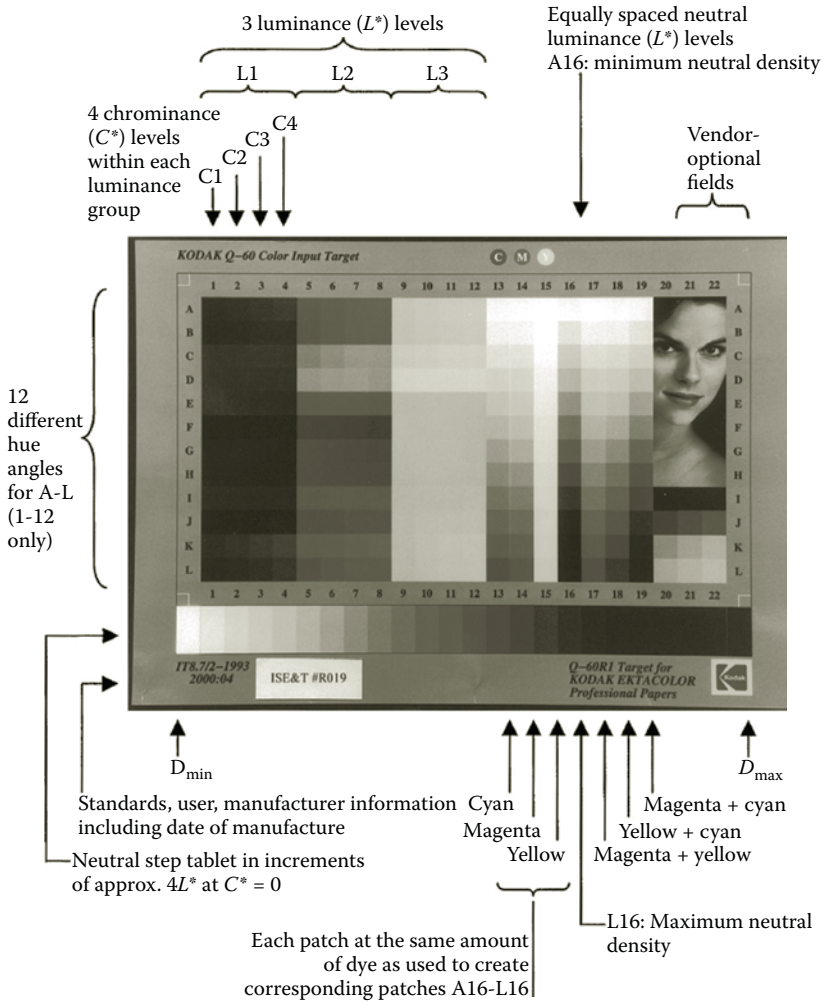
FIGURE 3.49
Layout of the IT8.7/1 (transmissive) and IT8.7/2 (reflective) scanner characterization targets. Details of colors are described in Table 5-1, 5-2, and 5-3 of Reference 23, or ISO IT8.7/1 and 2-1993 (from Reference 23). (*Note:* Do not attempt to use this reproduction as a test pattern.)

The MTF of a uniform slit or uniform image motion is defined as

$$T_{\text{slit}}(N) = \frac{\sin pDN}{pDN} \tag{3.37}$$

where D = width of slit in mm (or width of rectangular aperture or length of motion during image time) and N = cycles/mm.

Figure 3.52 is the MTF of a Gaussian spread function $S(r)$

$$T(N) = e^{-a^2 N^2} \tag{3.38}$$

Scales for geometric characteristics
(Skew, registration, magnification etc.)

Radial resolution targets: Pestrecov
Star, fans labeled in lines/inch, m's

Alternating B&W lines at 150 (3°), 200
(2°) and 100, 150, 200, 300, 400, 600 lpi

ISO character (hexag.) line patterns

Line width tapers and arrays
(20 m to 800 m) Note tilted lines

Halftone Patches at 65, 120 l/in.

NITS resolving power charts:
5 bars/freq., 1.4 - 18 c/mm

Various type styles and sizes
(2 - 12 point) English

Various type styles - non English

Full page diagonal line (pitch errors)

Label - IEEE Std 167A.1 - 1995

Continuous tone grey wedges

48 step density tablet in ~equal
visual steps to density 1.71

15 level step tablet from
density 0.06 to 1.94 and
uniformity test (across page)

High quality continuous tone
photographs

Label - IEEE Std 167A.2 - 1996

FIGURE 3.50

Images of two IEEE Standard Facsimile Test Charts (now "withdrawn" but still in use) which contain many elements valuable in assessing performance of scanning systems; (a) (top) IEEE Std. 167A. 1-1995-Bi-Level (black and white) chart, (b) (lower) IEEE Std. 167A.2-1996, High Contrast (gray scale) chart printed on glossy photographic paper. To identify what test pattern element each annotation refers to, project the relative vertical position of the bar in the specific annotation horizontally across the image of the test pattern. The bars are arranged from left to right in sequence. A composite using many parts of both (a) and (b) plus other elements is available today as the Eastman Kodak/Digital Science Imaging Test Chart (TL. 5003) from Applied Image Corp[106] under keyword Q4.60. See Figure 3.34 for other resolving power targets and Table 3.9 for pointers to other standard test patterns. (*Note:* Do not attempt to use these reproductions as test patterns.)

TABLE 3.9

Standards of Interest in Scanned Imaging

Principal Standards Group	Subgroup	Sampling of Areas of Work or Example Standards	Web site(s) and/or Address(es)
ANSI—American National Standards Institute (those assigned to NPES)	CGATS: Committee for Graphic Arts Technologies Standards	CGATS.4-1993 Graphic Technology—Graphic Arts Reflection Densitometry Measurements—Terminology, Equations, Image Elements and Procedures	http://www.npes.org/standards/cgats.html NPES Assoc. for Suppliers of Printing and Publishing Technologies, 1899 Preston White Dr., Reston, VA 22091
	IT8: Committee on Digital Data Exchange and Color Definition (assigned to CGATS in '94)	IT8.1 Exchange of Color Picture Data IT8.7/2-1993-Reaffirmed 2008 Color Reflection Target for Input Scanner Calibration	http://www.ansi.org American National Standards Inst, 1819 L St. NW #230 Washington, DC
Graphic Communications Association	GRACoL *Regrouped significantly in 2004*	General Requirements for Applications in Commercial Offset Lithography—specifications, best practices *(e.g. SWOP procedures)*	http://www.gracol.org/index.html IDEAlliance, 1421 Prince St. NW #230 Alexandria, VA
ICC—International Color Consortium	NA	ICC Color Management Profile Specification, e.g. ICC. 1:2003-09 ver 4.1.0	http://www.color.org/index.xalter see ANSI-NPES above
ISO/IEC	JTC 1/SC 28 (joint technical committee on office equipment)	ISO 13660-2001 image quality measurement for hard copy output: ISO 12653-2,2000 Test target for black and white scanning	http://www.iso.ch www.iec.ch ISO Secretariat International Organization for Standardization,1,CH.de la Voie-Creuse Case postale 56, CH-1211 Geneva 20, Switzerland IEC Central Office 3, Rue de Varembé, PO Box 131 CH-1211 Geneva 20, Switzerland
	JTC 1/SC 29. (coding of multimedia information)	Coding of audio, picture, multimedia and hypermedia information includes bilevel and limited bits-per-pixel still pictures	

Organization	Subgroup	Description	Contact
	ISO-TC42-WG18 (electronic still picture imaging covers scanners and digital still cameras) see Section 3.4.2	ISO 16067-1&2, 2003 &4: Electronic scanners for photographic images—Spatial Resolution Measurements; ISO 21550-2004 Dynamic range measurement; ISO 20462-3-2005, Psychophysical experimental methods for estimating image quality; ISO 12233 Photography—Electronic still picture cameras—resolution measurements	
INCITS (International Committee for Information Technology Standards); formerly NCITS '97–01 X3 '61–'96	W1 (office equipment	Includes copiers, multifunction, fax machines, page printers, scanners and other office equipment.[90] Collaborates, w JTC1/SC 28—for example on ISO 13660, above.	http://ncits.org/index/htm INCITS Secretariat C/O Information Technology Industry Council, 1250 Eye St. NW #200, Washington, DC 20005
ITU-T (International Telecommunication Union-Telecommunication Standardization Sector)	Joint Photographic Expert Group	ITU-T Rec. T.800 and ISO/IEC 15444, Information Technology—Digital compression and coding of continuous tone still images (JPEG 2000) [see References 93,167,168,180 D. Lee pp. 248–267]	http://www.itu.int/net/home/index.aspx (click on standardization) ITU-T/ International Telecommunications Union (ITU), Place des Nations, 1211 Geneva 20, Switzerland, ATTN. ITU-T
	Joint with ISO/IEC/JTC1 SC29 WG1	JBIG2 ITU-T Rec.T.88	
IEEE	Transmission Systems Committee of IEEE Communications Society	Facsimile imaging, which is of value to many general scanning areas of interest. See Figure 3.50 for example test patterns	http://www.ieee.com The Institute of Electrical and Electronics Engineers, Inc., Corporate Offices 3 Park Ave, 17th Floor New York, NY 10016

N = Frequency in c/mm
D = Diameter of uniform disc or width of a slit (motion)

$$Z = \pi N D$$
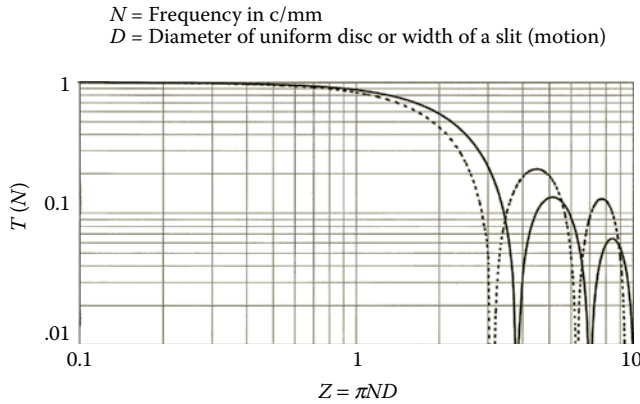
FIGURE 3.51

MTF of a uniform disk (solid line) and slit (dashed line) spread functions where N = frequency in cycles/mm and D = diameter of uniform disk, width of slit or rectangular aperture, or length of motion.

$$aN \text{ c/mm}$$
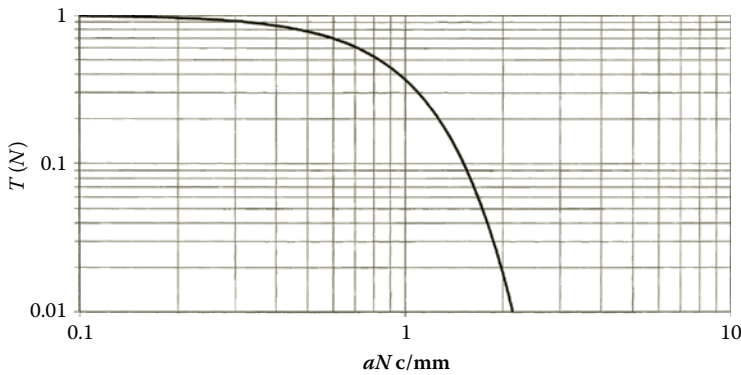
FIGURE 3.52

MTF for imaging system with Gaussian spread function.

where $a = \pi/c$, and c = width of Gaussian spread function $S(r)$ of the form

$$S(r) = 2c^2 e^{-c^2 r^2} = 2c^2 e^{-c^2(x^2+y^2)} \tag{3.39}$$

where r = radius such that $r^2 = x^2 + y^2$; all are in mm^2.

Figure 3.53 is the MTF of a diffraction-limited lens, where

$$T(N) = \frac{2}{p}[\cos^{-1} g - g\sqrt{1 - g^2}\,] \tag{3.40}$$

where $\gamma = N\lambda f$ (object at ∞), N = cycles/mm, λ = wavelength of light in mm, and f = aperture ratio = [focal length]/[aperture diameter].

It has been suggested (Reference 200) that terms in Equation 3.37 raised to the powers of 2, 3, and 4 (where D = the spacing between sensor elements) is useful approximations
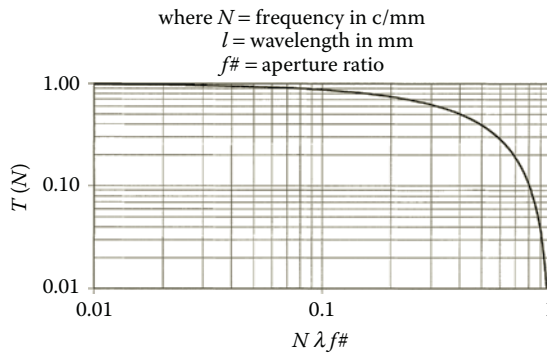
FIGURE 3.53

MTF of a diffraction-limited lens where N = frequency in cycles/mm, λ = wavelength in mm, and $f\#$ = aperture ratio.
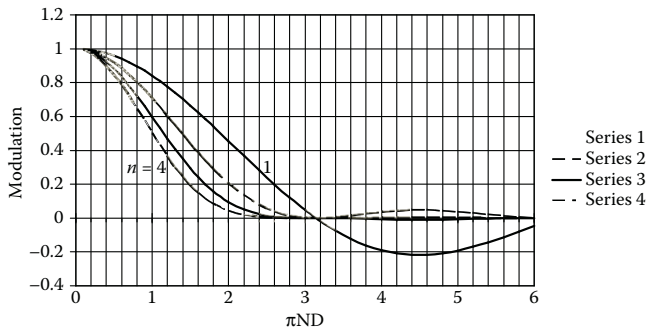
FIGURE 3.54

The general MTF family represented by $[\sin \pi DN/\pi DN]^n$ showing curves for n = 1–4. n = 1 case is shown previously in Figure 3.51 where the terms are explained. N = 3 and 4 approximate many real scanners.

to certain cases of actual scanner MTF performance. These are shown in Figure 3.54 where n = the power of the [sine $\pi DN/\pi DN$] term. In one case averaging over all sampling phases with an ideal sensor (where the sensor width = array spacing, i.e. 100% fill), the n = 2 case was a good approximation. In the case of several real film scanners (Reference 200) where other degradations from optics enter in, the n = 3 case was shown to be a good fit. Finally it appears that some inexpensive flatbed scanners which have even more degradation fit the n = 4 case.

Figure 3.55 presents data on four representative modem films, plotted here to provide perspective on the range of practical photographic characteristics. They are shown here to set scanning performance in perspective. These are not intended to be performance specifications of specific films.

Lastly we finish the reference curves with visual performance relationships. Figure 3.56 illustrates recently developed visual contrast sensitivity curves (related to MTF of linear systems) including color components of vision, after Fairchild,[4] drawn with scales relating to the earlier published visual frequency response characteristics shown in Figure 3.31. Figure 3.57 shows the line luminance visibility threshold as a function of line width, originally described as display "seam visibility" from display experiments after Alphonse and
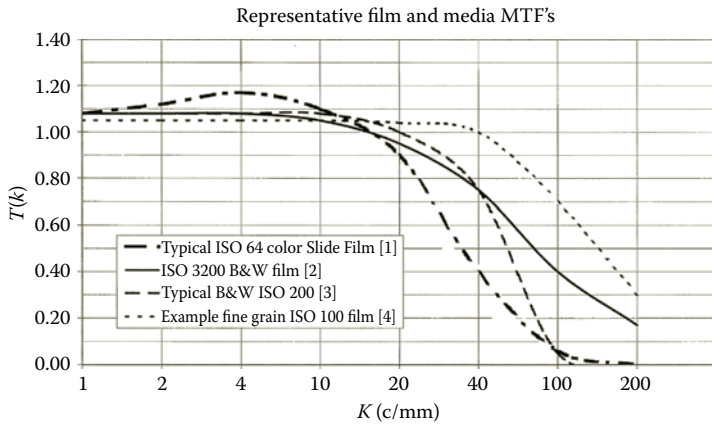
FIGURE 3.55
Data on four representative modern films.
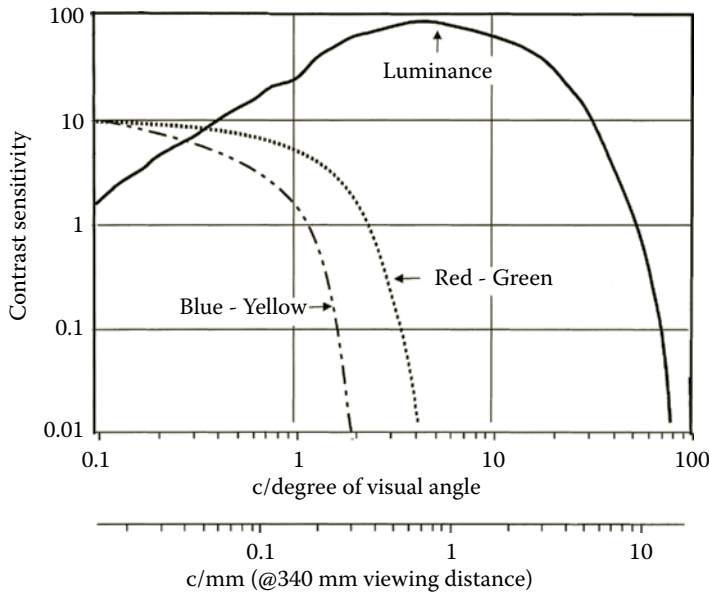
FIGURE 3.56
Typical visual spatial contrast sensitivity functions for luminance and indicated chromatic contrasts at constant luminance. (Adapted from Fairchild, M.D. *Color Appearance Models;* Addison-Wesley: Reading, MA, 1998.)

Lubin.[125] Figure 3.58 shows the edge contrast threshold visibility from display experiments of Lubin and Pica.[125]

The closing reference, Table 3.9, is a chart showing a *very* sparse cross section of the standards that intercept the digital and scanning image quality technical world. These enable an engineer to get some orientation and pointers to important standards organizations.
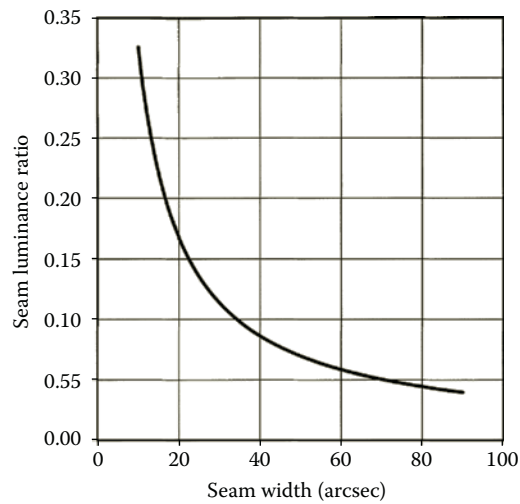
FIGURE 3.57

Line luminance visibility threshold as a function of line width for a black or white line on the opposite background derived from seam visibility for CRT displays. (Adapted from Lubin, J. The use of psychophysical data and models in the analysis of display system performance. In *Digital Images and Human Vision;* Watson, A.B., Ed.; MIT Press: Cambridge, MA, 1993; 163–178.)
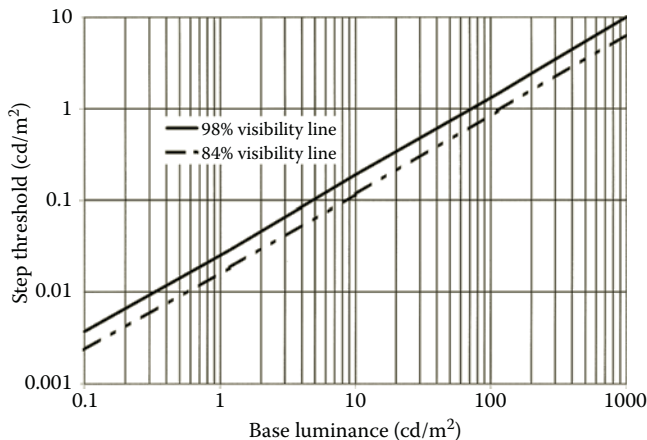
FIGURE 3.58

Thresholds for the visibility of a luminance difference at a step edge in a 17 by 5.25° CRT display where 84% detection = $dL_{85} = 0.01667L^{0.8502}$. (Adapted from Lubin, J. The use of psychophysical data and models in the analysis of display system performance. In *Digital Images and Human Vision;* Watson, A.B., Ed.; MIT Press: Cambridge, MA, 1993; 163–178.)

ACKNOWLEDGMENTS

scanning. His outstanding contributions to Xerox research and the careers and ideas of many he worked with at Xerox, as a consultant and in the general technical community, are widely regarded with the highest esteem. The previous edition of this chapter would not have been completed without his efforts, as he continued to help in its editing, even during his last days. We all miss his learned advice, special humor, and profound insights.

Contributions   We are indebted to Cherie Wright, John Moore, David Lieberman, Roger Triplett, and others on the staff of the Imaging Sciences Engineering and Technology Center of the Strategic Programs Development Unit at Xerox Corporation for their helpful participation and support in the preparation of enduring parts of the earlier version of this chapter, and to Xerox Corporation for the use of their resources. I'd also like to thank my colleagues in the Imaging and Photographic Technology Department[201] at RIT for their support as I reworked the current edition. All illustrations, except as noted, are original drawings created for purposes of this chapter. However, those inspired by another author's way of illustrating a complex topic or providing a collection of useful data, reference his or her contribution with the note "(From Reference___)." We are grateful for these authors' ideas or data, which we could build on here. We also wish to thank our reviewers Martin Banton, Guarav Sharma, Robert Loce, and Keith Knox for their time and many valuable suggestions, and our wives Jane Lehmbeck and Mary Urbach for their support and encouragement during work on the earlier edition and Jane during the current revision.

REFERENCES

1. Mertz, P.; Gray, F.A. A theory of scanning and its relation to the characteristics of the transmitted signal in telephotography and television. *Bell System Tech. J.* 1934, *13*, 464–515.
2. Hornack, J.P. *Visual Encyclopedia of Imaging Science and Technology;* J. Wiley: New York, 2002—In depth treatment of many topics including section on "Imaging Systems."
3. Dougherty, E.R. *Digital Image Processing Methods;* Marcel Dekker: New York, 1994.
4. Fairchild, M.D. *Color Appearance Models;* Addison-Wesley: Reading, MA, 1998.
5. Eschbach, R.; Braun, K. Eds. *Recent Progress in Color Science;* Society for Imaging Science & Technology: Springfield, VA, 1997.
6. Sharma, G. Ed. *Digital Color Imaging Handbook;* CRC Press: Boca Raton, FL, 2003.
7. Eschbach, R. Ed. *Recent Progress in Digital Halftoning I and II;* Society for Imaging Science & Technology: Springfield, VA, 1994, 1999.
8. Kang, H. *Color Technology for Electronic Imaging Devices;* SPIE Press: Bellingham, WA, 1997.
9. Dougherty, E.R. Ed. *Electronic Imaging Technology;* SPIE Press: Bellingham, WA, 1999.
10. MacAdam, D.L. Ed. *Selected Papers on Colorimetry—Fundamentals, MS 77;* SPIE Press: Bellingham, WA, 1993.
11. Pennebaker, W.; Mitchell, J. *JPEG Still Image Data Compression Standard;* Van Nostrand Reinhold: New York, 1993.
12. Engeldrum, P.G. *Psychometric Scaling: A Toolkit for Imaging Systems Development;* Imcotek Press: Winchester, MA, 2000.
13. Vollmerhausen, R.H.; Driggers, R.G. *Analysis of Sampled Imaging Systems Vol. TT39;* SPIE Press: Bellingham, WA, 2000.
14. Giorgianni, E.J.; Madden, T.E. *Digital Color Management Encoding Solutions;* Addison-Wesley: Reading, MA, 1998.
15. Watson, A.B. Ed. *Digital Images & Human Vision;* MIT Press: Cambridge, MA, 1993.

16. Sharma, G. *Digital Color Imaging Handbook;* CRC Press, Boca Raton, FL, 2003. An excellent in depth review of many topics including: fundamentals, psychophysics, color management, digital color halftones, compression and camera image processing and more.

17. Russ, J.C. *The Image Processing Handbook,* 5th Ed.; Taylor and Francis-CRC: Boca Raton, FL, 2007—In depth discussion of numerous image processing topics.

18. Graham, R. *The Digital Image;* CRC Press-Whittles Publishing: Boca Raton, FL 2005—Excellent tutorial on fundamentals of digital imaging and especially photography.

19. Cost, F. *Pocket Guide to Digital Printing;* Delmar Publishers: Albany, NY, 1997.

20. Ohta, N., Rosen, M. *Color Desktop Printing Technology;* Taylor & Francis-CRC Div: Boca Raton, FL, 2006—Comprehensive overview with useful detail.

21. Gann, R.G. Desktop Scanners Image quality Evaluation; (Prentice Hall PTR, Upper Saddle River, NJ, 1999) overall practical serious-user oriented with especially useful practical tests.

22. Matteson, R. Scanning for the SOHO Small Office and Home Office, Virtualbookworm.com Publishing PO Box 9949, College Station TX 2004—Excellent very basic tutorial on all aspects of scanning.

23. Adams, R.M.; Weisberg, J.B. *The GATF Practical Guide to Color Management;* GATF Press: Pittsburgh, PA, 2000.

24. Sharma, G.; Wang, S.; Sidavanahalli, D.; Knox, K. "The impact of UCR on scanner calibration." Proceedings of IS&T Image Processing, Image Quality, Image Capture, Systems Conference. Portland, OR, 1998; 121–124.

25. Knox, K.T. "Integrating cavity effect in scanners." Proceedings of IS&T/OSA Optics and Imaging in the Information Age, Rochester, NY, 1996; 156–158.

26. Sharma, G.; Knox, K.T. "Influence of resolution on scanner noise perceptibility." Proceedings of IS&T 54th Annual and Image Processing, Image Quality, Image Capture, Systems Conference, Montreal, Quebec, Canada, 2001; 137–141.

27. Loce, R.; Roetling, P.; Lin, Y. Digital halftoning for display and printing of electronic images. In *Electronic Imaging Technology;* Dougherty, E.R., Ed.; SPIE Press: Bellingham, WA, 1999.

28. Lieberman, D.J.; Allebach, J.P. "On the relation between DBS and void and cluster." Proceedings of IS&T's NIP 14: International Conference on Digital Printing Technologies, Toronto, Ontario, Canada, 1998; 290–293.

29. Sharma, G.; Trussell, H.J. Digital color imaging. *IEEE Trans. on Image Proc.* 1997, *6,* 901–932.

30. Sharma, G.; Vrhel, M.; Trussell, H.J. Color imaging for multimedia. *Proc. IEEE* 1998, *86,* 1088–12108.

31. Jin, E.W.; Feng, X.F.; Newell, J. "The development of a color visual difference model (CVDM)." Proceedings of IS&T Image Processing, Image Quality, Image Capture, Systems Conference, Portland, OR, 1998; 154–158.

32. Sharma, G.; Trussell, H.J. Figures of merit for color scanners. *IEEE Trans. on Image Proc*. 1997, *6,* 990–1001.

33. Shaw, R. "Quantum efficiency considerations in the comparison of analog and digital photography." Proceedings of IS&T Image Processing, Image Quality, Image Capture, Systems Conference, Portland, OR, 1998; 165–168.

34. Loce, R.; Lama, W.; Maltz, M. Vibration/banding. In *Electronic Imaging Technology;* Dougherty, E.R. Ed.; SPIE Press: Bellingham, WA, 1999.

35. Dalal, E.N.; Rasmussen, D.R.; Nakaya, F.; Crean, P.; Sato, M. "Evaluating the overall image quality of hardcopy output." Proceedings of IS&T Image Processing, Image Quality, Image Capture, Systems Conference, Portland, OR, 1998; 169–173.

36. Rasmussen, D.R.; Crean, P.; Nakaya, F.; Sato, M.; Dalai, E.N. "Image quality metrics: Applications and requirements." Proceedings of IS&T Image Processing, Image Quality, Image Capture, Systems Conference, Portland, OR, 1998; 174–178.

37. Loce, R.; Dougherty, E. Enhancement of digital documents. In *Electronic Imaging Technology;* Dougherty, E.R., Ed.; SPIE Press: Bellingham, WA, 1999.

38. Lieberman, D.J.; Allebach, J.P. "Image sharpening with reduced sensitivity to noise: A perceptually based approach." Proceedings of IS&T's NIP 14: International Conference on Digital Printing Technologies, Toronto, Ontario, Canada, 1998; 294–297.

39. Keelan, B.W. *Handbook of Image Quality Characterization and Prediction;* Marcell Dekker: New York, 2002—Comprehensive and detailed.

40. Engeldrum, P.G. "A new approach to image quality." Proceedings of the 42nd Annual Meeting of IS&T, 1989; 461–464.

41. Engeldrum, P.G. A framework for image quality models. *Imaging Sci. Technol.* 1995, *39,* 312–323.

42. Engeldrum, P.G. *Psychometric Scaling: A Toolkit for Imaging Systems Development;* IMCOTEK Press: Winchester, MA, 2000; chapter 2, 5–17.

43. Shannon, C.E. A mathematical theory of communication. *Bell System Tech. J.* 1948, *27,* 379, 623.

44. Roetling, "P.G. Visual performance and image coding." Proceedings of the Society of Photo-Optical Instrumentation Engineers on Image Processing, Vol. 74, 1976; 195–199.

45. Roetling, P.G.; Loce, R.P. Digital halftoning. In *Digital Image Processing Methods;* Dougherty, E.R., Ed.; Marcel Dekker: New York, 1994; 363–413.

46. Eyer, J.A. The influence of emulsion granularity on quantitative photographic radiometry. *Photog. Sci. Eng.* 1962, *6,* 71–74.

47. Dainty, J.C.; Shaw, R. *Image Science: Principles, Analysis and Evaluation of Photographic-Type Imaging Processes;* Academic Press: New York, 1974.

48. Selwyn, E.W.H. A theory of graininess. *Photog. J.* 1935, *75,* 571–589.

49. Siedentopf, H. Concerning granularity, resolution, and the enlargement of photographic negatives. *Physik Zeit.* 1937, *38,* 454.

50. Altman, J.H.; Zweig, H.J. Effect of spread function on the storage of information on photographic emulsions. *Photog. Sci. Eng.* 1963, *7,* 173–177.

51. Lehmbeck, D.R. Experimental study of the information storing properties of extended range film. *Photog. Sci. Eng.* 1967, *11,* 270–278.

52. Vaysman, A.; Fairchild, M.D. "Degree of quantization and spatial addressability tradeoffs in the perceived quality of color images." *Proc SPIE* on Color Imaging III 1998, *3300,* 250.

53. Marshall, G. *Handbook of Optical and Laser Scanning,* chapter 3; Marcell Dekker, NY, 2004 (previous edition this book & chapter).

54. Bryngdahl, O.J. *Opt. Soc. Am.* 1976, *66,* 87–98.

55. Loce, R.P.; Dougherty, E.R. *Enhancement and Restoration of Digital Documents;* SPIE Optical Engineering Press: Bellingham, WA, 1997.

56. Jorgensen, G.W. Preferred tone reproduction for black and white halftones. In *Advances in Printing Science and Technology;* Banks, W.H., Ed.; Pentech Press: London, 1977; 109–142.

57. Jones, L.A.; Nelson, C.N. The control of photographic printing by measured characteristics of the negative. *J. Opt. Soc. Am.* 1942, *32,* 558–619.

58. Jones, L.A. Recent developments in the theory and practice of tone reproduction. *Photogr. J.* Sect. B 1949, *89B,* 126–151.

59. Bartleson, C.J.; Breneman, E.J. Brightness perception in complex fields. *J. Opt. Soc. Am.* 1967, *57,* 953–957.

60. Nelson, C.N. Tone reproduction. In *The Theory of Photographic Process,* 4th Ed.; James, T.H., Ed.; Macmillan: New York, 1977; 536–560.

61. Nelson, C.N. The reproduction of tone. In *Neblette's Handbook of Photography and Reprography: Materials, Processes and Systems,* 7th Ed.; Sturge, J.M., Ed.; Van Nostrand Reinhold: New York, 1977; 234–246.

62. Holladay, T.M. An optimum algorithm for halftone generation for displays and hard copies. *Proceedings of the SID* 1980, *21,* 185–192.

63. Roetling, P.G.; Loce, R.P. Digital halftoning. In *Digital Image Processing Methods;* Dougherty, E.R., Ed.; Marcel Dekker: New York, 1994; 392–395.

64. Roetling, P.G. Analysis of detail and spurious signals in halftone images. *J. Appl. Phot. Eng.* 1977, *3,* 12–17.

65. ISO-TC-42, ISO 14524-1999 and 12232:2006(E), International Standards Organization, Geneva, Switzerland, 2006, See Table 10, this chapter. OECF stands for Opto-electronic Conversion Function—As applied to cameras in ISO 14524 which is conceptually the same for scanners. 12232 deals with speed metrics derived from OECF's.

66. Stoffel, J.C. *Graphical and Binary Image Processing and Applications;* Artech House: Norwood, MA, 1982; 285–350.

67. Stoffel, J.C.; Moreland, J.F. A survey of electronic techniques for pictorial image reproduction. *IEEE Trans. Comm.* 1981, *29*, 1898–1925.

68. Ulichney, R. *Digital Halftoning;* The MIT Press: Cambridge, MA, 1987.

69. Clapper, R.; Yule, J.A.C The effect of multiple internal reflections on the densities of halftone prints on paper. *J. Opt. Soc. Am.*, *43*, 600–603, 1953, as explained in Yule, J.A.C. *Principles of Color Reproduction;* John Wiley and Sons: New York, 1967; 214.

70. Yule, J.A.C.; Nielson, W.J. The penetration of light into paper and its effect on halftone reproduction. In *Research Laboratories Communication No. 416;* Kodak Research Laboratories: Rochester, NY, 1951 and in TAGA Proceedings, 1951, *3,* 65–76.

71. Lehmbeck, D.R. "Light scattering model for predicting density relationships in reflection images." Proceedings of 28th Annual Conference of SPSE, Denver, CO, 1975; 155–156.

72. Maltz, M. Light-scattering in xerographic images. *J. Appl. Phot. Eng.* 1983, *9*, 83–89.

73. Kofender, J.L. "The Optical Spread Functions and Noise Characteristics of Selected Paper Substrates Measured in Typical Reflection Optical System Configurations," MS thesis, Rochester Institute of Technology: Rochester, NY, 1987.

74. Klees, K.J.; Holmes, J. "Subjective evaluation of noise filters applied to bi-level images." 25th Fall Symposia of Imaging (papers in summary form only). Springfield, VA, *Soc. Phot. Sci. & Eng.,* 1985.

75. Hunt, R.W.G. *Reproduction of Colour in Photography, Printing & Television,* 5th Ed.; The Fountain Press: Tolworth, England, 1995.

76. Hunt, R.W.G. *Measuring Colour;* Ellis Horwood Limited, Halstead Press, John Wiley & Sons: NY, 1987.

77. Miller, M.; Segur, R. "Perceived IQ and acceptability of photographic prints originating from different resolution digital capture devices." Proceedings of IS&T Image Processing, Image Quality, Image Capture Systems (PICS) Conference, Savannah, GA, 1999; 131–137.

78. Smith, W.J. *Modern Optical Engineering;* McGraw Hill: New York, 1966; 308–324.

79. Bestenreiner, F.; Greis, U.; Helmberger, J.; Stadler, K. Visibility and correction of periodic interference structures in line-by-line recorded images. *J. Appl. Phot. Eng.* 1976, *2*, 86–92.

80. Sonnenberg, H. Laser-scanning parameters and latitudes in laser xerography. *Appl. Opt.* 1982, *21*, 1745–1751.

81. Firth, R.R.; Kessler, D.; Muka, E.; Naor, K.; Owens, J.C. A continuous-tone laser color printer. *J. Imaging Technol.* 1988, *14*, 78–89.

82. Goodman, N.B. "Perception of spatial color variation caused by mass variations about single separations." Proceedings of IS&T's NIP14: International Conference on Digital Printing Technologies, Toronto, Ontario, Canada, 1998; 556–559.

83. Shade, O. Image reproduction by a line raster process. In *Perception of Displayed Information;* Biberman, L.M., Ed.; Plenum Press: New York, 1976; 233–277.

84. Shade, O. Image gradation, graininess and sharpness in TV and motion picture systems. *J. SMPTE* 1953, 67, 97–164.

85. Biberman, L.M. Ed. *Perception of Displayed Information;* Plenum Press: New York, 1976.

86. Lehmbeck, D.R.; Urbach, J.C. "Scanned Image Quality," Xerox Internal Report X8800370; Xerox Corporation: Webster, NY, 1988.

87. Kipman, Y. Imagexpert Home Page, http://www.imagexpert.com; Nashua NH, 2003 (describes several scanning-based image quality tools).

88. Wolin, D.; Johnson, K.; Kipman, Y. "Importance of objective analysis in IQ evaluation." IS&T's NIP14: International Conference on Digital Print Technologies, Toronto, Ontario, Canada, 1998; 603.

89. Briggs, J.C; Tse, M.K. "Beyond density and color: Print quality measurement using a new hand-held instrument." Proceedings of ICIS 02: International Congress of Imaging Science, Tokyo, Japan, May 13–17, 2002, and describes other scanning-based image quality tools at QEA Inc., http://www.qea.com (accessed 2003).

90. Yuasa, M.; Spencer, P. NCITS-W1: "Developing standards for copiers and printers." Proceedings of IS&T Image Processing, Image Quality, Image Capture Systems (PICS) Conference, Savannah GA, 1999; 270.

91. Williams, D. "Debunking of specsmanship: Progress on ISO/TC42 standards for digital capture imaging performance." Proceedings of IS&T Processing Images, Image Quality, Capturing Images Systems Conference (PICS), Rochester, NY, 2003; 77–81.

92. Williams, D. "Benchmarking of the ISO 12233 slanted edge spatial frequency response plug-in." Proceedings of IS&T Image Processing, Image Quality, Image Capture Systems (PICS) Conference, Portland, OR, 1998; 133–136.

93. Hubel, P.M. "Color IQ in digital cameras." Proceedings of IS&T Image Processing, Image Quality, Image Capture Systems (PICS) Conference, 1999; 153.

94. Swing, R.E. *Selected Papers on Microdensitometry;* SPIE Optical Eng Press: Bellingham, WA, 1995.

95. Swing, R.E. *An Introduction to Microdensitometry;* SPIE Optical Eng Press: Bellingham, WA, 1997.

96. Lehmbeck, D.R.; Jakubowski, J.J. Optical-principles and practical considerations for reflection microdensitometry. *J. Appl. Phot. Eng.* 1979, *5*, 63–77.

97. Ptucha, R. "IQ assessment of digital scanners and electronic still cameras." Proceedings of IS&T Image Processing, Image Quality, Image Capture Systems (PICS) Conference, Savannah, GA, 1999; 125.

98. Gonzalez, R.C.; Wintz, P. *Digital Image Processing;* Addison, Wesley: Reading, MA, 1977; 36–114.

99. Jakubowski, J.J. Methodology for quantifying flare in a microdensitometer. *Opt. Eng.* 1980, *19*, 122–131.

100. Knox, K.T. "Integrating cavity effect in scanners." Proceedings of IS&T/OSA Optics and Imaging in the Information Age, Rochester, NY, 1996; 156–158.

101. Knox, K.T. US Patent #5,790,281, August 4, 1998.

102. Perrin, F.H. Methods of appraising photographic systems. *J. SMPTE* 1960, *69*, 151–156, 239–249.

103. Shade, O. *Image Quality, a Comparison of Photographic and Television Systems;* RCA Laboratories: Princeton, NJ, 1975.

104. Newell, J.T.; Triplett, R.L. An MTF analysis metric for digital scanners. *Proceedings of IS&T 47th Annual Conference/ICPS, Rochester, NY,* 1994; 451–455.

105. Lamberts, R.L. The prediction and use of variable transmittance sinusoidal test objects. *Appl. Opt.* 1963, 2, 273–276.

106. Applied Image, on line catalog, (Applied Image Inc., 1653 E. Main Street, Rochester NY USA, 2009, http://www.aig-imaging.com/. Nearly all test patterns referred to in this chapter and various standards are available from this source along with detailed descriptions in their on-line catalog. See also Reference 200.

107. Scott, F.; Scott, R.M.; Shack, R.V. The use of edge gradients in determining modulation transfer functions. *Photog. Sci. Eng.* 1963, *7*, 345–356.

108. Campbell, F.W. *Proc. Australian Physiol. Soc.* 1979, *10*, 1.

109. Gorog, I.; Carlson, C.R.; Cohen, R.W. "Luminance perception—Some new results." In Proceedings, SPSE Conference on Image Analysis and Evaluation; Shaw, R., Ed.; Toronto, Ontario, Canada, 1976; 382–388.

110. Bryngdahl, O. Characteristics of the visual system: Psychophysical measurements of the response to spatial sine-wave stimuli in the photopic region. *J. Opt. Soc. Am.* 1966, *56*, 811–821.

111. Watanabe, H.A.; Mori, T.; Nagata, S.; Hiwatoshi, K. *Vision Res.* 1968, *8*, 1245–1254.

112. Glenn, W.E.; Glenn, G.; Bastian, C.J. "Imaging system design based on psychophysical data." In Proceedings of the SID 1985, *26*, 71–78.

113. Dooley, R.P.; Shaw, R. A statistical model of image noise perception. In *Image Science Mathematics Symposium;* Wilde, C. O., Barrett, E., Eds.; Western Periodicals: Hollywood, CA, 1977; 10–14.

114. Patterson, M. In Proceedings of the SID 1986, *27*, 4.

115. Blakemore, C.; Campbell, F.W. *J. Physio.* 1969, *203*, 237–260.

116. Rogowitz, B.E. *Proceedings of the SID* 1983, *24*, 235–252.

117. Hufnagel, R. In *Perception of Displayed Information;* Biberman, L., Ed.; Plenum Press: New York, 1973; 48.

118. Oppenheim, A.V.; Schafer, R. *Digital Signal Processing;* Prentice-Hall: Englewood Cliffs, NJ, 1975; 413–418.

119. Jones, R.C. New method of describing and measuring the granularity of photographic materials. *J. Opt. Soc. Am.* 1955, *45,* 799–808.

120. Lehmbeck, D.R. *Imaging Performance Measurement Methods for Scanners that Generate Binary Output.* 43rd Annual Conference of SPSE, Rochester, NY, 1990; 202–203.

121. Vollmerhausen, R.H.; Driggers, R.G. *Analysis of Sampled Imaging Systems Vol. TT39;* SPIE Press: Bellingham, WA, 2000; 50–72.

122. Kriss, M. Image structure. In *The Theory of Photographic Process,* 4th Ed.; James, T.H., Ed.; Plenum Press: New York, 1977; Chap. 21, 592–635.

123. Carlson, C.R.; Cohen, R.W. A simple psychophysical model for predicting the visibility of displayed information. *Proc. of SID* 1980, *21,* 229–246.

124. Barten, P.G.J. "The square root integral (SQRI): A new metric to describe the effect of various display parameters on perceived image quality." Proceedings of SPIE conference on Human Vision, Visual Processing, and Digital Display, Los Angeles, CA, 1989; Vol. 1077, 73–82.

125. Lubin, J. The use of psychophysical data and models in the analysis of display system performance. In *Digital Images and Human Vision;* Watson, A.B., Ed.; MIT Press: Cambridge, MA, 1993; 163–178.

126. Daly, S. The visible differences predictor: an algorithm for the assessment of image fidelity. In *Digital Images and Human Vision;* Watson, A.B., Ed.; MIT Press: Cambridge, MA, 1993; 179–206.

127. Scott, F. Three-bar target modulation detectability. *J. Photog. Sci. Eng.* 1966, *10,* 49–52.

128. Charman, W.N.; Olin, A. Image quality criteria for aerial camera systems. J. *Photogr. Sci. Eng.* 1965, *9,* 385–397.

129. Burroughs, H.C.; Fallis, R.F.; Warnock, T.H.; Brit, J.H. *Quantitative Determination of Image Quality,* Boeing Corporation Report D2: 114058–1, 1967.

130. Snyder, H.L. "Display image quality and the eye of the beholder." Proceedings of SPSE Conference on Image Analysis and Evaluation, Shaw, R., Ed.; Toronto, Ontario, Canada, 1976; 341–352.

131. Barten, P.G.J. The SQRI method: A new method for the evaluation of visible resolution on a display. *Proc. SID* 1987, *28,* 253–262.

132. Barten, P.G.J. "Physical model for the contrast sensitivity of the human eye." Proceedings of the SPIE on Human Vision, Visual Processing, and Digital Display III, San Jose, CA, 1992; Vol. 1666, 57–72.

133. Daly, S. "The visible differences predictor: an algorithm for the assessment of image fidelity." Proceedings of the SPIE on Human Vision, Visual Processing, and Digital Display III, San Jose, CA, 1992; Vol. 1666, 2–15.

134. Frieser, H.; Biederman, K. Experiments on image quality in relation to modulation transfer function and graininess of photographs. *J. Phot. Sci. Eng.* 1963, *7,* 28–46.

135. Biederman, K. *J. Photog. Korresp.* 1967, *103,* 41–49.

136. Granger, E.M.; Cupery, K.N. An optical merit function (SQF) which correlates with subjective image judgements. *J. Phot. Sci. Eng.* 1972, *16,* 221–230.

137. Kriss, M.; O'Toole, J.; Kinard, J. "Information capacity as a measure of image structure quality of the photographic image." Proceedings of SPSE Conference on Image Analysis and Evaluation, Toronto, Ontario, Canada, 1976; 122–133.

138. Miyake, Y.; Seidel, K.; Tomamichel, F. Color and tone corrections of digitized color pictures. *J. Photogr. Sci.* 1981, *29,* 111–118.

139. Crane, E.M. *J. SMPTE* 1964, *73,* 643.

140. Gendron, R.G. *J. SMPTE* 1973, *82,* 1009.

141. Granger, E.M. Visual limits to image quality. *J. Proc. Soc. Photo-Opt. Instr. Engrs* 1985, *528,* 95–102.

142. Natale-Hoffman, K.; Dalai, E.; Rasmussen, R.; Sato, M. Proceedings of IS&T Image Processing, Image Quality, Image Capture Systems (PICS) Conference, Savannah, GA, 1999; 266–273.

143. Dalal, E.; Rasmussen, R.; Nakaya, F.; Crean, P.; Sato, M. "Evaluating the overall image quality of hardcopy output." Proceedings of IS&T Image Processing, Image Quality, Image Capture Systems (PICS) Conference, Portland, OR, 1998; 169–173.

144. Inagaki, T.; Miyagi, T.; Sasahara, S.; Matsuzaki, T.; Gotoh, T. Color image quality prediction models for color hard copy. *Proceedings of SPIE* 1997, *2171*, 253–257.

145. Goldman, S. *Information Theory*; Prentice Hall: New York, 1953; 1–63.

146. Felgett, P.B.; Linfoot, E.H.J. *Philos. Trans. R. Soc. London* 1955, *247*, 369–387.

147. McCamy, C.S. On the information in a photomicrograph. *J. Appl. Opt.* 1965, *4*, 405–411.

148. Huck, F.O.; Park, S.K. Optical–mechanical line-scan image process—Its information capacity and efficiency. *J. Appl. Opt.* 1975, *14*, 2508–2520.

149. Huck, F.O.; Park, S.K.; Speray, D.E.; Halyo, N. Information density and efficiency of 2-dimensional (2-D) sampled imagery. *Proc. Soc. Photo-Optical Instrum. Engrs* 1981, *310*, 36–42.

150. Burke, J.J.; Snyder, H.L. Quality metrics of digitally derived imagery and their relation to interpreter performance. *SPIE* 1981, *310*, 16–23.

151. Sachs, M.B.; Nachmias, J.; Robson, J.G. *J. Opt. Soc. Am.* 1971, *61*, 1176.

152. Stromeyer, C.F.; Julesz, B. Spatial frequency masking in vision: critical bands and spread of masking. *J. Opt. Soc. Am.* 1972, *62*, 1221.

153. Miyake, Y.; Inoue, S.; Inui, M.; Kubo, S. An evaluation of image quality for quantized continuous tone image. *J. Imag. Technol.* 1986, *12*, 25–34.

154. Metz, J.H.; Ruchti, S.; Seidel, K. Comparison of image quality and information capacity for different model imaging systems. *J. Photogr. Sci.* 1978, *26*, 229.

155. Hunter, R.; Robinson, A.H. International digital facsimile coding standards. *Proc. IEEE* 1980, *68*(7), 854–867.

156. Rabbani, M. *Image Compression. Fundamentals and International Standards, Short Course Notes*; SPIE: Bellingham, WA, 1995.

157. Rabbani, M.; Jones, P.W. *Digital Image Compression Techniques, TT7*; SPIE Optical Engineering Press: Bellingham, WA, 1991.

158. Joint BiLevel Working Group. *ITU-T Rec. T.82 and T.85*; Telecommunication Standardization Sector of the International Telecommunication Union, March 1995, August 1995.

159. Buckley, R.; Venable, D.; McIntyre, L. "New developments in color facsimile and internet fax." Proceedings of IS&T 5th Annual Color Imaging Conference, Scottsdale, AZ, 1997; 296–300.

160. Huffman, D. A method for the construction of minimum redundancy codes. *Proc. IRE* 1962, *40*, 1098–1101.

161. Lempel, A.; Ziv, J. Compression of 2 dimensional data. *IEEE Trans Info. Theory* 1986, *IT-32* (1), 8–19.

162. Lempel, A.; Ziv, J. Compression of 2 dimensional data. *IEEE Trans Info. Theory* 1977, *IT-23*, 337–343.

163. Lempel, A.; Ziv, J. Compression of 2 dimensional data. *IEEE Trans Info. Theory* 1978, *1T-24*, 530–536.

164. Welch, T. A technique for high performance data compression. *IEEE Trans Comput.* 1984, *17*(6), 8–19.

165. Beretta, G. Compressing images for the internet. *Proc. SPIE, Color Imaging*, III, 1998, *3300*, 405–409.

166. Lee, D.T. Intro to color facsimile: Hardware, software, standards. *Proc. SPIE* 1996, *2658*, 8–19.

167. Marcellin, M.W.; Gornish, M.J.; Bilgin, A.; "Boliek, M.P. An overview of JPEG 2000." SPIE Proceedings of 2000 Data Compression Conference, Snowbird, Utah, 2000, *2658*, 8–30.

168. Sharpe II, L.H.; Buckley, R. "JPEG 2000.jpm file format: a layered imaging architecture for document imaging and basic animation on the web." Proceedings SPIE 45th Annual Meeting, San Diego, CA, 2001; 4115, 47.

169. Dougherty, E.R. Ed. *An Introduction to Morphological Image Processing*; SPIE Optical Engineering Press: Bellingham, WA, 1992.

170. Hamerly, J.R. An analysis of edge raggedness and blur. *J. Appl. Phot. Eng.* 1981, *7*, 148–151.

171. Tung, C. "Resolution enhancement in laser printers." Proceedings of SPIE Conference on Color Imaging: Device-Independent Color, Color Hardcopy, and Graphic Arts II, San Jose, CA, 1997.

172. Tung, C. Piece Wise Print Enhancement. US Patent 4,847,641, July 11, 1989, US Patent 5,005,139, April 2, 1991.

173. Walsh, B.F.; Halpert, D.E. Low Resolution Raster Images, US Patent 4,437,122, March 13, 1984.

174. Bassetti, L.W. Fine Line Enhancement, US Patent 4,544,264 October 1, 1985, Interacting Print Enhancement, US Patent 4,625,222, November 25, 1986.

175. Lung, C.Y. Edge Enhancement Method and Apparatus for Dot Matrix Devices. US Patent 5,029,108, July 2, 1991.

176. Tuijn, W.; Cliquet, C. "Today's image capturing needs: going beyond color management." Proceedings IS&T/SID 5th Color Imaging Conference, Scottsdale, AZ, 1997; 203.

177. Gonzalez, G.; Hecht, T.; Ritzer, A.; Paul, A.; LeNest, J.F.; "Has, M. Color management—How accurate need it be." Proceedings IS&T/SID 5th Color Imaging Conference, Scottsdale, AZ, 1997; 270.

178. Frazier, A.L.; Pierson, J.S. Resolution transforming raster based imaging system, US Patent 5,134,495, July 28, 1992, Interleaving vertical pixels in raster-based laser printers, US Patent 5,193,008, March 9, 1993.

179. Has, M. Color management—Current approaches, standards and future perspectives. IS&T, 11NIP Proceedings, Hilton Head, SC, 1995; 441.

180. Buckley, R. *Recent Progress in Color Management and Communication;* Society for Imaging Science and Technology (IS&T): Springfield, VA, 1998.

181. Newman, T. "Making color plug and play." Proceedings IS&T/SID 5th Color Imaging Conference, Scottsdale, AZ, 1997; 284.

182. Chung, R.; Kuo, S. "Colormatching with ICC Profiles—Take one." Proc. IS&T/SID 4th Color Imaging Conference, Scottsdale, AZ, 1996; 10.

183. Rickmers, A.D.; Todd, H.N. *Statistics, an Introduction;* McGraw Hill: New York, 1967.

184. Dvorak, C.; Hamerly, J. Just noticeable differences for text quality components. *J. Appl. Phot. Eng.* 1983, *9,* 97–100.

185. Hamerly, J. Just noticeable differences for solid area. *J. Appl. Phot. Eng.* 1983, *9,* 14–17.

186. Bartleson, C.J.; Woodbury, W.W. Psychophysical methods for evaluating the quality of color transparencies III. Effect of number of categories, anchors and types of instructions on quality ratings. *J. Photo. Sci. Eng.* 1965, *9,* 323–338.

187. Stevens, S.S. *Psychophysics: Introduction to Its Perceptual, Neural and Social Prospects;* John Wiley and Sons: New York, 1975; Reprinted: Transactions Inc.: New Brunswick, NJ, 1986.

188. Thurstone, L.L. Rank order as a psychophysical method. *J. Exper. Psychol.* 1931, *14,* 187–195.

189. Stevens, S.S. On the theory of scales of measurement. *J. Sci.* 1946, *103,* 677–687.

190. Kress, G. *Marketing Research,* 2nd Ed.; Reston Publishing Co. Inc.: a Prentice Hall Co.: Reston, VA, 1982.

191. Morrissey, J.H. New method for the assignment of psychometric scale values from incomplete paired comparisons. *JOSA* 1955, *45,* 373–389.

192. Bartleson, C.J.; Breneman, E.J. Brightness perception in complex fields. *JOSA* 1967, *57,* 953–960.

193. Bartleson, C.J. The combined influence of sharpness and graininess on the quality of color prints. *J. Photogr. Sci.* 1982, *30,* 33–45.

194. Bartleson, C.J.; Grum, F. Eds. Visual measurements. In *Optical Radiation Measurements,* Academic Press: Orlando, FL, Vol. 5, 1984.

195. Gescheider, G.A. *Psychophysics: The Fundamentals,* 3rd Ed.; Lawrence Erlbaum Assoc. Inc.: Mahwah, NJ, 1997.

196. Guilford, J.P. *Psychometric Methods;* McGraw Hill Book Co.: New York, 1954.

197. Malone, D. Psychometric methods. In *SPSE Handbook of Photographic Science and Engineering,* Chapter 19.4; A Wiley Interscience Publication, John Wiley &Sons: New York, 1973; 1113–1128.

198. Nunnally, J.C.; Bernstein, I.R. *Psychometric Theory,* 3rd Ed.; McGraw Hill Book Co.: New York, 1994.

199. Torgerson, W.S. *Theory and Methods of Scaling;* J. Wiley &Sons: New York, 1958.
200. Koren, N. Making fine prints in your digital darkroom, Understanding image sharpness and MTF; http://www.normankoren.com, validated. 2/11/2009—Good practical and theoretical insights with pointers to a software product "Imatest" used to measure these and many other characteristics of cameras and scanners.
201. A. Davidhazy, Ed. http://www.rit.edu/cias/photo/ipt-faculty/ (Rochester Institute of Technology. Rochester, NY, 2009) which lists faculty and related courses that influenced updates herein.
202. Hunter, R.S.; Harold, R.W. *The Measurement of Appearance,* 2nd Ed.; John Wiley and Sons: New York, 1987; 191.
203. Scheff'e, H. An analysis of variance for paired comparisons. *J. Am. Statist. Assoc.* 1952, *47,* 381–395.
204. Daniels, C.M.; Ptucha, R.W.; Schaefer, L. "The necessary resolution to zoom and crop hardcopy images." Proceedings of IS&T Image Processing, Image Quality, Image Capture Systems (PICS) Conference, Savannah, Georgia, 1999; 143.
205. Nakamura, J. *Image Sensors and Signal Processing for Digital Still Cameras;* Taylor and Francis: Boca Raton, FL, 2006; Mizoguchi, T. Ch 6: Evaluation of image sensors, 179–203.

4

Polygonal Scanners: Components, Performance, and Design

Glenn E. Stutz

Lincoln Laser Company
Phoenix, Arizona, USA

CONTENTS

4.1 INTRODUCTION

Polygonal scanners have found a role in a wide range of applications including inspection, laser printing, medical imaging, laser marking, laser radar, and displays, to name a few. Ever since the laser was first discovered, engineers have needed a means to move the laser output in a repetitive format.

The term "polygonal scanner" refers to a category of scanners that incorporate a rotating optical element with three or more reflective facets. The optical element in a polygonal scanner is usually a metal mirror. In addition to the polygonal scanner other scanners can have as few as one facet such as a pentaprism, cube beam splitter, or "monogon." This section will concentrate on scanners that use a metal mirror as the optical element.

Polygonal scanners are not the only technology available to move an optical beam. These other technologies include galvanometers, micromirrors, hologons, piezo mirrors, and acousto-optic deflectors. Each technology has a niche where it excels. Polygonal scanners excel in applications requiring unidirectional scans, high scan rates, large apertures, large scan angles, or high throughputs. The polygonal scanner in most applications is paired with another means for beam steering or object motion to produce a second axis. This creates a raster image with the polygonal scanner producing the fast scan axis of motion.

This chapter will provide information on types of scan mirrors, fabrication techniques to create these mirrors, and typical specifications for these mirrors. The motor and bearing systems used with the mirror to construct a scanner are covered. A section on properly specifying a polygonal scanner as well as the cost drivers in the scanner design is included. The incorporation of the scanner into a scan system including system level specifications and design approaches is reviewed. The final section covers system image defects and methods used to compensate for these defects in a scanning system.

4.2 TYPES OF SCANNING MIRRORS

There are many types of scan mirrors, but most can be included in the following categories:

1. Prismatic polygonal scanning mirrors
2. Pyramidal polygonal scanning mirrors

3. "Monogons"
4. Irregular polygonal scanning mirrors

4.2.1 Prismatic Polygonal Scanning Mirrors

A regular prismatic polygon is defined as one having a number of plane mirror facets that are parallel to, equidistant from, and face away from a central rotational axis (Figure 4.1). This type of scan mirror is used to produce repetitive scans over the same image plane. It is the most cost effective to manufacture and therefore finds its way into the vast majority of applications including barcode scanning and laser printing. An illustration of why the manufacturing cost can be lower than other types of scan mirrors is shown in Figure 4.2.
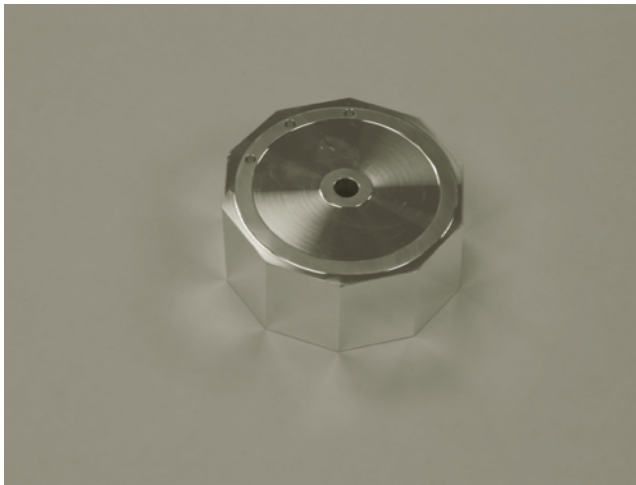
FIGURE 4.1
Regular prismatic polygonal scanning mirror.
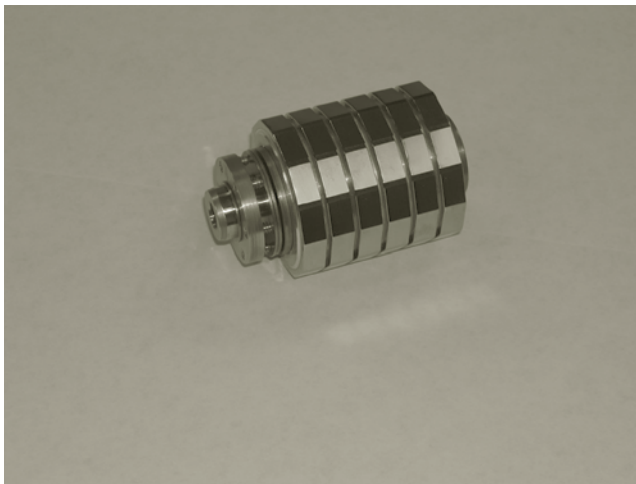
FIGURE 4.2
Mirror stack reduces fabrication costs.

Here we see a stack of mirrors that can be moved through the manufacturing process as a single piece resulting in less handling, more consistency, and less machining time.

4.2.2 Pyramidal Polygonal Scanning Mirrors

A regular pyramidal polygon is defined as one having a number of facets inclined at the same angle, usually 45°, to the rotational axis (Figure 4.3). This type of polygon is expensive to manufacture since one cannot stack mirrors together to process at the same time as is done with regular prismatic polygons.

A significant feature of the 45° pyramidal polygon is that it can produce half the output scan angle of a prismatic polygon for the same amount of shaft rotation. This feature can be used to the system designer's advantage by reducing data rates for a given polygon rotation speed. Prismatic polygons are used primarily with the input beam perpendicular to the rotation axis whereas pyramidal polygons are used primarily with the input beam parallel to the rotation axis (Figure 4.4).

4.2.3 "Monogons"

"Monogons" are scan mirrors where there is only one facet centered on the rotational axis. Because there is only one facet, a monogon is not a true polygon but is an important subset of the scan mirror family. Monogons are also referred to as truncated mirrors and find application in internal drum scanning. In a typical system employing a monogon, the laser is directed toward the monogon along the rotation axis and the output sweeps a circle on an internal drum as the scanner rotates. This type of scan system can produce very accurate spot placement and very high resolution and finds application in the prepress market. An example of a monogon scan mirror is shown in Figure 4.5.

4.2.4 Irregular Polygonal Scanning Mirrors

An irregular polygonal scanning mirror is defined as one having a number of plane facets that are at a variety of angles with respect to, and face away from, the rotational axis
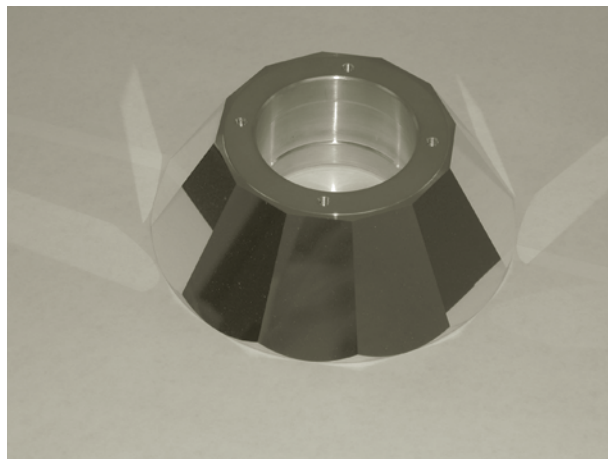
FIGURE 4.3
Regular pyramidal polygonal scanning mirror.
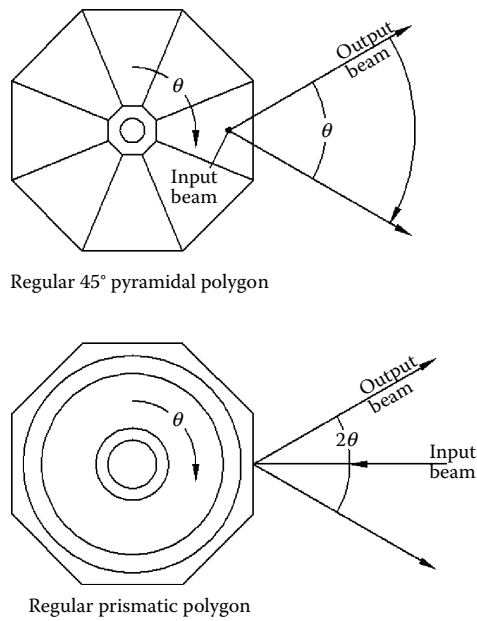
FIGURE 4.4
Scan angle versus rotation angle.

FIGURE 4.5
"Monogon."

(Figure 4.6). The unique feature of this type of scan mirror is that it can produce a raster output without a second axis of motion. The resulting output scans are nonsuperimposing if the facets are at different angles. This type of scanner finds its way into coarse scanning applications such as:

1. Point-of-sale barcode readers
2. Laser heat-treating systems
3. Low resolution writing and display systems
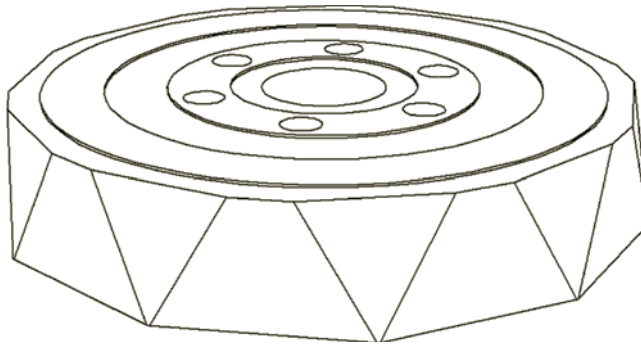
FIGURE 4.6
Irregular polygonal scanning mirror.

These polygons typically cost significantly more than regular polygons because their asymmetry prevents cost savings from stacking. Another disadvantage of these scanners is the inherent dynamic imbalance of the polygon during rotation. This limits use to low-speed applications. A special case where equal and opposing facets are used on each side of the polygon helps with the balance problem. The result is the scan pattern is generated twice each revolution.

Now that the types of scan mirrors have been covered, a logical next step is to consider the materials used to fabricate the mirrors. The following section addresses the most common materials in use today.

4.3 MATERIALS

Material selection for polygonal mirrors is driven by considerations of performance and cost. The most common materials for polygonal mirrors are aluminum, plastic, and beryllium. Facet distortion/flatness is a key performance consideration when choosing a material.

Aluminum represents a good trade-off between cost and performance. This material has good stiffness, is relatively light, and lends itself to low-cost fabrication methods. The upper limit for the use of aluminum mirrors without the risk of facet distortion beyond $\lambda/10$ is on the order of a tip velocity of 76 m/s. Above this speed the size of the facet, the disc shape, and the mounting method all play a role in the distortion of the facet. It is recommended that a finite element analysis be performed if you intend to operate above this level. An example of the shape change due to high-speed rotation, for a six-faceted polygon, is shown in Figure 4.7.

Plastic is used in applications where cost is the primary concern and performance is good enough for the application. An example is in the hand-held barcode market and other short-range, low-resolution scanning applications. Injection molding techniques have come far in the past few years but it is still difficult to reliably produce plastic mirrors larger than 2-mm diameter with facets flat to better than 1 wave.

Beryllium has been used successfully in applications where high speed and low distortion are required. It is a very expensive substrate and produces toxic dust when machined, requiring specialized extraction and filtration equipment. Therefore it does not find wide usage and is a very expensive solution. Beryllium is typically nickel plated prior to polishing. The nickel plating seals in the beryllium and removes the risk of toxic dust.
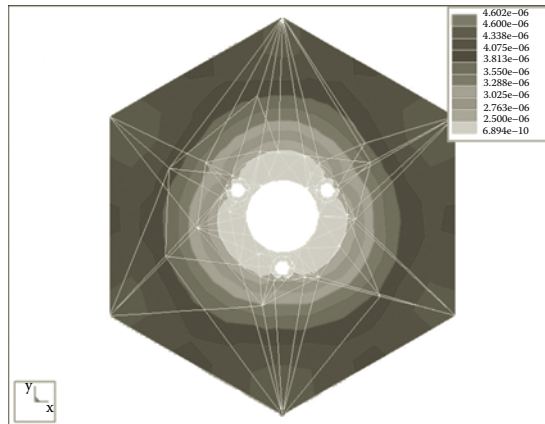
FIGURE 4.7
Finite element analysis of polygon rotating at 30,000 rpm.

In some high-speed applications, distortions in facet flatness can be tolerated. In these cases the structural integrity of the polygonal mirror must be considered. The speed at which the dynamic stress will reach the yield strength (causing permanent distortion and dangerously close to the breaking speed) is found using the formula below.[1]

$$B = \sqrt{\frac{S}{(7.1e-6)w[(3+m)R^2 + (1-m)r^2]}} \tag{4.1}$$

where B = maximum safe speed (rpm), S = yield strength (lb/in^2), w = weight of material (lb/in^3), R = outer tip radius (in), r = inner bore radius (in), and m = Poisson's ratio. This formula does not have a margin of safety, so it would be wise to consider this and back off the results by an appropriate margin.

4.4 POLYGONAL MIRROR FABRICATION TECHNIQUES

Aluminum is the most common substrate for the fabrication of polygonal mirrors. There are two techniques for fabricating aluminum polygons that are widely used. These techniques are conventional polishing and single point diamond turning. Each technique has its advantages and the application will usually dictate the technique used.

4.4.1 Conventional Polishing

Conventional polishing in this context is pitch lapping in much the same manner as glass lenses and prisms are polished. A polishing tool is covered with a layer of pitch and a polishing compound is used that is a slurry composed of iron oxide and water. The pitch lap rubs against the optic using the polishing compound to remove material. Pitch lapping can be used to produce high-quality surfaces on a number of materials. Unfortunately, aluminum is not one of them. The aluminum surface is too susceptible to scratches during the

polishing process. New techniques have been developed, but these rely on minimal abrasive mixtures and therefore material removal rates that are too slow to be cost effective.

Because one cannot polish the aluminum directly, a plating must be applied prior to polishing. Electroless nickel is the most common plating applied. This combination provides the low cost and ease of machining of the aluminum substrate with the superior polishing properties and durability of nickel. The mirror facets are polished individually, blocked up in a surround as shown in Figure 4.8. If the polygonal mirror is regular then a stack of polygons can be polished in one setup.

4.4.2 Single Point Diamond Turning

Single point diamond machining is a process of material removal using a finely sharpened single-crystal diamond-cutting tool. Diamond machining centers are available in the form of lathes and mills. The use of ultra-precise air-bearing spindles and hydrostatic table ways, coupled with vibration isolating mounting pads, enables machining to optical quality surface specifications. Figure 4.9 shows a diamond machining center with a polygon in process.

Diamond machining has proven to be an efficient process for generating optical surfaces since it can be automated and the process time is a small fraction of the time required for conventional polishing. The diamond machined mirror is typically fabricated from aluminum, but satisfactory results have been obtained on other substrates. The diamond machined mirror face appears to be a perfect mirror, but upon close inspection the residual tool marks on the surface are apparent. These tool marks create a grating pattern on the surface. This grating pattern can increase the scatter coming from the surface, particularly at wavelengths below 500 nm.

4.4.3 Polishing versus Diamond Turning

Diamond turned aluminum scan mirrors are by far used in the highest volumes. This is due to the low manufacturing cost and good performance characteristics. Polished mirrors,
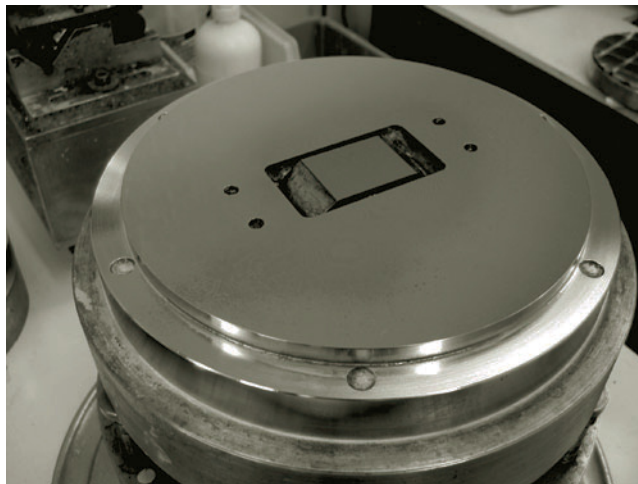
FIGURE 4.8
Conventional polishing of polygonal mirrors.

FIGURE 4.9
Diamond turning center.

however, have found a niche where they outperform diamond turned mirrors and justify the higher cost. These applications tend to be very scatter sensitive, such as writing on film. A polished mirror can approach surface roughness levels of 10 Å rms whereas diamond turned mirrors are limited to roughness levels of about 40 Å rms. Short-wavelength applications may also require the lower scatter of a polished mirror surface. Applications below 400 nm frequently need the lower scatter level of polished mirrors and the scatter can be a problem in applications up to about 500 nm.

There is also a difference in the type of scatter produced by the two surface types. Polished surfaces tend to produce lambertian scatter pattern which has a large wide-angle component whereas diamond turned surfaces produce a large percentage of low angle scatter. The low angle scatter travels along a path close to the specular reflection so it can be difficult to mask out in an optical system.

4.5 POLYGON SPECIFICATIONS

In addition to selecting the type of polygon, the material to use, and the fabrication technique, several mechanical specifications need to be established. In a perfect world the polygon would have exactly the dimensions and angles that we specify on a print. Real-world manufacturing limitations cause us to have to add in a practical set of tolerances on the polygon and evaluate how these imperfections would affect system performance. Some of the items that need to be specified on a polygonal mirror include:

1. Facet-to-facet angle variance
2. Pyramidal error
3. Facet-to-axis variance (total and adjacent facet)

   4. Facet radius:
       i.   nominal tolerance,
       ii. variation of all facets tolerance;
   5. Surface figure (composed of power and irregularity)
   6. Surface quality and scatter

4.5.1 Facet-to-Facet Angle Variance

The definition of facet-to-facet angle variance (Δ) is the variation in the angle between the normals (⊠) of the adjacent facets on the polygon (Figure 4.10). This variation in angle causes timing errors from one facet to the next as the polygon rotates. Typical values for this angle range from ±5 arc s to ±30 arc s. Most scanning systems are not sensitive to errors in this range because of the use of start of scan sensors and/or encoders.

   This, as well as all other angular tolerances, is a mechanical measure. The system designer needs to ensure that everyone is discussing these errors in mechanical rather than optical terms since there is a factor of two involved between the mechanical tolerances and the optical effects.

4.5.2 Pyramidal Error

Pyramidal error is defined as the average variation (Ω) from the desired angle between the facet and the mirror datum (Figure 4.11). This variation results in a pointing error of the output beam and can also cause scan line bow. Typical values for this specification are ±1 arc minute (mechanical).

4.5.3 Facet-to-Axis Variance

This is defined as the total variation of the pyramidal error from all the facets within one polygon (Figure 4.12). This is a critical specification for the mirror and contributes to a scanner specification of dynamic track, discussed later. Typical values for this specification range from 2 arc s to 60 arc s (mechanical).
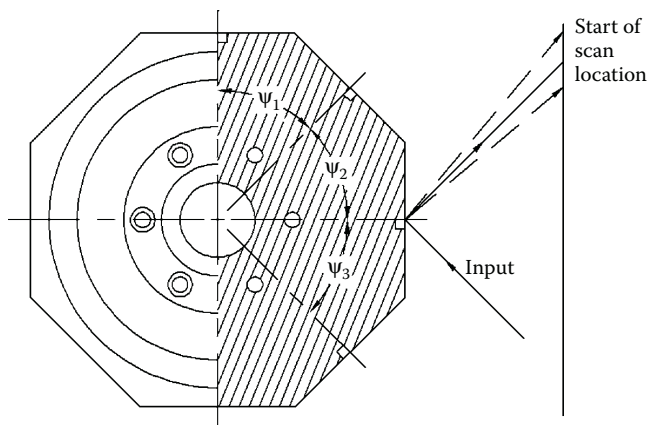
FIGURE 4.10
Facet-to-facet angle variance.
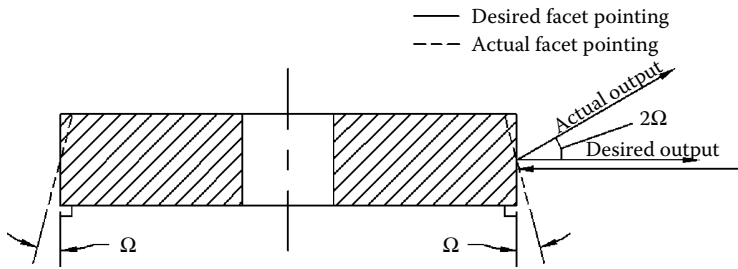
FIGURE 4.11
Pyramidal error.
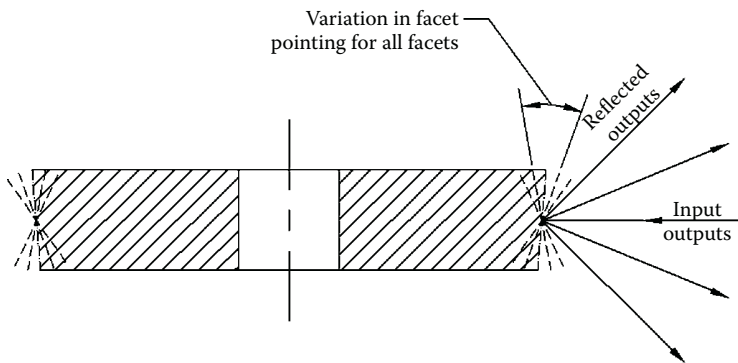
FIGURE 4.12
Facet-to-axis variance.

Another parameter related to this is the adjacent facet-to-axis variance. This is defined as the largest step in the pyramidal angle from one facet to the next within a polygon. This is important to control in order to reduce banding artifacts in the final system. Typical values for this specification are in the range of 1–30 arc s (mechanical).

Optical scanning systems may employ correction devices that allow this value to be reduced. System resolution plays a large part in determining the actual value required. Film writing applications tend to have the tightest requirements and passive reading systems tend to have the loosest requirements.

4.5.4 Facet Radius

The facet radius (referred to as facet height by some manufacturers) is the distance from the center of the polygon to the facet. The variation in this radius within the polygon and the tolerance on the average radius are important to specify. The average facet radius is important because it locates the facet in the optical system. The variation of this radius within a polygon causes errors in the focal plane location from one facet to the next. It also causes velocity variations within the scan line, which are usually small and show up as jitter errors. Typical values for these parameters are ±60 microns for the facet radius average position and ±25 microns for the facet radius variation within a polygon.

4.5.5 Surface Figure

Surface figure is the macro shape of the polygon facet and is measured as the deviation from an ideal flat surface. The flatness of polygon facets will have an impact both on the aberrations in the beam as well as the pointing of the beam. The aberrations can affect the final focused spot size in the scan system. The pointing error results in velocity variations across the scan.

Several factors influence the flatness of polygon facets:

   1. Initial fabrication tolerances
   2. Distortion due to mounting stresses
   3. Distortion due to forces induced when rotating at high speeds
   4. Distortion due to long-term stress relief

Interferometers are commonly used to measure static flatness. The flatness is specified in wavelengths, λ, (or fractions thereof) of light. A typical flatness specification is: $\lambda/8$ at 633 nm. Departure from flatness can have a variety of forms, depending on how the surface was fabricated. For example, conventionally polished mirror surfaces tend to depart from flat in a regular spherical form, either convex or concave. Diamond machined surfaces usually depart from flat in a regular cylindrical form, either convex or concave. A polygon will typically have two specifications related to flatness, a surface power specification and irregularity. The irregularity is defined as the deviation from a best-fit sphere. Another common way of specifying the optical surface is in terms of power and pv—power (peak to valley error minus power), which separate the regular and irregular shapes. Most polygons used in printing applications are specified in the $\lambda/8$ to $\lambda/10$ range at the wavelength of interest.

4.5.6 Surface Quality and Scatter

Ideally a reflective optical surface will reflect all of the incident light without introducing any scattered components. In reality an optical surface has multiple defects of various sizes. The U.S. military developed a scratch and dig specification for surface defects, which is included in MIL-PRF-13830B and is in broad use within the optics industry. This method of quality determination involves close examination of a surface and identifying a scratch and dig level in a given unit area. A typical high-quality conventionally polished polygon will have a quality level of 40–20 scratch and dig.

Machined optical surfaces on the other hand, are made up of a precise regular pattern of machine tool marks, which are sufficiently high in frequency and low in height errors as to behave as a plane mirror at most visible and infrared wavelengths. The scratch and dig specification must be supplemented with an additional measure of surface quality here. A more representative definition for the overall surface quality is the rms surface roughness. The rms surface roughness can be measured directly by mechanical or optical profilometry means or indirectly by measuring the scatter from the surface.

A special test system is required to measure the scatter from the surface and correlate this to an rms roughness value. Different tests must be used for the measurement of scatter of diamond turned and conventionally polished mirrors. Diamond turned surfaces produce a significant fraction of scattered energy in a narrow cone around the reflected beam. Conventionally polished mirrors have the majority of their scattered energy in a cone significantly greater than the divergence angle of the reflected beam.

Conventionally polished mirrors can be tested using an integrating sphere that gathers a wide cone angle (Figure 4.13). Scattered light in the cone of 4–180° is gathered with this test method. The diamond turned mirrors can be tested using a combination of the integrating sphere and a near angle test (Figure 4.14). Scattered light in the cone of 0.4–4° is gathered using this test.

A correlation has been developed between the rms surface roughness and the total integrated scattered incident light:[2]

$$\text{rms roughness} = \frac{1\sqrt{ln\left(1 - TIS\right)}}{4p} \tag{4.2}$$

where *TIS* is the total integrated scatter.
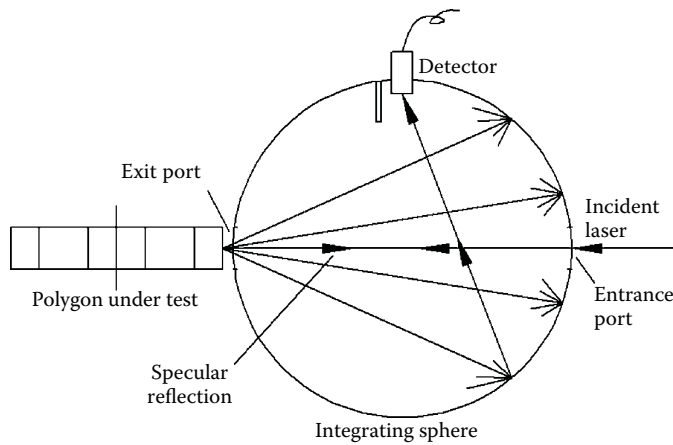
FIGURE 4.13
Test for wide angle scatter.
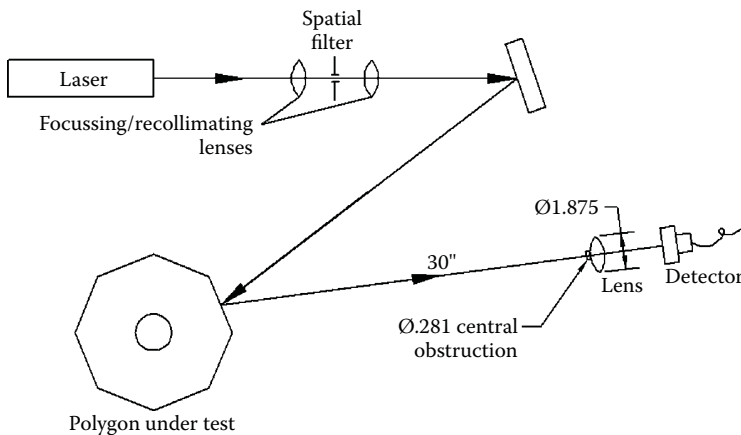
FIGURE 4.14
Test for near angle scatter.

The combination of scratch and dig along with rms surface roughness provide a good description of the surface structure higher in frequency than surface figure.

4.6 THIN FILM COATINGS

There are two major functions of optical coatings on polygons: to improve the reflectance of the surface and/or to improve durability. In the case of diamond machined polygons, the substrate is usually aluminum (in itself a good reflector over most of the visible spectrum). This aluminum surface is too soft without a coating. It is easily scratched during even a light cleaning so one solution is to apply a thin layer of silicon monoxide, a dielectric material, as a surface protector. The optical thickness is usually about one-half wavelength at the wavelength of interest. This material is more durable than the base aluminum and can be readily cleaned. The coating just described is typically referred to as a protected aluminum coating. This coating has a reflectivity of >88% across the 450–650 nm range.

The protected aluminum coating is fine for many applications and because of its simplicity is relatively inexpensive. Many applications, however, require enhancement coatings due to needs for higher reflectivity performance at various wavelengths.

The first layer deposited in most applications is a binding layer, then a metal, such as aluminum, silver, or gold. The layer or layers above the metal are composed of dielectric materials. The metal is selected based on the wavelengths of interest. As mentioned earlier, aluminum is a good choice in the visible region. It is also selected for ultraviolet applications since its reflectivity can be enhanced in this region with a dielectric stack. Gold is often selected as the base metal in applications above 600 nm. It has good reflectivity at 600 nm (90%) and very good reflectivity from 1 micron, out past 10.6 microns (>98%).

Silver exhibits very desirable reflectance characteristics over a broad spectrum and is frequently considered as a material for polygon coating. In practice, however, it is frequently a disappointing choice for the long term. The slightest pinhole (or minute scratch from cleaning) will expose the silver to reactive contaminants from the atmosphere, which over time (several days or weeks) will diffuse into the silver, producing an expanding blemish. Aluminum, which initially exhibits somewhat lower values of reflectance than silver, is far superior in terms of durability.

Common dielectric materials used to protect the surface and enhance reflectivity are silicon monoxide, silicon dioxide, hafnium oxide, and titanium dioxide. A quarter wave stack combining high refractive index and low refractive index materials is used to enhance the reflectivity in the wavelength region of interest. The term quarter wave stack refers to an alternating series of high and low index materials that are one-quarter of an optical wavelength thick at the wavelength of interest. The design of this quarter wave stack can be tailored to raise the reflectivity of the base metal significantly in various regions of interest. Most companies will offer a variety of standard reflectivity enhancing coatings for different wavelengths.

Thin film coatings are applied in vacuum deposition chambers such as the one shown in Figure 4.15. The tooling to coat a polygon is specialized because the polygon has optical surfaces around its periphery. The polygons are stacked onto coating arbors that are placed in the chamber above the evaporant sources. The arbors are rotated with a drive mechanism during the deposition process at a constant rate so that all facets will see the same thickness of coating material. The rotation rate must be fast enough that the time for

FIGURE 4.15
Thin film deposition chamber.

one revolution is a small fraction of the deposition time for a single layer. Otherwise there will be significant variation around the polygon depending on when the shutter is opened and closed.

Reflectance uniformity both within a facet and facet-to-facet is an important polygon specification. The reflectance uniformity can impact the accuracy of written or read images. In practice, the reflectance uniformity of rather large polygon facets (e.g., a few inches square) is more difficult to achieve than if the facets are small (e.g., $\frac{1}{2}$in square). This has to do with the consistency of the cleaning of the surface prior to coating and the variations in deposition rates with location and time within the coating chamber.

Aluminum polygons cannot be heated up to high temperatures during a coating run as one would typically do for good adhesion and layer density. The shape of the polygons makes them susceptible to slight stress changes during this heating cycle. This results in changes to the facet flatness. The coating process should be designed to keep the polygons below a temperature of 225°C to maintain the flatness.

Crucial to all of the desired characteristics of the coating of a polygon is its cleanliness prior to coating. Irrespective of the fabrication methods, polygons will be handled prior to coating during the inspection processes, transport, and installation into the coating chamber. The polygon must be cleaned thoroughly to remove foreign material that will degrade surface quality, prevent good coating adhesion, or outgas in the coating system.

Several tools are available to measure the optical performance of the coating. Common measuring tools to determine reflectance are spectrophotometers and laser reflectometers. Spectrophotometers are used to provide information on the reflectance versus wavelength. The majority of spectrophotometers with a reflectance measuring attachment are limited to small sample sizes on the order of 1 to 2 in in diameter. This fact usually precludes measuring the polygon itself. A witness sample is coated at the same time as the polygon and can be used to represent the actual part performance. This can be a reliable method of ascertaining the performance of the polygon as long as the witness sample has a similar surface preparation and quality level to the polygon. This means that diamond turned witness samples should be used with diamond turned polygons and polished witness samples used with polished mirrors.

Laser reflectometers compare the reflected beam to the incident beam at a specific wavelength and can be designed to test over a range of angles with either S or P polarization. Reflectometers are useful for determining performance at one specific wavelength but cannot provide broadband information.

4.7 MOTORS AND BEARING SYSTEMS

The polygonal mirror requires a bearing system and a drive mechanism to turn it into a functional scanner. Drive mechanisms include pneumatic, AC hysteresis synchronous, and brushless DC. Bearing systems used in most applications are ball bearing, aerostatic air bearings, or aerodynamic air bearings.

4.7.1 Pneumatic Drives

Much of today's scan mirror technology has evolved from the development of ultra-high-speed polygon/turbine motors for the high-speed photography industry. Compressed air turbines continue to offer an attractive method of rotating a polygonal mirror at speeds beyond the capability of electric motors. The advantages of turbine drives are:

1. Substantial horsepower can be delivered to the scan mirror to produce rapid acceleration and very high speed (up to 1,000,000 rpm).
2. Compact in size and low in weight in proportion to delivered power.
3. Can be equipped with shaft seals so that the scan mirror can be used in a partial vacuum.

The disadvantages of turbine drives are:

1. Require a compressed air source
2. Asynchronous devices
3. Relatively high in cost
4. Relatively short total running life

Pneumatic drives are only recommended for short duty cycles and where ultra-high speed is essential.

4.7.2 Hysteresis Synchronous Motors

The rotor of a hysteresis synchronous motor is usually fabricated from a single piece of hardened steel selected out of a group (predominantly alloyed with cobalt) that exhibits substantial hysteresis loss. This resistance to the movement of magnetic flux in the material imparts torque to a rotor out of sync with the drive current. This torque is responsible for the motor's ability to start rotation. When the rotor approaches the speed of the stator flux, it becomes permanently magnetized and "locks in" to synchronism with the drive. If the motor is turned off and restarted, the stator flux demagnetizes the rotor and hysteresis

takes over again. The synchronous mode of operation is more efficient than the hysteresis or startup mode, and in many systems a sync detector is used to reduce drive current after the motor is locked in to save energy and reduce heating.

AC hysteresis synchronous motors exhibit a characteristic called phase jitter (hunting). The rotors behave as though they were coupled to the drive waveform by a spring. Within synchronism the rotor springs forward and back in phase at a rate determined by the spring rate (flux density) and the torque/inertia ratio of the system. Typically, the frequency of this phase jitter is in the range of 0.5–3 Hz, at an amplitude of a few degrees (1–6° peak to peak). Under perfect conditions this jitter damps to zero values of amplitude. However, perfection is seldom seen and continual recurrence of jitter may be expected, caused by electrical transients on the input, mechanical shock to the assembly, variable resistance torque of the motor bearings, and so on. For many systems the 0.5–0.01% velocity error contribution of phase jitter is acceptably small. If this is not the case then a feedback loop is needed to reduce this level.

4.7.3 Brushless DC Motors

Brushless DC motors are by far the most common motor used to drive polygonal scanners. These motors use a permanent motor magnet and a stator that supplies the varying magnetic force. Motor magnets are composed of various materials including neodymium and ferrite depending on the application. Stators can be iron-based or ironless, with or without teeth. The number of magnetic poles is usually determined by the operating speed. Low-speed motors tend to have higher pole counts (8–12) while higher speed motors (>10,000 rpm) tend to have lower pole counts (4–6). The reason for the large number of poles at low speed is to achieve smoother rotation. At higher speeds this is not required and the lower pole count motors have less losses because the stator flux speed is lower.

Brushless motors do not have the hunting problem associated with AC hysteresis synchronous motors. The motor controls used to drive these motors can hold a tighter control loop. These motors can exhibit more high-frequency variations due to the torque available to rapidly change speed. This high-frequency velocity change is referred to as jitter. The amount of jitter is related to the rotor inertia and the number of feedback pulses per revolution. At higher speeds the inertia smoothes out the rotation and limits the amount of jitter. At lower speeds the number of feedback pulses helps keep the control loop errors small and therefore less velocity jitter when the motor has a correction torque applied. Hall effect devices are used at higher speeds to provide magnetic position feedback to the controller. Hall effect devices at lower speeds, where inertia is lower, can induce jitter as the controller chases the positional and triggering errors of the Hall effect devices. At lower speeds an encoder on the rotor may be required to achieve low jitter levels. Even incremental encoders can induce jitter errors due to disc quality, alignment, and component quality.

4.7.4 Bearing Types

Polygonal scanners require a bearing support system to allow the rotor to rotate. The most common bearings used in scanners are:

1. Ball bearings
2. Aerostatic air bearings
3. Aerodynamic air bearings

These three types of bearing systems are discussed in detail in other chapters in this book. Ball bearings are used where possible due to their low cost. Applications requiring speeds less than 20,000 rpm and that can tolerate bearing nonrepeatable errors, both in scan and cross scan, are candidates for ball bearings.

Aerodynamic air bearings have made large inroads in laser scanning since the 1980s. An aerodynamic bearing generates its own air pressure as it rotates. It is commonly designed with two close-fitting cylinders for the radial bearing. The axial bearing can be either an air thrust bearing or magnetic bearing. These systems have many advantages over both conventional ball bearing systems and aerostatic air bearings. The speed range for aerodynamic bearings is from approx. 4000 rpm to over 100,000 rpm. These bearings are only slightly more costly than an equivalent ball bearing system. There is no wear while operating and they require no external pressure support equipment. These bearings have been developed to withstand over 20,000 start/stop cycles. Aerodynamic bearings do have some limitations that limit their application. They are not well suited to dirty environments. Many designs exchange outside air frequently during operation, thereby ingesting the outside debris. Most designs cannot withstand high shock loads because the bearing stiffness is limited. The mass of the optic is limited in many applications due to both the lack of support and the need to withstand constant starting and stopping. Additional mass causes added wear to the bearing during startup and shutdown.

Aerostatic air bearings provide the ultimate in performance at a high cost. An aerostatic bearing uses pressurized air and closely spaced axial and radial bearing surfaces to float the rotor. When pressurized, the bearing has no contacting parts, resulting in extremely long life. These bearings are very stiff and have wobble errors less than 1 arc s. They are capable of supporting heavy loads and do not suffer from wear at startup and shutdown. They do require external components to supply the pressure to the bearing. This increases system complexity as well as cost.

4.8 SCANNER SPECIFICATIONS

Once the polygon, motor, and bearing system have been decided on, the packaging of the assembly becomes the next concern. One of the key elements in attaining high scanner performance is the mounting of the scan mirror to the rotating spindle.

To preserve the facet flatness achieved during initial polygon fabrication, it is necessary to fasten the polygon to its drive spindle with care, particularly if $\lambda/8$ or better flatness is required. The interface between the mirror and the rotor must not induce stress in the mirror that is translated out to the facets.

A typical mounting scheme is shown in Figure 4.16. In this case the datum surface of the polygon and the locating annulus of the mounting hub are lapped to optical quality so that when the two are firmly held together, distortions are minimized.

Equally important to the accurate mounting of mirror datum and rotor hub surfaces is cleanliness at assembly and the appropriate torque levels of the fastening screws. Polygons can be attached in the manner described in low- and medium-speed applications. When tip velocities approach 76 m/s, other methods of mounting need to be considered.

In many applications the facets can be allowed to distort as long as they all change by the same amount. A symmetrical mounting method with screws aligned with every apex will work in this type of application. Other applications cannot stand significant shape change

on the facets and require a true radially symmetric mounting method such as clamping. Clamping has been used successfully but this also requires a radial attachment means that may consist of an elastic material or aluminum shaped to have a spring force.

Once the polygonal mirror is integrated with the motor and the bearing system it can be referred to as a polygonal scanner. The scanner assembly is defined by performance specifications that include:

1. Dynamic track
2. Jitter
3. Speed stability
4. Balance
5. Perpendicularity
6. Time to sync

4.8.1 Dynamic Track

Dynamic track is defined as the total mechanical angular variation of the facets perpendicular to the scanning direction. This is illustrated in Figure 4.17. An optical beam illuminating a polygon with a dynamic track of 10 arc s will have a scan envelope from all the
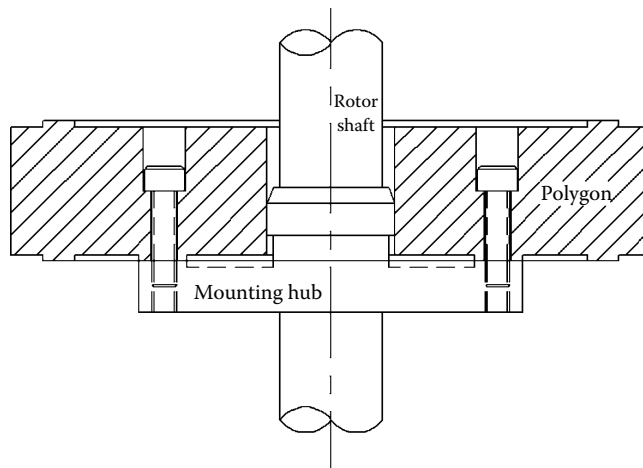
FIGURE 4.16
Mirror/rotor interface.
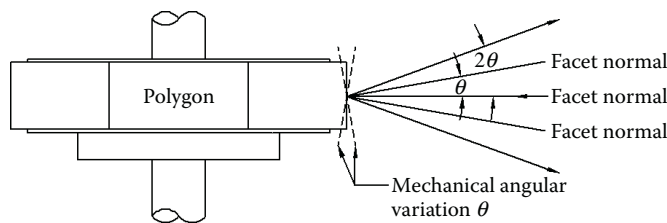
FIGURE 4.17
Dynamic track errors.

facets of 20 arc s perpendicular to the scan direction. This is caused by an angle doubling effect on reflection from the rotating mirror.

There are four significant contributors to dynamic track error. The first is the polygon itself, which has a variation in the angle of each facet and can have a residual pyramidal (squareness) error. The second contribution comes from the mounting of the polygon to the rotating shaft. If the polygon is not perfectly perpendicular to the rotating shaft, then the facets will change their pointing in a sinusoidal manner with a period of one revolution. These first two contributions are fixed and repeatable. The third contribution is a random nonrepeatable error caused by the bearing support system. The random component of the dynamic track error will be 1–2 arc s for a ball bearing assembly and less than 1 arc s for air bearing assemblies. The fourth contribution, which is a repeatable error, comes from the grinding operation performed on the mounting hub of the spindle. This hub needs to be perpendicular to the rotation axis of the spindle and typically a few arc s error is present due to the grinding tolerance.

The repeatable component of dynamic track (which tends to be larger) will show up in a laser writing system as a banding artifact. The line spacing will not be uniform and will repeat the pattern on each revolution of the polygon. Dynamic track errors can be reduced, if needed, through either active or passive correction means. These are methods discussed in Section 4.12.1.

4.8.2 Jitter and Speed Stability

Velocity errors from a polygonal scanner are important to minimize because they affect the pixel placement in a writing application and the receiving angle in a reading application. Velocity errors have both repeatable and nonrepeatable components. The repeatable components are easier to deal with than the nonrepeatable errors.

Specifications for velocity errors are broken into both high-frequency (jitter) and low-frequency (speed stability) components. The high-frequency components range from pixel-to-pixel to once per revolution. The low-frequency components are over multiple revolutions. There are many elements of the scanner system that contribute to either jitter or speed stability errors. These contributing elements are shown in Table 4.1.

This is a long, but certainly not exhaustive, list of causative elements contributing or potentially contributing to velocity errors. It becomes obvious that the entire scanner optical system is involved and influences the speed stability measurement and result.

4.8.3 Balance

Polygonal scanners are rotational devices that can operate at high speeds. As such, they need to be properly balanced to reduce the amount of imbalance forces generated during operation. This includes compensating for both static and dynamic imbalance. This requires that a two-plane balancing system be used. In a two-plane balancing system sensors are located at two separated planes where correction weights are to be applied. The sensors record the magnitude and phase of the imbalance.

Various methods of either adding or removing weight are used to balance scanners. The most common techniques are:

1. Drill balancing
2. Epoxy balancing
3. Screw balancing
4. Grind balancing

TABLE 4.1

Elements Contributing to Jitter or Speed Stability
Errors

Primary causes

- Optical system

Fixed geometric errors of the scan lens

- Electronic driver stability

Frequency and phase stability

Voltage stability

Noise

- Motor characteristics

AC motor hunting (low frequency)

Cogging (high frequency)

- Bearing behavior

Varying resistance torque from lube migration

Roughness from wear and/or dirt

Bearing preload

- Polygonal mirror characteristics

Flatness

Facet radius variation (distance from center of rotation)

- Environmental (external shocks and vibrations)
- Encoder errors—sine wave errors due to disk centering

Secondary causes

- Reflectance uniformity
- SOS detector/amplifier noise
- Facet (polygon) surface roughness
- Air turbulence in the optical path (high-speed systems)
- Polygon/motor tracking accuracy
- Laser pointing errors (dynamic)

The preferred approach for high-speed operation is either grinding or drilling to remove material. The addition of material always brings risk of improper attachment and slinging of bonding agents.

Unbalance is typically measured in mg mm, a mass multiplied by the distance from the rotational axis. An unbalance of 100 mg mm, for example, indicates one side of the rotor has excess mass equivalent to 100 mg at a 1-mm radius. Typical values for small high-speed scanners range from 10 to 100 mg mm. The impact of unbalance on a scanner is vibration. This vibration can be measured and from this the actual scanner unbalance can be calculated.

4.8.4 Perpendicularity

Another important scanner parameter is the perpendicularity of the rotation axis to the mounting datum. This is important to ensure proper pointing of the beam after reflection from the polygon and to minimize the bow that can be created by striking the polygon out of the rotation plane. A typical specification for perpendicularity is 3 to 5 arc minutes.

4.8.5 Time to Synchronization

The time that it takes for the scanner to reach operating speed from a stopped condition can be important in some applications. This is a function of the motor/winding and the available current as well as the rotor inertia and the windage that must be overcome as the scanner approaches operating speed. Typical values range from 3 to 60 s.

4.9 SCANNER COST DRIVERS

Polygonal scanners can range from low-cost, easy-to-manufacture units, to high-cost state-of-the-art devices. It is important when designing a scan system to understand the cost drivers. One should try to minimize the overall cost through system level trade-offs. The scanner assembly has many cost drivers including:

1. Polygon shape
2. Number of facets
3. Fabrication method, conventionally polished or diamond turned
4. Optical specifications including surface figure, surface roughness, and scratch/dig
5. Coating requirements
6. Polygon size
7. Type of bearing system
8. Speed
9. Velocity stability
10. Dynamic track specification

In an earlier section the various shapes of polygons were discussed. In order to reduce costs it is advisable when possible to select either a regular polygon or a monogon. The other polygon shapes have cost penalties that may or may not be justified based on the application. While polygons can be manufactured with any number of facets, fewer facets result in lower cost. This is not a large cost component in a diamond turned mirror but has a large impact on the cost of a polished mirror.

The selection of diamond turned or polished mirror has a major impact on scanner cost. Diamond turned mirrors are the lowest cost and have surface roughness values greater than 40 Å rms. Conventionally polished mirrors are more costly but can bring the surface roughness down to 10 Å rms. All but the most scatter-sensitive short-wavelength systems can use diamond turned mirrors.

The optical specification of surface figure can also have a large influence on cost. Optical surface figure values of $\lambda/4$ per inch at 633 nm are common but surface figure values down to $\lambda/20$ can be achieved at additional cost. A scratch/dig specification of 80/50 is a typical standard, but specifications down to 10/5 can be achieved at significantly higher cost.

The optical coating chosen for the polygon can have a minor impact on the cost. The lowest cost option is a simple aluminum coating with a silicon monoxide overcoat. As the reflectivity specifications get higher, more dielectric layers are needed to enhance the reflectivity, which can increase chamber time and therefore costs. Gold is another

expensive coating option for the infrared. The inherent high reflectivity of gold across the IR spectrum is often worth the cost of the material.

Bearing selection can have a significant effect on cost. In the speed range of 500–4000 rpm, the choice is between ball bearings and aerostatic air bearings. Ball bearing scanners are relatively low in cost and are the appropriate solution for many applications, but are susceptible to damage, generate many vibration frequencies, and can create motor speed instability. Aerostatic scanners are costly and require support equipment, but offer the ultimate in scanning performance.

The bearing choice in the speed range of 4000–20,000 rpm includes ball bearings, aerodynamic air bearings, and aerostatic air bearings. The selection is based on cost and performance criteria such as velocity stability and dynamic track. Above 20,000 rpm, aerodynamic air bearings are usually the best solution. These bearings are relatively low in cost and have long life operating at this speed. Ball bearings start to have life issues above 20,000 rpm and aerostatic bearings usually are not cost effective.

Velocity stability standard specifications are a function of speed and mirror load. If speeds are too low or mirror loads too small then an encoder is required to achieve tight velocity stability. Velocity stability in this context is a measurement of the variation in the time for a beam reflected from the same facet of a scanner to cross two stationary detectors in an image plane over 500–1000 revolutions. Scanners operating faster than 4000 rpm can easily achieve 0.02% velocity stability. On most units this can be improved upon down to 0.002% at additional cost. Below 4000 rpm the mirror load becomes very important. The lighter the mirror and slower the speed, the more difficult it is to achieve tight velocity stability.

A final significant cost driver is the track specification placed on the assembly. Mechanical track values of 45 arc s results in low-cost assemblies, but specifications as tight as 1 arc s can be achieved by some vendors at much higher costs. This specification is a serious cost driver, so it is recommended that you review your actual needs carefully to obtain the most cost-effective design.

4.10 SYSTEM DESIGN CONSIDERATIONS

Laser scanning systems based on polygon technology can take on a variety of forms. Systems can range from very simple to extremely complex based on the performance level required. The first system consideration is whether it will be a reading or writing system. Writing systems tend to have much tighter performance requirements than reading systems. This is due to the fact that writing system errors tend to be visible, whereas the same level of error in a reading system will not typically be great enough to impact data integrity. A reading system, however, has the additional complexity of collecting the scattered light back from the target.

Reading systems will either use an external collection system that is separate from the scan system or an internal collection system where the scattered light passes back through the scan system and is derotated by the polygonal scanner. The internal collection system places increased demands on the scan system by requiring less backscattered light and reduced ghost images. For laser radar systems, one often has a scanning system for transmitting the laser, and a separate receiver, with a synchronized scanner to avoid this problem. This, however, is a very expensive solution. Another approach taken with laser

radar is to increase the facet width and separate the transmission and receive apertures. Care must be taken in the design to ensure that the receiver instantaneous field of view encompasses the transmitter output over the distance range desired. This topic is covered in greater detail in the chapter on laser radar later in this book.

Beyond having knowledge of the basic system configuration it is important to develop a thorough list of performance specifications when starting the system design process. A list of key parameters and typical values are shown in Table 4.2.

The list in Table 4.2 covers the majority of specifications that are placed on a scanning system. Some scan systems will require additional specifications based on the unique nature of the writing or reading application.

The optical system used in laser scanners can be separated into two generic types: pre-objective and post-objective. Pre-objective is a term used to describe the use of a polygon to deflect a ray bundle, which after deflection is imaged by a lens or curved mirror (Figure 4.18). This method of scanning places the function of focal plane definition on the lens, referred to as a scan lens, rather than on the scanning facet. Several desirable characteristics can be designed into the scan lens when employed in pre-objective scanning. An example is a lens design referred to as F-Theta. An F-Theta lens has the following characteristics:

1. A flat focal plane
2. Uniform spot diameter over the entire scan
3. Linear spot velocity at the scan plane (assuming constant angular velocity of the polygon)

Usually it is desirable to have the scanning spot move with a highly accurate and constant velocity in the scan plane. Polygonal mirror deflectors provide angular velocity stability in the range 0.002%–0.05%, depending on the speed and inertia of the scanner. Without the aid of an F-Theta lens, however, the spot velocity variation on a flat focal surface will be

TABLE 4.2

List of Key Parameters

Wavelength	350–10,600 nm
Number of resolvable points	100–50,000
Spot size	1 micron–25 mm
Spot size variation across scan	≤5%–15%
Scan length	1 mm–2 m
Telecentricity	0.5–30°
Bow	≤0.001% of scan line length
Scan efficiency	30%–90%
Intensity nonuniformity	≤2% to ≤10%
Pixel placement accuracy	
• Jitter	≤0.002% to ≤0.02%
• Cross-scan error	≤1% to ≤25% of line spacing
Scatter	≤0.2% to ≤5%
Data rate	
Laser noise levels	
Environmental factors and system interfaces	
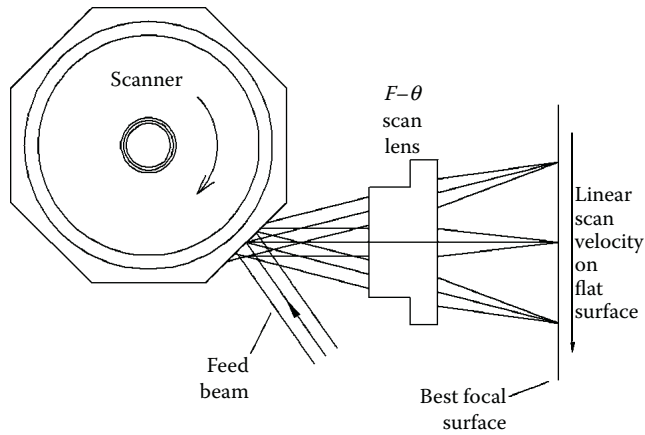
FIGURE 4.18
Pre-objective scanning system.
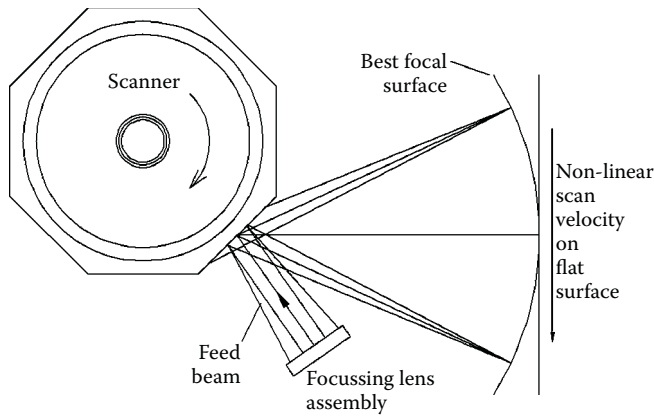
FIGURE 4.19
Post-objective scanning system.

proportional to the tangent of the scan angle, which for systems involving several degrees of scan means several percent variation.

Post-objective scanning is a term used to describe the use of a polygon to deflect a focusing ray bundle over a focal surface (Figure 4.19). This method places the function of focal plane definition on the polygonal mirror, and the imaging (spot forming) lens is a relatively simple component located prior to the polygon.

The focal surface of a post-objective scanner is curved. The center of curvature is the center of the polygon facet. This type of scan system is typically used when the scan plane can be curved to match the focal surface. Otherwise there are problems with spot size and velocity variations across the scan. The most popular system design incorporating post-objective scanning is a drum scanner. A drum scanner uses a monogon mirror, usually at 45°, with the source on the scan axis. As the monogon rotates, the focal surface is generated on the inside of a drum. Film or other flexible medium is located on this drum for image generation.

Post-objective scanning finds application in very high-resolution systems requiring greater than 25,000 points across the scan. Scanners designed for the prepress industry use this design technique quite often.

Another factor to consider when designing a scan system is the degree of telecentricity required. A system is considered to be telecentric if the output from the scan system strikes the image plane at 90° for all points across the scan line. A post-objective scanner can be telecentric if the image plane can be curved to intercept the output from the scanner. If a flat image plane is required, a pre-objective scan system will need to have a scan lens that is slightly larger than the scan plane to meet the telecentric requirement. This can drive up the scan lens costs and result in a prohibitively expensive system. Normally, some level of deviation from telecentricity is given in a system specification.

In a writing application a decision as to how to use the available polygon facet is needed. Systems can either be under- or overfilled. Underfilled designs are the most common and do not waste available laser energy because the facet is sized such that the beam footprint on the facet never crosses over the edges of the facet during the full system scan angle. On the other hand, in an overfilled design the polygon facet is sized such that the beam completely fills the polygon facet over the entire full scan angle. Underfilled designs are preferred in many applications because there is less wasted energy and there is minimal diffraction from the facet edges. Overfilled designs have the one advantage that the system duty cycle can approach 100%. The duty cycle is the ratio of the active scan time to the full facet time.

4.11 POLYGON SIZE CALCULATION

Once a system concept is chosen, and the optical design completed, the polygon size needs to be calculated. A few key parameters must be known in order to size the polygon:

1. Scan angle, θ
2. Beam feed angle, α
3. Wavelength, λ
4. Desired duty cycle, C

θ is the full extent of the active scan measured in degrees as illustrated in Figure 4.20. This value is usually in the range of 5–70°. α is the beam feed angle measured in degrees between the input beam to the polygon and the center of the scan exiting the polygon. It will be cost effective to keep this angle as small as possible in order to reduce polygon size. In certain scanner applications the beam feed angle is zero. The beam is brought in through a beamsplitter in the center of scan or at a slight angle relative to the exiting scanned beam. λ is the operating wavelength expressed in microns and to be used in the calculation of the beam size on the polygon with a known desired spot size in the scan plane. C is the duty cycle, which is the ratio of active scan time to total time. Duty cycles in the range of 30%–90% are common. However, the greater the duty cycle, the larger and more costly the polygon. With all conventional scan systems with the exception of monogon drum scanners some portion of the time will be spent transitioning from one facet to the next. We will assume that the design being considered is underfilled. This means that only one facet is being used to scan the image plane at any given time.
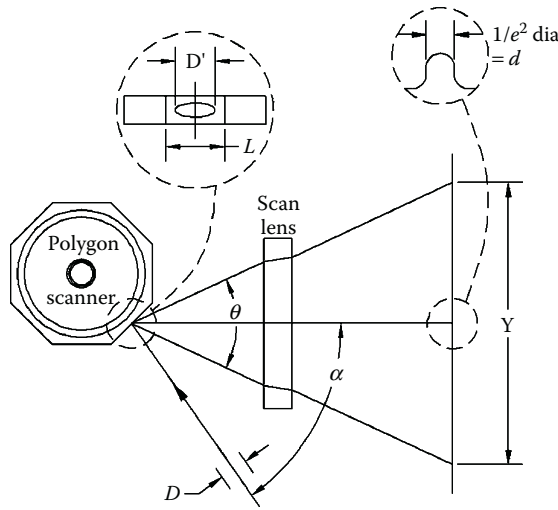
FIGURE 4.20
Illustration of scan angles.

The number of facets, *n*, to be used is a trade-off that needs to be addressed. The formula for the number of facets is given by:

$$n = \frac{720C}{q} \tag{4.3}$$

If this equation produces a noninteger answer, this means that there is no exact solution to provide the duty cycle desired at the same time as the optical scan angle requirement is satisfied. A next logical step is to fix the number of facets to an integer value near the result from the previous calculation and fix either the scan angle or the duty cycle and solve for the remaining variable.

$$C = \frac{nq}{720} \tag{4.4}$$

For a writing application, once the duty cycle, scan angle, and number of facets are determined, the beam diameter *D* incident to the facet can be calculated. The following formulas assume a Gaussian beam profile and the beam size defined at the $1/e^2$ intensity points.

$$D(\text{mm}) = \frac{1.27 \, 1F}{d} \tag{4.5}$$

where *F* is the focal length of the scan lens in mm, and *d* is the $1/e^2$ beam diameter in the scan plane in microns.

The polygon can be sized without a scan lens by using the following formula:

$$D(\text{mm}) = \frac{1.27 \, 1T}{d} \tag{4.6}$$

where T is the distance from the polygon to the focal surface in mm, and d is the $1/e^2$ beam diameter on the focal surface in microns.

For a reading system, D is a selected value based on the system-limiting aperture. The intensity profile across the diameter is no longer Gaussian but top hat instead.

Since the size of the facet depends on the actual beam footprint on the facet, the feed angle effect on D must be taken into account. The value D' is the projected footprint on the polygon facet. It takes into account the truncation diameter and the cosine growth of the beam on the facet due to the beam feed angle. The formula for calculating the beam footprint is:

$$D' = \frac{1.5D}{\cos(a/2)} \tag{4.7}$$

The calculations assume a TEM00 Gaussian beam that is truncated at the $1.5 \times 1/e^2$ diameter. If the application can tolerate more clipping at the start and end of scan the polygon size can be reduced.

The length of the facet (L) can be approximated from the beam footprint using the following:[3]

$$L(\text{mm}) = \frac{D'}{1 - C} \tag{4.8}$$

The polygon diameter can now be approximated as follows:

$$\text{Diam}_{\text{inscribed}} \oplus \frac{L}{\tan(180/n)} \tag{4.9}$$

If the polygon diameter is too large then there are three options. The first is to reduce the duty cycle and suffer a higher speed and burst data rate. The second is to reduce the beam feed angle. The third is to allow more intensity variation across the scan by reducing the 1.5 multiplier. This in turn, reduces the facet length.

4.12 MINIMIZING IMAGE DEFECTS IN SCANNING SYSTEMS

In order to design a scanning system that accurately reproduces information, knowledge of the types of artifacts that the scan system can produce and visibility thresholds of these artifacts is needed. The specifications required to reduce the artifacts to acceptable levels vary by application; for example, a prepress imager has different requirements from a laser printer. This section is written with writing applications in mind. Many of the defects associated with writing applications can also be present in reading applications.

4.12.1 Banding

Banding is one of the most common scan artifacts that will show up in scanning systems. Banding is a periodic variation in the line-to-line separation or intensity of the output. The human eye is very sensitive to periodic errors. The sensitivity is frequency dependent and great care must be taken to ensure that scan errors in the peak frequency range are minimized.[4] In continuous tone and halftone printers the line-to-line placement errors need to be reduced to less than 0.5% of the line spacing. In other applications this can be as large as 10%–20% before banding becomes visible.

The sources of banding include reflectivity variations, dynamic track errors between facets, mechanical vibrations, electrical noise and secondary axis translation errors. Polygon reflectivity variations can be easily eliminated by properly specifying the polygon coating such that these errors will not be visible. A specification of less than 1% variation of reflectivity on all facets will be adequate for all but the most demanding applications.

Either improving the polygon itself or compensating for the error can reduce dynamic track errors. Either approach will increase system costs so the trade-off between the brute force approach of improving the polygonal scanner must be weighed against the costs of additional system complexity.

Dynamic track error compensation can either be active or passive. Active techniques rely on using an active component to move the beam to compensate for the error whereas passive techniques rely on optics to minimize the errors. Active correction techniques will compensate for repeatable errors, but not errors that vary throughout the scan line. Passive techniques will compensate for both repeatable and nonrepeatable errors.

Active correction techniques are usually based on sampling the beam position errors perpendicular to the scan direction (cross scan) between scans and applying a beam steering correction in the system prior to the polygon to change the beam pointing. These techniques are used primarily in low-speed systems due to the frequency response limitations of the beam steering components. Active correction systems are rare because there is added mechanical complexity, higher cost, and the lack of correction for changes that occur during scan.

Passive correction techniques are quite common and the basic concept is illustrated in Figure 4.21. The polygon facet is reimaged with some magnification to the scan plane in the cross-scan axis. A cylindrical lens element is typically used to create a line focus on the polygon facet. The reimaging of this line in the cross-scan axis can be accomplished using a variety of components. Common methods include using a toroidal element near the polygon, a cylindrical lens near the scan plane, or a cylindrical mirror near the scan plane.[5,6] These passive methods will provide a significant reduction, but not perfect compensation due to pupil shifting. The pupil shift is due to the fact that the polygon rotates about its center rather than rotating about the facet. The facet vertex changes during rotation so the object point moves in and out as the facet rotates.

Banding does not necessarily result from optical effects. Other sources such as vibration or electrical noise can contribute to banding. Mechanical vibrations can be introduced by the rotating device and amplified by the scan system platform. If the platform is not rigidly coupled to the image plane, then relative motion between the scanner and the image can result in a banding artifact.
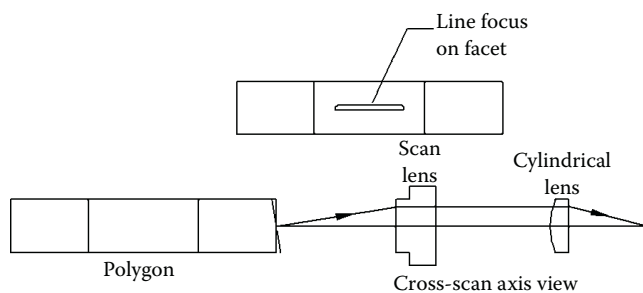
FIGURE 4.21
Passive cross-scan correction.

Electrical noise can be generated by lasers or laser power supplies. It can modulate the laser output directly or it can affect the performance of an external modulation device such as an acousto-optic modulator. Continuous tone applications can be particularly sensitive to electrical noise. Repeatable noise on the order of 0.5% of peak power can be visible. The electrical noise can appear to be banding if the frequency is near to or a multiple of the revolution rate.

In most scan systems the second axis of the image is controlled by a mechanical device, such as a translation stage, direct drive rollers, or a belt drive. The velocity stability of this second axis must be specified to the same level of requirements as the scanner device. Velocity errors in this second axis directly impact the banding in the image.

This section has shown that there are a variety of sources of banding. Care must be taken in the design phase to properly specify all components that contribute to this problem since it can be difficult to isolate the root cause when this defect appears in a scan system.

4.12.2 Jitter

Jitter is the high-frequency variation in the pixel placement along the scan direction. Various systems can tolerate different levels of jitter before artifacts become visible. Output scanners that place a premium on pixel placement will typically require 0.1-pixel accuracy whereas visual image outputs can tolerate up to 1 pixel in many applications. Jitter has both random and repeatable components. Random jitter is visually less objectionable than periodic jitter.

Random jitter errors can be produced by the ball bearings used in most low-speed scanning systems. The magnitude of these errors is dependent on the inertia of the rotor, the ball bearings chosen, and the bearing mounting method. Errors are usually small enough not to be of concern. Aerostatic air bearings offer an alternative if the system is sensitive to the ball bearing errors.

Motor cogging with brushless DC motors can also create jitter errors that repeat once per revolution. The motor controller can reduce these errors with proper feedback rates (encoders or start of scan feedback), but they cannot be eliminated. One method to overcome these errors in low-speed applications is to retime the output data, based on actual scanner position information provided by an encoder. The only way to eliminate these errors is to find a motor with zero cogging torque. There is a class of toothless motors that have close to zero cogging torque. They are expensive, but may become more affordable as they further penetrate the market.

Polygon facet flatness variations result in a periodic jitter with a frequency of once per revolution or higher. The curvature causes small deviations in the angle of reflection from the facet. If the curvature of each facet varies, this causes the time between start of scan and end of scan to vary. A special case exists where there is no contribution to jitter if all facets have the same curvature. A facet flatness specification on the order of $\lambda/8$ is adequate for most applications.

Facet radius variations in systems using post-objective scanning result in beam displacement in the scan plane.[7] A facet radius variation specification of less than 25 microns is acceptable for most scanning applications.

4.12.3 Scatter and Ghost Images

There are many sources of scattered light and ghost images in an optical system. The majority of ghost images can be controlled through proper coatings and the placement of baffles. For example, if the strays are out of the plane of the scanned image an exit slit does

a good job of eliminating them. Scan lenses can create problems with ghost images. The interior surfaces in these lens systems set up ghost images that are difficult if not impossible to eliminate with baffles. The antireflection coatings need to be high quality, reducing reflections to near zero.

As mentioned earlier in the chapter, the polygon surface can contribute to scatter. In extremely sensitive applications a diamond turned surface may produce too much scatter. This type of surface also produces a large percentage of near angle scatter that is difficult to baffle. A conventionally polished polygon produces wide-angle scatter and a much lower magnitude of total integrated scatter.

If the system contains an exit window or has a lens close to the scan plane, then the cleanliness of this element is important. Dust particles on this element can cause localized scatter in the scan plane since the spot is typically small at this point in the optical system. If repeated each scan, this will result in a line being produced down the image.

Adjacent polygon facets tend to be problematic in many systems where there is significant reflection from the target surface. The beam can find its way back through the system to the next facet. The problem with this type of stray light is that it will be on axis. The best solution is to tilt the scan plane a few degrees relative to the scan system so scan plane reflections are out of plane. Another solution is to mask the polygon sufficiently to leave only the active scan aperture open.

The time between scans when the beam is passing over the tips of the polygon is another source for scatter. Light will scatter from the tips of the polygon and from the side of scan lens mounts. Turning off the beam between scans and using a time interval counter to turn the beam on just prior to the start of scan sensor will eliminate this possible source of problems.

Acousto-optic modulators can produce several undesirable effects. Scatter from the crystal can limit the extinction ratio. Long decay times may result in tails when transitioning from black to gray in a continuous tone application. The crystals used can also suffer from sound field reflections that show up as ghost images. Working with an application engineer at the modulator supplier is the best way to avoid these issues from affecting a scan system.

4.12.4 Intensity Variation

Variations in laser intensity can produce a variety of image artifacts depending on the frequency of the variation. A slowly varying fluctuation is much less objectionable than a high-frequency variation. Whereas intensity variation on the order of a few percent may be tolerable over an entire image, local intensity variations may need to be controlled to less than 0.5%.

Scan lens coatings and the coatings on any other elements located after the polygon can cause variations in the scan plane intensity across the scan. A transmission or reflection uniformity specification is needed to control this variable.

4.12.5 Distortion

Scanning systems typically employ a scan lens that has an F-theta characteristic. The lens distortion is controlled to produce image height that is proportional to the scan angle. This F-theta characteristic ensures linear scans with constant velocity. These lenses are not perfect but they do reduce the nonlinearity down to 0.01%–0.1% range. This is adequate for all but the most critical applications. These residual errors are repeatable, therefore; intensity compensation for dwell time differences or variable clocking schemes for pixel placement differences can be employed to remove the residual error.

4.12.6 Bow

Bow is defined as the variation from straightness of a scan line. Bow is usually a slowly varying function across the scan. A considerable amount of bow can be tolerated before becoming visually objectionable. In most applications 0.05% of the scan line length of bow is an adequate specification. If the system is designed to have the beam brought in on the same axis as the scan then bow is caused by the errors in beam alignment. This can normally be adjusted to very fine levels and is therefore usually not a serious problem in a polygonal scan system. An equation for bow is given by:

$$E = F \sin b \left[\frac{1}{\cos q} - 1 \right] \tag{4.10}$$

where F is the focal length of the scan lens, E is the spot displacement as a function of field angle, θ is the field angle, and β is the angle between the incoming beam and the plane that is perpendicular to the rotation axis.[8]

4.13 SUMMARY

This chapter has covered the components, performance characteristics, and design approaches for polygonal scanners and systems based on these scanners. This technology continues to evolve and thrives among increasing competition from other technologies both in writing and reading applications. I fully expect that the performance values that are stated in this chapter will be significantly improved on in the near future. However, the system level artifacts that a system designer must be careful to avoid tend to remain a constant. An in-depth knowledge of these artifacts and their root causes will help reduce development time for new systems.

ACKNOWLEDGMENTS

REFERENCES

1. Oberg, E. *Machinery's Handbook,* 23rd Ed; Industrial Press: New York, 1988; 196.
2. Bennett, J.M.; Mattsson, L. *Introduction to Surface Roughness and Scattering;* Optical Society of America: Washington, DC, 1989; 50–52.
3. Beiser, L. Design equations for a polygon laser scanner. In *Beam Deflection and Scanning Technologies*; Marshall, G.F., Beiser, L., Eds; *Proc. SPIE* 1454; 1991; 60–65.

4. Bestenreiner, F.; Greis, U.; Helmberger, J.; Stadler, K. Visibility and corrections of periodic interference structures in line-by-line recorded images. *J. Appl. Phot. Eng.* 1976, 2, 86–92.

5. Fleischer, J. Light Scanning and Printing Systems. US Patent 3,750,189, July 1973.

6. Brueggemann, H. Scanner with reflective pyramid error compensation. US Patent 4,247,160, January 1981.

7. Horikawa, H.; Sugisaki, I.; Tashiro, M. Relationship between fluctuation in mirror radius (within polygon) and the jitter. In *Beam Deflection and Scanning Technologies*; Marshall, G.F., Beiser, L., Eds; *Proc. SPIE* 1454; 1991; 46–59.

8. Hopkins, R.; Stephenson, D. Optical systems for laser scanners. In *Optical Scanning*; Marcel Dekker: New York, 1991; 46.

5

Motors and Controllers (Drivers) for High-Performance Polygonal Scanners

Emery Erdelyi
Axsys Technologies, Inc.
San Diego, California, USA

Gerald A. Rynkowski
Axsys Technologies, Inc.
Rochester Hills, Michigan, USA

CONTENTS

5.1 INTRODUCTION

This chapter updates and expands upon the material covered by Gerald A. Rynkowski in *Optical Scanning,*[1] with greater emphasis being placed on brushless DC motors and the associated control electronics designed specifically for rotary scanning applications. Some background topics related to polygon scanning have been carried over to this chapter in order to clarify or illustrate control concepts. Other topics not directly related to motors and controllers, such as the discussion of air bearing design, have been omitted since these areas are covered in greater detail elsewhere in this book.

The availability of low-cost brushless DC motors and the continuing improvement in scanner control, as well as miniaturization of the drive electronics, have contributed greatly to the viability of optomechanical scanning in many new applications. Optomechanical scanning continues to be a cost-effective alternative to competing solid-state technologies. Several new application examples have been added that highlight the trend toward brushless DC motors and compact integrated control systems designed for military and commercial use.

Polygonal scanners have been designed, developed, and manufactured in all shapes and configurations during the past 35 years. These devices have been employed in military reconnaissance and earth resources studies, thermal imaging systems, film recorders, laser printers, flight simulators, and optical inspection systems, to name a few of the well-known applications. A common characteristic of all of these scanners is the requirement for precise control of polygon rotation, and consequently, the control of the beam scan. Recent advances in motor and control technologies have greatly improved the performance and efficiency of these scanners while simultaneously reducing the cost and size of the system.

This chapter explores the trends in motor and control technologies that are being utilized in today's polygonal scanners through the discussion of specific applications. In addition, motor characteristics, control techniques, and system models are presented to aid the optomechanical engineer in understanding these critical areas of scanning system design.

5.2 POLYGONAL SCANNER BASICS

Although a more thorough discussion of polygon geometry and scanning optics can be found elsewhere in this book, it will be useful to review some of the basic polygon configurations and optical designs as they influence the selection of the scan motor and control system. Also, a film recording system is presented in some detail in the next section in order to illustrate the influence of the scanner motor and controller characteristics on the overall system performance.

5.2.1 Polygon Configurations

In general, three types of scanner mirror configurations are popularly utilized in collimated or convergent, passive, or laser scanning optical systems. These rotating mirror spinners (polygons) are at the center of the electro-optical system. They direct incoming optical signals to a detector or steer outgoing modulated laser beams by virtue of their geometry and rotation about an axis.

The three most utilized scanner beam deflector configurations are the regular polygon, pyramidal, and the single-faceted cantilever design. The regular polygonal scanner is generally the most popular with system designers and can be utilized with either the collimated-beam or the convergent-beam scanning configurations. Figures 5.1 and 5.2 illustrate the two configurations using six-sided polygons having the spin axis projected into the page. In Figure 5.1 the facets are illuminated with a collimated beam and reflected to a concave mirror that focuses at a curved focal plane.

Figure 5.2 shows a lens system that focuses the collimated beam prior to being reflected at the scanner facets, and then converging at a focus. Comparing the two configurations, it is obvious that both focal image surfaces are curved. This presents a problem to the system designer, but is usually corrected optically with a suitable field-flattening correction lens system, or perhaps the recording surface is curved to conform to the focal surface.
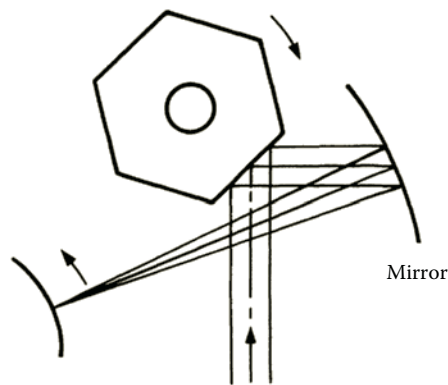
FIGURE 5.1
Collimated beam scanning. (From Speedring Systems Group. "Ultra precise bearings for high speed use," 102–1; "Gas bearing design considerations," 102–2; "Rotating mirror scanners," 101–1, 101–2, 101–3. Technical Bulletins: Rochester Hills, MI. With permission.)
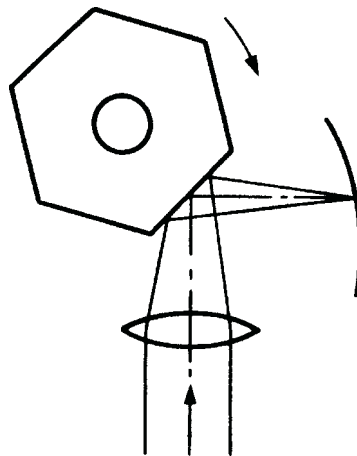
FIGURE 5.2
Convergent beam scanning. (From Marshall, G.F.; Rynkowski, G.A. Eds. *Optical Scanning*; Marcel Dekker, Inc.: New York, 1991. With permission.)

Another difference with regard to the polygon is that the convergent-beam bundle (Figure 5.2) uses less area of the facet than the collimated configuration. However, maximum utilization of the facet area is desirable because the facet surface flatness irregularities tend to be averaged out, and therefore, minimize modulation of the exiting-beam scanning angles. Most precision polygonal scanning systems use the collimated beam scanning configuration, which utilizes a larger proportion of the area of the facets.

Figure 5.3 illustrates a pyramidal mirror scanner commonly used with the rotational and optical axes parallel, but not coincident. Note that either the collimated or convergent configuration can be utilized with this design.

Figure 5.4 illustrates a regular polygonal scanner in which the optical axis is normal to the rotational axis, or where the angle is acute to normal. Note that when the two axes, optical and rotational, are normal, the beam can reflect back upon itself.

Shown in Figure 5.5 is a single-faceted cantilevered scanner with the beam and rotational axes coincident and reflecting from a 45° facet, and thereby generating a continuous 360° scan angle and a circular focused scan line.

This configuration has been used in passive infrared scanning systems having long focal length and requiring a large aperture. Nine-inch, clear-aperture scanners have been manufactured in this configuration for high collection efficiency.

This type of beam deflector design is also popular in scanning systems that are used in the image setting machines sold to the printing market. Many of these image setting machines have an "internal drum" design which involves the placement of a large piece of film on the inside surface of a cylinder. The single-faceted "monogon" beam deflector rotates at high speed and scans a laser spot across the width of the film as it travels the length of the drum. Figure 5.6 illustrates a monogon beam deflector designed for rotational speeds exceeding 30,000 rpm.
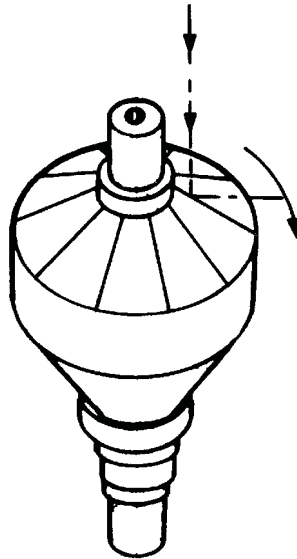
FIGURE 5.3
Parallel, but not coincident, optical and rotational axes. (From Speedring Systems Group. "Ultra precise bearings for high speed use," 102–1; "Gas bearing design considerations," 102–2; "Rotating mirror scanners," 101–1, 101–2, 101–3. Technical Bulletins: Rochester Hills, MI. With permission.)
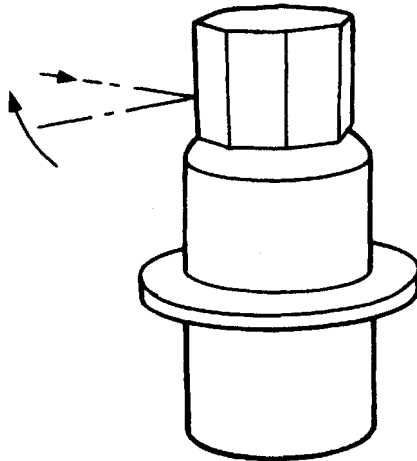
FIGURE 5.4
Optical and rotational axes normal. (From Marshall, G.F.; Rynkowski, G.A. Eds. *Optical Scanning*; Marcel Dekker, Inc.: New York, 1991. With permission.)

Detector
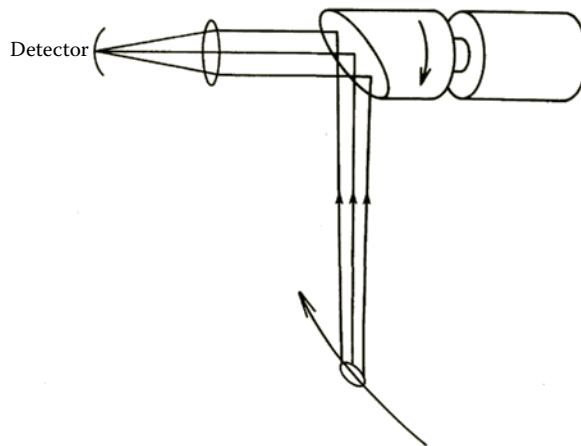
FIGURE 5.5
Parallel and coincident optical and rotational axis. (From Marshall, G.F.; Rynkowski, G.A. Eds. *Optical Scanning*; Marcel Dekker, Inc.: New York, 1991. With permission.)

5.2.2 Polygon Rotation and Scan Angle Relationship

At this point, it is noteworthy to realize the relation between the facet angle and scan angle. The facet angle is defined as $360°/N$ (N = number of facets). The optical scan angle may be expressed as:

$$\text{Scan angle (degrees)} = \frac{720}{N}$$

for $N \geq 2$. Observe that for $N \geq 2$, the optical scan angle is two times the shaft angle. This angle-doubling effect must obviously be considered when relating the shaft and facet parameters and their effects on the angular position of the focused spot at the focal plane.
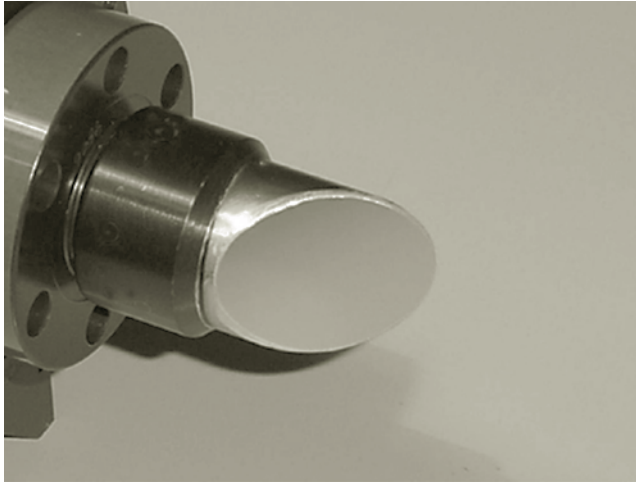
FIGURE 5.6
Monogon beam deflector.

The optical angle-doubling effect places an even greater demand on the control of the polygon rotational velocity and must be carefully considered if the desired system accuracy is to be achieved.

The scanner rotational speed may be expressed as:

$$rpm = \frac{60W}{N}$$

where W = line scans/s, and N = number of facets.

An increase in the number of facets reduces the motor speed requirements as well as the maximum scan angle. However, the usable scan angle may in some cases also be limited by aperture size and the allowable vignette effect. In practice, the optical design will usually dictate the number of facets and consequently the motor and controller will have to be selected to accommodate the optical system designer. Figure 5.7 illustrates a small 12-faceted polygon mirror.

5.2.3 Polygon Speed Considerations

When a range of polygon rotation speeds are allowed by the optical design, it is best to avoid configurations that require speeds that are very low or very high. Problems with motor controllability may occur at speeds below about 60 rpm and motor efficiency may suffer when speeds exceeding 60,000 rpm are specified. To achieve good speed regulation at very low polygon rotation speeds, special motor and control designs are often required that will lead to increased system cost. A polygon rotating at 60 rpm and specified for l0 ppm (10 parts per million, or 0.001%) speed regulation will likely require a sinusoidally driven slotless brushless DC motor and complex controller design to achieve this level of performance.

Another complication associated with low-speed designs involves the selection of the velocity feedback device. Often the feedback device will be an optical encoder (Figure 5.8),
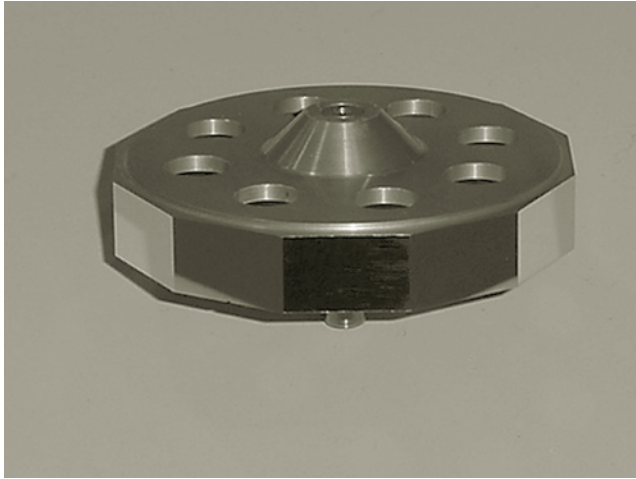
FIGURE 5.7
Twelve-faceted polygon.
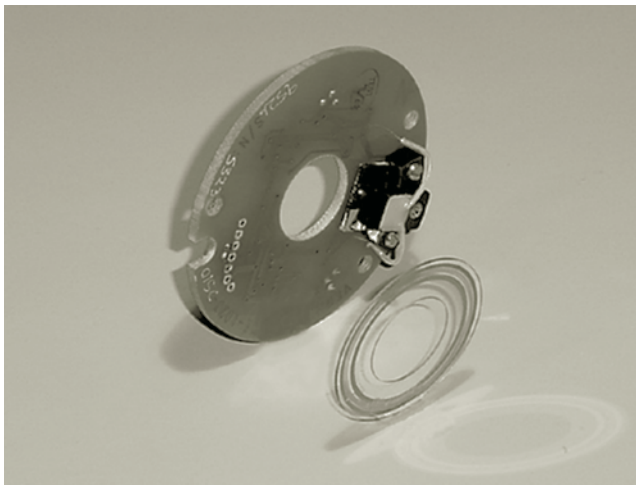
FIGURE 5.8
Optical encoder, disc, and readout electronics.

which functions both as a tachometer to monitor the polygon speed, and in some systems, as a position sensor for reporting the true position of the polygon to the scan processing electronics. Today's high-performance speed control systems operate using phase-lock loop techniques that provide excellent short-term as well as long-term speed regulation. These systems operate by comparing the frequency and phase of a stable reference signal with that of the polygon encoder, thereby generating an error signal, which is used to adjust the polygon speed. In order to achieve good speed control at low speeds, a high-resolution/high-accuracy encoder is necessary.

Depending on the polygon inertia, bearing friction, and the level of disturbances present, an encoder line density greater than 10,000 lines (counts) per revolution may be required for a 60 rpm scanner. Optical encoders having greater than a few thousand line counts per

revolution will also contribute to increased system cost since the disc pattern and the edge detection functions must be more precise.

As the polygon operating speed increases, fewer pulses per revolution from the encoder are required to achieve the same level of performance, all other factors being equal. This is due to the fact that at higher speeds, lower encoder resolutions can still produce an adequate number of pulses or "speed updates" per second from the encoder and allow for a reasonably fast control loop bandwidth. Higher control system bandwidth is desirable because the control loop can more readily react to short duration disturbances in speed, which may not be adequately attenuated by the polygon inertia.

The control system will receive some average polygon speed between encoder pulses from the phase detector, but it is essentially operating open loop until the next pulse arrives and updates the speed based on the encoder pulse phase relative to the reference clock. During this period between speed updates the polygon speed will generally decrease from the set point (and the last update) until the next encoder pulse arrives and the control system makes a correction. Therefore to achieve a specified update rate or bandwidth for the control system, the slower polygon will require a higher resolution encoder with correspondingly better pattern accuracy.

The actual speed change between encoder updates is a function of many factors, including the operating speed, total rotating inertia of the scanner, the amplitude and frequency of disturbances, the control system gain, and the friction. In general, higher polygon/motor inertia and lower bearing and windage friction are beneficial and will reduce the short-term speed variations present in the scanner. This will be discussed in greater detail in Sec. 5, which deals with the control system design.

At the other extreme, high operating speeds require that the polygon and the entire rotating assembly exhibit exceptionally low imbalance, ideally less than 10 μin-oz. Often this requires that the scanner must be balanced in two planes: at the polygon and at the motor/encoder location. This is necessary not only for maintaining the mechanical integrity and life of the rotating components, but also for achieving precise speed control within one revolution of the scanner. Any imbalance that causes concentricity error at the encoder will result in a sinusoidal scanner speed variation. This is especially true for air bearing scanners where the bearing stiffness is lower and allows for larger concentricity errors to be produced.

Also, the motor efficiency is adversely affected at high speeds, especially when the commutation frequency exceeds 1000 cycles/s. Special motor designs are required for efficient high-speed operation and these designs usually require expensive development time and have higher unit cost.

5.3 CASE STUDY: A FILM RECORDING SYSTEM

A film recording scanning system has been selected as a reference subsystem for purposes of discussing the scanner parameters as well as the dynamic performance requirements involving the motor and control system design. Figure 5.9 depicts a laser recording system capable of recording high-resolution video or digital data on film. The rotating polygon (spinner) generates a line scan at the film plane using a focused, intensity-modulated laser beam. The laser exposes the film in proportion to the intensity modulation in the video or digital signal. Line-to-line scan is accomplished by moving the photographic
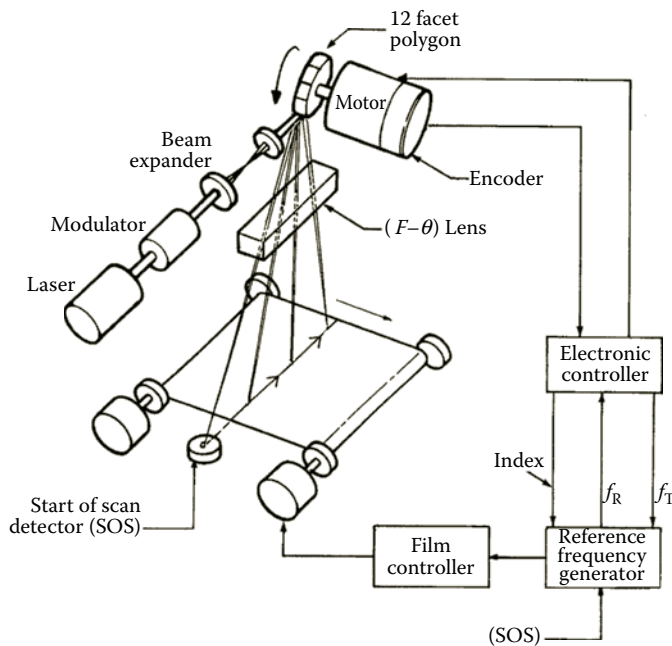
FIGURE 5.9
Film recording system. (From Marshall, G.F.; Rynkowski, G.A. Eds. *Optical Scanning*; Marcel Dekker, Inc.: New York, 1991. With permission.)

film at constant velocity and recording a continuous corridor of data limited only by the length of film. The film controller provides precise control of film velocity, which is locked to the polygon rotation. The expanded and collimated laser beam is intensity modulated with video or digital data and then scanned by the facets of the spinner to the film plane. A field correction lens (F–θ) is used to focus the beam, to linearize the scan line with respect to the scan angle, and thereby provide a uniform spot size along the line at the film plane.

The scanner assembly contains a 12-faceted polygon required to perform the optical scan function. The rotating mirror, its drive motor rotor, and a precision optical tachometer are supported by externally pressurized gas bearings. The electronic controller provides precise motor speed control and synchronization between the reference frequency sync generator and the high-density data track of the optical encoder. The encoder also supplies an index pulse used for facet identification and derivation of the synchronized field frequency that is required for some raster scanning systems, as well as pixel registration and control of the film drive motor.

5.3.1 System Performance Requirements

The system performance parameters for our example digital film recorder are discussed in the following paragraphs. These parameters are summarized in Table 5.1 and are typical and representative of a recently manufactured film recording system.

The system resolution requirements are defined at the film plane since all optical, scanning, and film transport errors will become evident on the recording media. The application requires 10,000 pixels of digital data per line at the scan plane using a 10-μm

TABLE 5.1

Film Recorder System Requirements

Line resolution (both directions)	10,000 pixels/line
Line scan length	13.97 cm
Scan rate	1200 lines/s
Film speed	1.6764 cm/s
Pixel frequency (clock)	12 MHz
Pixel diameter	10 μm at $1/e^2$
Pixel-to-pixel spacing	13.97 μm

Source: Marshall, G.F.; Rynkowski, G.A. Eds. *Optical Scanning*; Marcel Dekker, Inc.: New York, 1991.

spot diameter measured at the $1/e^2$ irradiance level. One can calculate that for a 13.97-cm line scan length the pixel spacing, center to center, must be 13.97 μm, as must the spacing between scan lines. The line spacing variance, or film transport jitter, is conservatively specified to be less than ±5 ppm (half the nominal spot size). For a reasonable throughput, the application dictates a scan rate requirement of 1200 lines/s. From this information the calculated pixel frequency is found to be 12 MHz (1200 lines/s × 10,000 pixels/line), and the film speed needed is 1.6764 cm/s (1200 lines/s × 13.97 μm/line).

5.3.2 Spinner Parameters

The polygon specification is determined and driven by the form, fit, and functional performances set by the optical requirements of the system. At this point, the optical engineer must optimize the design of the optical elements, which includes specifying the polygon type, facet number, facet width and height, inscribed diameter of the polygon, facet flatness and reflectance, and rotating speed. Owing to the interactions of the specifications and the high-accuracy requirements, the spinner is addressed as a scanner subsystem to allow the scanner designer to make trade-off decisions within the limits imposed by the optical and system performance requirements.

The scanner subsystem in our film recorder example consists of a one-piece beryllium scan mirror and shaft, suspended on hydrostatic gas bearings, and driven by a servo-controlled AC synchronous motor. The *F*–θ lens is designed to function with a 60° optical scan entrance angle which dictates a 30° facet angle specification for the polygon. The number of facets is therefore calculated to be 12, and a motor speed of 100 rev/s, or 6000 rpm, generates 1200 scan lines per second. Depicted in Table 5.2 is a summary of the scanner polygon requirements.

5.3.3 Scanner Specification Tolerances

Scanner specification tolerances are determined by the permissible static and dynamic pixel position errors acceptable at the film plane. These worst-case errors are referenced back through the optical system and scanner subsystem to be distributed and budgeted between the operational elements and reference datum. The acceptable variances are often specified as a percentage of pixel-to-pixel angle, pixel diameter, pixel-to-pixel spacing, or the motor speed regulation (stability) over one or more revolutions. The conversion of these variances to meaningful and quantifiable units is necessary for manufacturing, measurement, inspection, and testing of the scanning system components.

TABLE 5.2

Film Recorder Polygon Requirements

Number of facets	12
Facet angle	30°
Inscribed diameter	4.0 in
Facet height	0.5 in
Facet reflectance	89%–95%
Facet flatness	$\lambda/20$
Facet quality	MIL-F-48616
Scan rate	1200 scans/s
Rotational speed	6000 rpm

Source: Marshall, G.F.; Rynkowski, G.A. Eds. *Optical Scanning*; Marcel Dekker, Inc.: New York, 1991.

TABLE 5.3

Film Recorder Scanner Characteristics and Tolerances

Characteristic	Tolerance	Comments
Polygon data		
Number of facets	N.A.	Determined by scan angle
Facet angle	±10 arc sec	One pixel-pixel angle
Diameter	N.A.	Controlled by facet width dimension
Facet width	1.035 in mm	0.020-in roll-off
Facet height	0.5 in mm	0.020-in roll-off
Flatness	$\lambda/20$ max	Spot control
Reflectance	±3%	Tolerance
Apex angle	1.00 arc s	Total variation—10% of line–line angle
Speed regulation		
1 revolution	±10 ppm	±1.08 arc s/line
Long term	±50 ppm	±5.40 arc s/line

Note: Scan error for any 12 scans ≤ ±12.96 arc s. N.A., not applicable.
Source: Marshall, G.F.; Rynkowski, G.A. Eds. *Optical Scanning*; Marcel Dekker, Inc.: New York, 1991.

Our primary concern here is to relate the scanning system specifications, which ultimately determine the quality of the scanned image, to the performance requirements that are imposed on the motor and control system. It is apparent (Table 5.3) that the polygon rotation regulation plays a major role in the scanning system performance and that the design of the motor and control system must be given careful consideration in order to achieve the precise pixel placement accuracy.

Precision closed-loop control of the motor speed is essential if the performance targets outlined in Table 5.1 are to be met. To achieve the level of speed regulation required by the 13.97-μm pixel spacing specification, a phase-lock loop control system with a quartz oscillator frequency reference will be required. The 13.97-μm pixel spacing translates to a polygon speed regulation requirement of 0.001% (10 ppm) within one rotation of the scanner and over the time required to write the full page of the image onto the film. Speed

TABLE 5.4

Scanner Performance Comparison

Characteristics	Reference system	State-of-art system
Facet number	12	20
Facet tolerance	±10 arc s	±1 arc s
Apex angle error	±0.4 arc s	±0.2 arc s
Speed	6000 rpm	28,800 rpm
Scan rate	1200 scans/s	9600 scans/s
Speed regulation/rev	<10 ppm	<1 ppm
Pixels/scan	10,000	50,000
Pixel/jitter/rev	<±25 ns	<±2 ns
Pixel clock	12 MHz	480 MHz

Source: Marshall, G.F.; Rynkowski, G.A. Eds. *Optical Scanning*; Marcel Dekker, Inc.: New York, 1991.

variations within one turn of the polygon will produce an uneven scan line length, which will be visible as a variation along the edge of the film opposite the start of scan. Slower speed variations that occur over many revolutions but still within the same film frame may produce other undesirable image artifacts and distortions.

Subtle variations in the printed image such as shading and banding can also result from short-term speed changes within the scanner, which may occur over a few revolutions, and the human eye has a remarkable ability to detect these otherwise minor variations within the image.

5.3.4 High-Performance, Defined

Table 5.4 depicts the performance of the film recorder reference system in comparison to a state-of-the-art recording system considered by many as the highest resolution and fastest system manufactured to date. Note that the polygon speed regulation required in the state-of-the-art system is less than one part per million, or 0.0001%.

5.4 MOTOR CONSIDERATIONS

5.4.1 Motor Requirements

In the most demanding scanning applications, which require high speed and exceptional accuracy, the precision of the integral polygon, shaft, and bearing assembly must not be degraded by the introduction of the motor rotor. This places a heavy burden on motor rotor selection.

Any motor rotor attached to the shaft must have very stable and predictable characteristics with regard to strength and temperature, and if possible, the rotor material should be homogeneous. If the rotor is of a complex mechanical configuration and consists of laminations and windings, the assembly may not maintain a precision balance (less than 20 μin-oz) when operating at high speeds. Additionally, thermal expansion and high

centrifugal forces may shift and reposition the rotor and perhaps cause a catastrophic failure in an air bearing scanner.

Two motor designs are considered for use with high- and low-speed air bearing and ball bearing scanners, respectively; they are the hysteresis synchronous and the DC brushless. Pictured in Figure 5.10 is a high-speed DC brushless motor with integral Hall effect sensors for commutation and the associated rotor magnet mounted on a brass encoder hub assembly. Figure 5.11 depicts the two main components of a hysteresis synchronous motor: the stator and the hysteresis ring rotor.
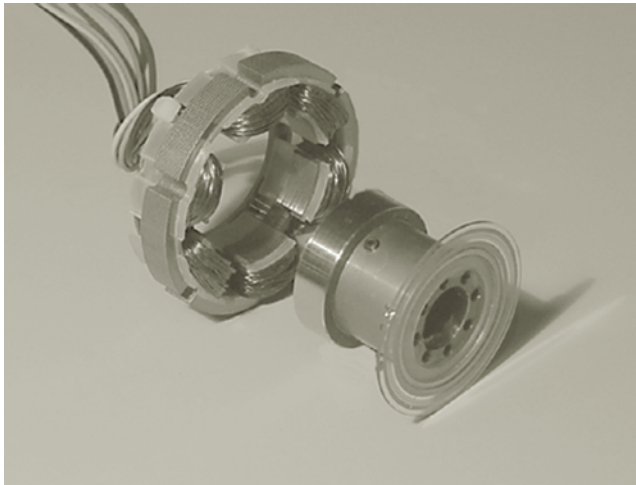
FIGURE 5.10
Brushless DC motor: stator and rotor.
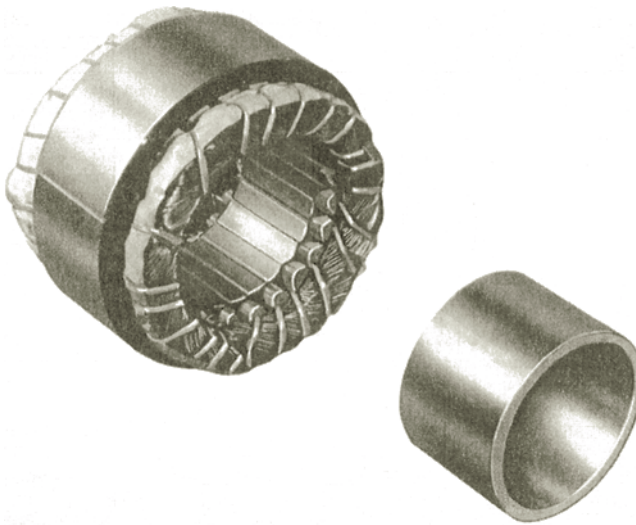
FIGURE 5.11
Hysteresis synchronous motor: stator and rotor. (From "Rotors, H.C. "The hysteresis motor—advances which permit economical fractional horsepower ratings," AIEE Technical Paper 47–218, 1947. With permission.)

5.4.2 Hysteresis Synchronous Motor

The difficult rotor mechanical stability requirements for high-speed scanner operation are easily achieved with the use of a hysteresis synchronous motor (Figure 5.11). The hysteresis rotor is uncommonly simple in design and consists of a cylinder of hardened cobalt steel that is heat shrunk onto the rotor shaft assembly. Careful calculations are required with regard to the centrifugal forces and thermal expansion influences on the rotor and shaft for safe and reliable operation. This type of motor is well suited for operation at speeds ranging from 1000 rpm to 120,000 rpm. Motors having output power as large as 2.2 kw have been successfully used on large-aperture IR scanning systems operating at 6000 rpm.

The operation of a hysteresis synchronous motor relies on the magnetic hysteresis characteristics of the rotor material. As the magnetizing force from a suitably powered stator (not unlike that used with reluctance-type motors) is applied to a cobalt steel rotor ring or cylinder, the induced rotor magnetic flux density will follow the stator coil current, as illustrated in Figure 5.12.

The sinusoidal current is shown to increase from zero, along the initial magnetization curve, to point (a), thereby magnetizing the material to a corresponding flux level at the peak of the sinusoid. As the current decreases to zero, the rotor remains magnetized at point (b). If the current at this point in time were to remain at zero, the rotor would be permanently magnetized at the point (b) flux level. However, as the current reverses direction, the flux reduces to zero at some negative value of current as shown at point (c). Further decreases in current (negative direction) reverse the direction of flux as shown at
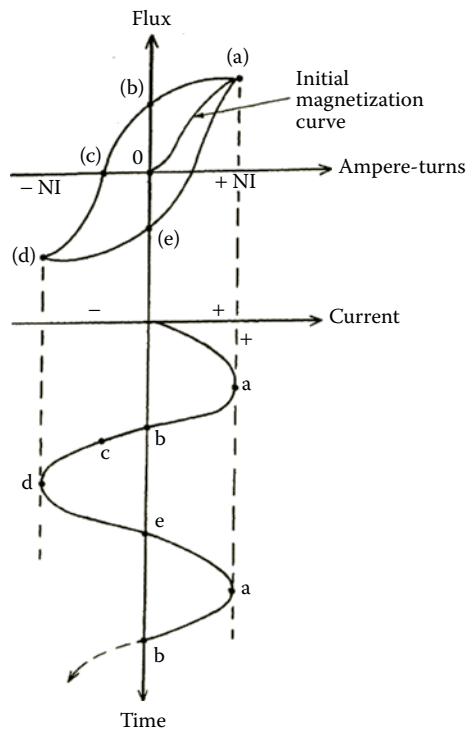
FIGURE 5.12
Magnetic hysteresis curve. (From Lloyd, T.C. *Electric Motors and Their Applications;* Wiley: New York, 1969. With permission.)

point (d), corresponding with the negative peak of the current. The process continues to point (e) and back to point (a), completing the loop for one cycle of current. The figure generated is called a magnetic hysteresis loop. In physics, hysteresis is defined as a lag in the magnetization behind a varying magnetizing force.

By analogy, as the axis of the magnetizing force rotates, the axis of the lagging force of the rotor will accelerate the rotor in the same direction as the rotating field. As the rotor accelerates, its speed will increase until it reaches the synchronous rotating frequency of the field. At this point, the rotor becomes permanently magnetized and follows the rotating field in synchronism.

The synchronous speed of the rotor can be calculated with the following expression:

$$\text{rpm} = \frac{120\,f}{N}$$

where f = line frequency (Hz) and N = number of poles.

Figure 5.13 depicts a typical speed versus torque characteristic curve for a hysteresis synchronous motor.

If a fixed line voltage and frequency is applied to the stator winding of the motor, an accelerating torque is developed equal to the starting torque, T_s, shown at point A. As the speed of the rotor increases, the operating point on the curve moves through the maximum torque developed at point B, and continues through point C, at which time synchronous speed is reached. The final operating point, D, is determined by the operating load torque presented to the shaft at torque level T_0. Note that if the operating load torque is greater than the in-sync torque, synchronous speed will not be reached.

Figure 5.14 is a vector representation of the rotating magnetizing field and the magnetized rotor field while in synchronism. Note that the rotor field vector lags the magnetizing field by an angle α. The operating torque (as developed by the motor in synchronism) is in proportion to the sine of the angle α in electrical degrees. If the load torque and stator frequency are absolutely constant, their frequencies will be precisely equal.

However, should the torque angle be modulated sinusoidally, as indicated in Figure 5.14 (with a torque variance of $\pm\beta$), the rotor vector will advance and retard as indicated about the average angle α. The long-term average speed will be as constant as the applied stator source frequency, but the instantaneous speed will follow the derivative of the sine wave
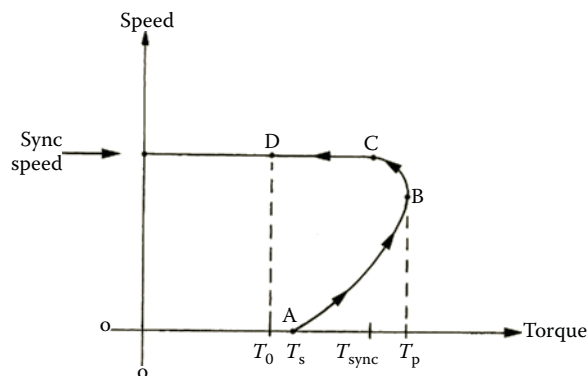
FIGURE 5.13
Speed/torque performance curve. (From Lloyd, T.C. *Electric Motors and Their Applications;* Wiley: New York, 1969. With permission.)
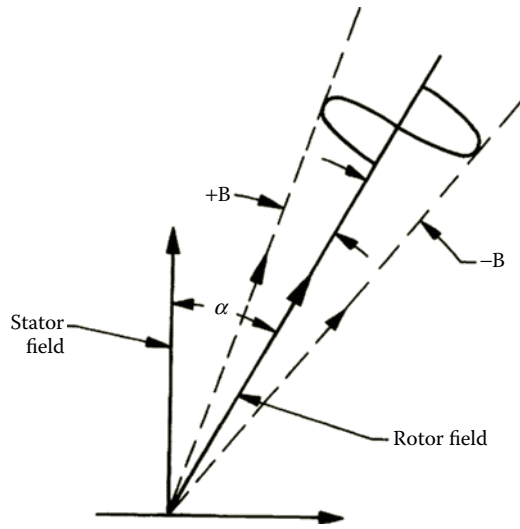
FIGURE 5.14
Stator/rotor field vectors. (From Lloyd, T.C. *Electric Motors and Their Applications;* Wiley: New York, 1969. With permission.)

on a one-to-one basis. The effect of torque perturbations that are not attenuated by system inertia will also modulate the shaft speed accordingly.

A characteristic of hysteresis synchronous motors (and other second-order devices and systems) called "hunting" may be observed when operating the motor in a system having low losses and damping factor. The motor rotor will oscillate in a sine wave fashion, not unlike that depicted in Figure 5.14, if perturbed by applied forces internal or external to the system. If the perturbations are sustained, the oscillations will also be sustained.

However, if the perturbations are not sustained, the oscillations will diminish in amplitude to essentially zero. The internal damping factor of the motor can be influenced by the rotor resistivity, rotor-to-stator coupling coefficient, and driver source and stator impedance. In a typical open-loop operation (no external velocity or position feedback) the oscillations may not be predictable and, therefore, may suddenly appear due to an unknown source or sources of perturbing forces. The amplitude of the oscillations in shaft degrees can typically range from 1° to 10°, and, as a practical matter, is very difficult to calculate; however, the rotar oscillating (hunting) frequency (W_n) can be estimated as

$$W_n = \sqrt{\frac{K}{I}}$$

where W_n = natural resonant frequency (rad/s), K = motor stiffness (in-oz/rad, or $\partial T/\partial a$), and I = shaft moment of inertia (in-oz s²).

The maximum instantaneous speed is determined by setting the derivative of the sine function to zero, and then calculating the maximum positional rate of change in radians per second:

$$\text{Change in speed (rad/s)} = \pm A_p W_n$$

where $A_p = \pm b$ rekey and W_n = natural resonant frequency (rad/s).
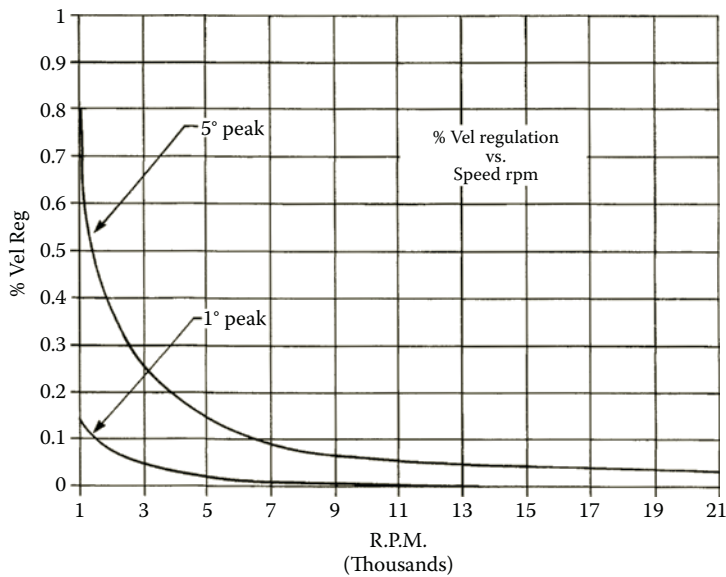
FIGURE 5.15
Velocity regulation versus speed. (From Marshall, G.F.; Rynkowski, G.A. Eds. *Optical Scanning*; Marcel Dekker, Inc.: New York, 1991. With permission.)

The maximum change in speed is often expressed as a percentage change relative to the nominal operating speed of the motor:

$$\text{Velocity regulation (\%)} = 100\left(\frac{A_p W_n}{W_s}\right)$$

where W_s = nominal operating speed (rad/s).

Figure 5.15 shows two curves of percent velocity regulation versus speed in rpm, for a typical open-loop scanning system having peak angular displacements of 1° and 5°. A four-pole motor with a peak torque of 10 oz-in and having a total inertia of 0.076 oz-in s² was used for the calculations in Figure 5.15. Note that for peak angular displacements of 1°, velocity regulation of 0.05% could be claimed for all speeds greater than 3000 rpm. However, should the peak angular displacements increase to 5°, then 0.05% velocity regulation is only obtainable at speeds greater than 14,000 rpm. Figure 5.15 illustrates that, for a typical open-loop scanning system operating at speeds above 3000 rpm, the system designer can expect variance in velocity ranging from less than 0.05% to as high as 0.25%. In conclusion, if system speed regulation requirements must be guaranteed to be less than 0.05% (500 ppm), closed-loop control of the motor speed using phase position feedback must be incorporated.

Another advantage of hysteresis synchronous motors in precision scanners is the near absence of rotor eddy currents when the rotor is synchronized with the rotating stator field. These currents can produce increased I^2R? rotor losses that can cause rotor/shaft distortions due to generated temperature gradients and produce adverse effects, especially in air bearing designs. The primary source of rotor eddy current losses is from the spurious flux changes that occur as the rotor passes the stator slots. These parasitic losses are often referred to as "slot effect losses" and can be very significant at high

speeds, rendering the device very inefficient as is often noted in older designs. Careful stator design can minimize these losses to the extent that the primary source of rotor/shaft heating is through the air bearing gap from the stator or due to air friction. Any residual heat generated by stator losses may be further reduced and diverted away from the bearing/rotor system by water or air cooling of the stator housing to minimize the rotor/shaft temperature rise.

In summary, AC hysteresis-synchronous motors were a natural choice in early polygon scanners, especially for high-speed applications. Simplicity of construction and reliable, maintenance-free operation were key advantages. Also, since the long-term shaft speed was precisely determined by the excitation frequency, no speed control system was required in order to produce acceptable results in low-cost systems. However, as scanner speed control requirements became more critical, feedback devices such as optical encoders were added in order to control the short-term speed variation or "hunting" found in this type of motor. This required increase in control complexity paved the way for the entry of DC brushless motors for precision scanning applications.

5.4.3 Brushless DC Motor Characteristics

The brushless DC motor (Figure 5.10) is well suited for speeds ranging from near zero to as high as 80,000 rpm. These motors exhibit the same characteristic as brush commutated types and can therefore be used in the same applications. They are also suited for velocity and position servo applications since they have a near ideal, linear control characteristic, meaning that the torque produced by the motor is in direct proportion to the applied current.

The elimination of brushes and commutating bars provides reduced electromagnetic interference, higher operating speeds and reliability, with no brush material debris from brush wear. The commutating switching function is accomplished by using magnetic or optical rotor position sensors that control the electronic commutating logic switching sequence. In actual operation, a DC current is applied to the stator windings, which generates a magnetic field that attracts the permanent magnets of the rotor, causing rotation. As the rotor magnetic field aligns with the stator field, the field currents are switched, thereby rotating the stator field and the rotor magnets follow accordingly.

The rotor will continue to accelerate until the motor output torque is equal to the load torque. Under no load conditions, the motor speed will increase until the back electromotive force (BEMF) generates a voltage equal to the stator supply voltage minus the DC winding resistance voltage drop. At this point, the rotor speed reaches an equilibrium level as determined by the BEMF motor constant.

The open-loop speed stability and regulation under controlled power supply and temperature conditions is usually 1%–5%, so the device is typically used with closed-loop feedback control. In the closed-loop mode of operation, particularly in the phase-lock loop configuration, short-term speed stability of 1 ppm is obtainable. However, on a long-term basis, the speed stability and accuracy are only as good as the reference source, which is routinely specified at 50 ppm or less for quartz crystal oscillator references, which are used in phase-lock loop speed control systems.

The brushless DC motor, when properly commutated, will exhibit the same performance characteristics as a brush commutated DC motor, and for servo analysis the two may be considered equivalent devices. Both motor types may be characterized by the same set of parameters as described in the following discussion.

5.4.3.1 Torque and Winding Characteristics

The basic torque waveform of a brushless DC motor has a sinusoidal or trapezoidal shape. It is the result of the interaction between the rotor and stator magnetic fields, and is defined as the output torque generated relative to rotor position when a constant DC current is applied between two motor leads. With constant current drive, the torque waveform follows the shape of the BEMF voltage waveform generated at any two motor winding leads. The frequency of the BEMF voltage waveform is equal to the number of pole pairs in the motor times the speed in revolutions per second. The BEMF waveform is easily observed by rotating the motor at a constant speed and is in fact often used to characterize the motor during testing.

The brushless DC motor exhibits torque/speed characteristics similar to a conventional brush-type DC motor. The stator excitation currents may be square wave or sinusoidal and should be applied in a sequence that provides a constant output torque with shaft rotation. Square wave excitation results in a small ripple in the output torque due to the finite commutation angle.

The commutation angle is defined as the angle the rotor must rotate through before the windings are switched. Ripple torque is typically expressed as a percentage of average-to-peak torque ratio and is present whenever the windings are switched by a step function either electrically via solid-state switches or mechanically via brushes. In brushless DC motors designed for square wave excitation, the ripple torque can be minimized by reducing the commutation angle through the use of a larger number of phases, which also improves motor efficiency. For a two-phase brushless motor, the commutation angle is 90 electrical degrees, which yields the largest ripple torque of about 17% average-to-peak. A three-phase delta-wound motor is shown in Figure 5.16. The commutation angle is 60° and the ripple torque is approximately 7% average-to-peak. Since two-thirds of the available windings are used at any one time, compared to one-half for the two-phase motor, the three-phase system is more efficient.

The torque waveforms shown in Figure 5.16 have a sinusoidal shape. For square wave excitation of the motor a trapezoidal torque waveform will produce improved torque uniformity. A trapezoidal torque waveform can be obtained by using a salient pole structure in conjunction with the necessary lamination/winding configuration. In practice, the trapezoidal torque waveform does not have a perfectly flat top and the benefit to torque ripple reduction may be small.

The commutation points and output torque for a three-phase brushless motor are shown in Figure 5.16. Each phase (winding) is energized in the proper sequence and polarity to produce the sum torque shown at the bottom of Figure 5.16 and is calculated to be equivalent to the current times the torque sensitivity of the motor (IK_T).

5.4.3.2 Brushless Motor Circuit Model

The equivalent electrical circuit model for a DC brushless motor is shown in Figure 5.17. This model can be used to develop the electrical and speed–torque characteristic equations, which are used to predict the motor performance in a specific application.

The electrical equation is

$$V_T = IR + L \, dI/dt + K_B w \tag{5.1}$$

where V_T = the terminal voltage across the active commutated phase, I = sum of the phase currents into the motor, R = equivalent input resistance of the active commutated phase,
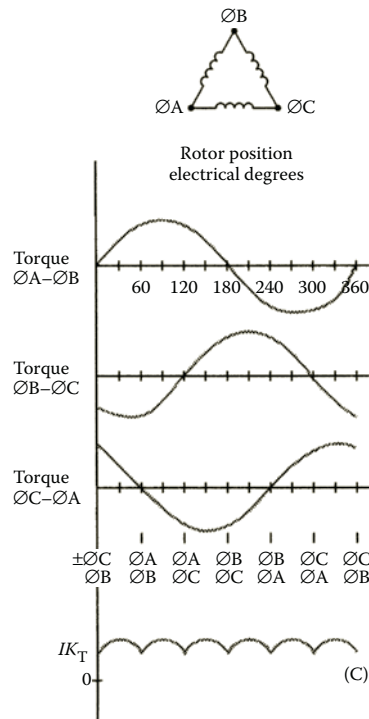
FIGURE 5.16
Three-phase motor torque and commutation points. (From Axsys Technologies Motion Control Products Division, San Diego, CA, *Brushless Motor Sourcebook;* Axsys: San Diego, CA, 1998. With permission.)
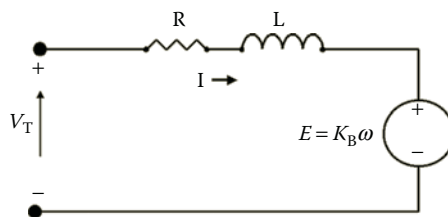
FIGURE 5.17
DC brushless motor equivalent circuit. (From Axsys Technologies Motion Control Products Division, San Diego, CA, *Brushless Motor Sourcebook;* Axsys: San Diego, CA, 1998. With permission.)

L = equivalent input inductance of the active commutate phase, K_B = BEMF constant of the active commutated phase, and ω = angular velocity of the rotor.

If the electrical time constant of the brushless DC motor is substantially less than the period of commutation, the steady-state equation describing the voltage across the motor may be written as:

$$V_T = IR + K_B w \tag{5.2}$$

The torque developed by the brushless DC motor is proportional to the input current such that:

$$T = IK_T$$

where K_T = the torque sensitivity (oz-in/A).

Solving for I and substituting into Equation 5.2 yields:

$$V_T = T/K_T R + K_B w \tag{5.3}$$

The first term represents the voltage required to produce the desired torque, and the second term represents the voltage required to overcome the BEMF of the winding at the desired operating speed. If we solve Equation 5.3 for rotor speed we obtain:

$$w = \left(\frac{V_T}{K_B}\right) - \left(\frac{TR}{K_B K_T}\right) \tag{5.4}$$

Equation 5.4 is the speed–torque relationship for a permanent magnet DC motor. A family of speed–torque curves represented by Equation 5.4 is shown in Figure 5.18. The no-load speed may be obtained by substituting $T = 0$ into Equation 5.4:

$$w \text{ (no load)} = \frac{V_T}{K_B}$$

Stall torque can be found by substituting $\omega = 0$ into Equation 5.4:

$$T(\text{stall}) = \frac{K_T V_T}{R} = IK_T$$

The slopes of the parallel line speed curves of Figure 5.18 may be expressed by:

$$\frac{R}{K_B K_T} = w(\text{no load})T(\text{stall})$$
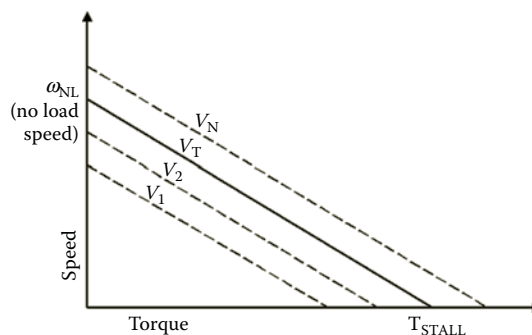
FIGURE 5.18
DC motor characteristic curves. (From Axsys Technologies Motion Control Products Division, San Diego, CA, *Brushless Motor Sourcebook;* Axsys: San Diego, CA, 1998. With permission.)

Since the speed–torque curves are linear, their construction is not required for predicting motor performance. The system designer can calculate the required information for servo performance from the basic motor parameters given by the manufacturer.

5.4.3.3 Winding Configurations

Almost all of the brushless motors and drives produced today will be of the three-phase configuration, although two-phase motors have unique advantages, which are exploited in very low-speed applications. Three-phase windings can be connected in either a "delta" or "Y" configuration as shown in Figure 5.19.

Excitation currents into the windings can be switched full on, full off, or be applied as a sinusoidal function depending on the application. The switch mode drive is the most commonly used system because it results in the most efficient use of the electronics. Two switches per phase (winding) terminal are required for the switch mode drive system. Therefore, only six switches are required for either the "Y" or "delta" configuration.

The delta windings form a continuous loop, so current flows through all three windings regardless of which pair of terminals are connected to the power supply. Since the internal resistance of each phase is equal, the current divides unequally, with two-thirds flowing through the one winding connected directly between the switched terminals, and one-third flowing through the two series-connected windings appearing in parallel with it. This results in switching only one-third of the total current from one winding to another as the windings are commutated.

For the "Y" connection, current flows through the two windings between the switched terminals. The third winding is isolated and carries no current. As the windings are commutated, the full load current must be switched from terminal to terminal. Owing to the electrical time constant of the windings, it takes a finite amount of time for the current to reach full value. At high motor speeds, the electrical time constant ($T_E = L/R$) may limit the switched current from reaching full load value during the commutation interval, and thus limits the generated torque. This is one of the reasons the delta configuration is preferred for applications requiring high operating speed. Other considerations are manufacturing factors, which permit the delta configuration to be fabricated with lower BEMF constant, resistance, and inductance. A lower BEMF constant allows the use of more common low-voltage power supplies, and the solid-state switches will not be required to switch high voltage. For other than high-speed applications, the "Y" connection is preferred because

Delta winding
configuration

"WYE" winding
configuration
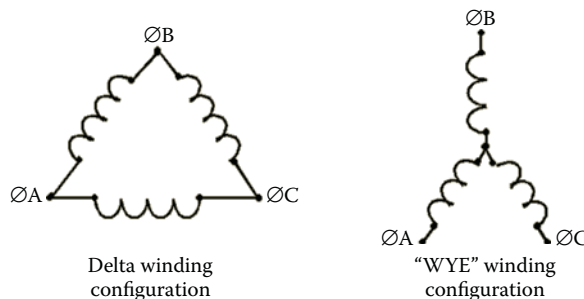
FIGURE 5.19
Three-phase motor winding configurations.

it provides greater motor efficiency when used in conjunction with brushless motors designed to generate a trapezoidal torque waveform.

5.4.3.4 Commutation Sensor Timing and Alignment

A brushless DC motor duplicates the performance characteristics of a DC motor only when its windings are properly commutated, which means that the winding currents are applied at the proper time, polarity, and in the correct order. Proper commutation involves the timing and sequence of stator winding excitation. Winding excitation must be timed to coincide with the rotor position that will produce optimum torque. The excitation sequence controls the polarity of generated torque, and therefore the direction of rotation. Rotor position sensors provide the information necessary for proper commutation. Sensor outputs are decoded by the commutation logic electronics and are fed to the power drive circuit, and activate the solid-state switches that control the winding current.

A useful method to achieve correct commutation timing is to align the position sensor to the BEMF waveform. Since the BEMF waveform is qualitatively equivalent to the torque waveform, the test motor can be driven at a constant speed by another motor, and the position sensors aligned to the generated BEMF waveform. The sensor transition points relative to the corresponding BEMF waveforms should be coincident when correct commutation has been achieved. For critical applications that require the commutation points to be optimized, the motor should be operated at its rated load point, and then the position sensors should be adjusted until the average winding current is at its minimum value.

The commutation points and output torque for a three-phase brushless motor are shown in Figure 5.16. The commutation angle is 60 electrical degrees. The windings are switched "on" at 30 electrical degrees before the peak torque position, and switched "off" at 30 electrical degrees after the peak torque position. The current polarity must be reversed for negative torque peaks to produce continuous rotation. To identify each of the six commutation points, a minimum of three logic signals are required, as shown in Figure 5.20. The logic signals are generated by three sensors which are spaced 60 electrical degrees apart and produce a 50% duty cycle.

As indicated in Figures 5.16 and 5.20, sensor S1 can be readily aligned to the ØA–ØB zero torque position. This can be accomplished by applying a constant current to the ØA–ØB terminals. The rotor will rotate to the ØA–ØB zero torque position and then stop. Sensor Sl should then be positioned so that its output just switches from a low to high logic state. Sensors S2 and S3 may then be positioned 120 and 240 electrical degrees respectively from S1 (either CW or CCW direction depending on the direction of rotation) and basic commutation will be established for the motor.

5.4.3.5 Rotor Configurations

The brushless DC motor rotor configurations (Figure 5.21) most often used for scanners consist of rare-earth samarium–cobalt permanent magnets that are contained with a rigid ring or cup for the outer rotor configurations, and are usually bonded to a machined hub for the inner rotor configurations. The inner rotor configuration is generally used at the higher speeds because of the lower centrifugal forces resulting from a reduced rotor diameter.

Owing to the elastic characteristics of epoxy and other adhesives, and the need for stable and reliable precision balancing, the operating speed of DC brushless spinner rotors is generally less than what can be expected for the hysteresis motors. However, DC brushless
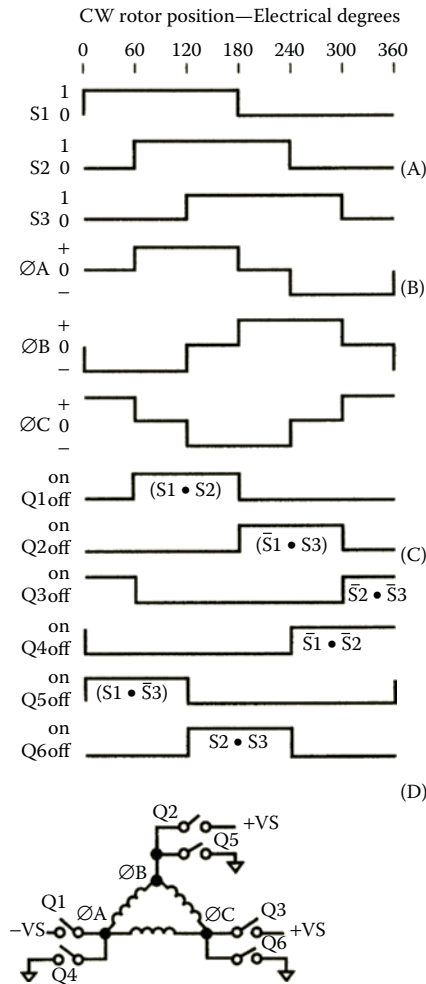
FIGURE 5.20
Three-phase motor commutation logic and excitation. (From Axsys Technologies Motion Control Products Division, San Diego, CA, *Brushless Motor Sourcebook;* Axsys: San Diego, CA, 1998. With permission.)

motor technology is steadily displacing other motor types in many high-speed applications. Stainless steel sleeves have been used to aid in the retention of the rotor magnets and improve the rotor mechanical characteristics at high speeds. In addition, ring magnet rotors have been developed for high-speed designs, allowing DC brushless motors to operate in excess of 80,000 rpm. In this rotor configuration, a ring of the rotor material is magnetized by the pole pieces of a powerful electromagnet, which imprints the pole locations and polarities into the ring. This type of rotor design has in fact been employed to drive precision polygonal scanners for laser projection systems which operate at 81,000 rpm.

Pictured in Figure 5.10 is a low-cost, high-speed DC brushless motor, which has been successfully used in several scanner designs. The simple stator design, which allows for machine winding of the coils, and the inexpensive ring magnet rotor, account for the low manufacturing cost of this rugged and reliable motor. The rotor magnet is further strengthened by an outer stainless steel sleeve.
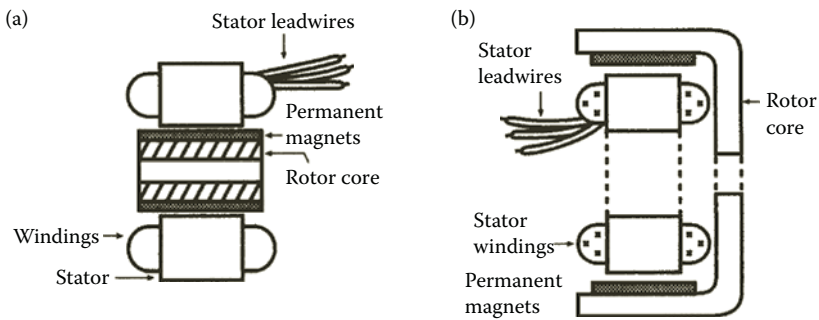
FIGURE 5.21
Brushless DC motor configurations. (a) inner rotor brushless DC motor; (b) outer rotor brushless DC motor. (From Axsys Technologies Motion Control Products Division, San Diego, CA, *Brushless Motor Sourcebook;* Axsys: San Diego, CA, 1998. With permission.)

FIGURE 5.22
Low-speed DC brushless motor.

The commutation sensors (Hall effect devices) are placed within the stator slots and do not require any timing adjustments. This motor is capable of delivering at least 50 W on a continuous duty basis. Also visible in Figure 5.10 is the rotor hub assembly onto which the ring magnet and the optical encoder disc are mounted.

Pictured in Figure 5.22 is a miniature eight-pole brushless DC motor, which is suitable for many low-power scanner drive applications. The stator configuration is noticeably more complex than the low-cost motor pictured in Figure 5.10 and is intended for lower speed applications where cogging torque must be minimized. Commutation of the windings in this design is performed by dedicated channels of a shaft-mounted optical encoder rather than Hall effect sensors.

Commutation timing adjustment is made possible by rotating the encoder pattern relative to the rotor magnet position. Optimum timing at the desired operating speed is required in order to produce the lowest torque ripple and power consumption. The rotor assembly is of a more conventional design, where individual magnet pieces are bonded to a machined hub, which is then ground to the final dimensions.

For low-speed applications it is desirable to construct the motor using the highest pole count possible, that is consistent with the physical size and the electrical parameters specified. The smaller commutation angle and the resulting higher ripple torque frequency may be more readily attenuated by the rotating inertia in the system and produce a more constant speed within one revolution of the rotor.

5.5 CONTROL SYSTEM DESIGN

The basic control requirements for a precision polygonal spinner are to provide synchronization and velocity control for precise scan registration, whether on a film plane or detector array, or at a distant target being illuminated. To this end, the principles of feedback control are utilized for synchronization, velocity, and phase position (shaft angle) control.

In our film recorder example (Figure 5.9), speed control is required for accurate pixel positioning, repeatability, and linearity, as well as line-to-line pixel registration and synchronization. In order to accurately position data pixels in a line at the film plane, the system must generate precision pulses spatially related to the facet angles. These pulses, occurring on a one per pixel basis, are used to gate in and turn off the intensity-modulated video being projected to the light-sensitive film surface. Because there are 120,000 pixels in the film recorder example (12 times 10,000) per revolution, one clock pulse would be required for every 10 arc s of shaft rotation.

Optical encoders are well suited to the task of generating accurately timed and positioned pulses. However, incremental, high-density data track encoders are expensive, large in diameter, and difficult to mechanically interface with an integral spinner/motor/shaft assembly.

To overcome this problem, a smaller and inexpensive low-density optical encoder have 6000 pulses per revolution (PPR) was designed into the system. The required 120,000 PPR pixel clock pulses are obtained by electronically multiplying the encoder data track frequency (6000 PPR) by a factor of 20. Scanner speed control is accomplished by frequency/phase locking the encoder data track (600 KHz) to an accurate and stable crystal oscillator. An index pulse is accurately positioned at the normal of a facet on a second encoder track, thereby providing start of scan (refer to Figure 5.9, SOS detector) synchronization and pixel registration.

This system of generating pixel clock pulses places a heavy burden on the performance accuracy of the rotating shaft assembly, the encoder design and adjustment, and the stability of the crystal oscillator. Nevertheless, the net speed control and jitter performance has been optically measured to be less than ±10 ppm for one revolution of the scanner.

5.5.1 AC Synchronous Motor Control

Velocity control of the hysteresis synchronous motor is intrinsic to its design, that is, the long-term speed is as accurate as the applied frequency. With reference to Figure 5.14, the stator and rotor fields rotate together (at an integer submultiple of the applied

frequency) with the rotor lagging by the torque angle α and with possible modulations of $\pm\beta$, as was previously discussed. The control systems' task is to fix the rotor vector position, and therefore eliminate hunting and other speed variances.

To implement phase-lock control, the shaft rotational frequency and phase position are measured with a shaft-mounted incremental encoder. The encoder pulses are frequency and phase compared with a stable reference frequency using a frequency/phase comparator, which has the transfer characteristics as shown in Figure 5.23. The frequency/phase comparator has a unique transfer characteristic that allows the device to produce an output that is in saturation until the two input frequencies are equal. This is the frequency detection mode of the phase comparator. The saturation level is either positive or negative in value, and is useful in determining whether the motor speed is too high or too low. Assuming that the motor has reached synchronous speed, the tachometer frequency will equal the reference frequency, and the frequency/phase comparator will operate in the phase comparator mode. In this mode of operation, the output of the phase comparator is an analog voltage proportional to the phase difference between f_T and f_R. At zero frequency and phase differences, the two signals are edge locked, and the phase comparator output voltage is zero. Should the shaft advance or retard for any reason, the phase comparator error voltage will be in direct proportion to the phase difference within ± 360 electrical degrees of the reference frequency.

With reference to Figure 5.24, the phase comparator error is processed through a proportional-integral-derivative (PID) controller compensation scheme and fed to the control input of the phase modulator.

The phase error-corrected phase modulator output frequency, f_M is applied to the motor, thereby completing the position control loop from the encoder to the frequency/phase comparator. The open-loop DC gain of the system is primarily determined by the product of the encoder, frequency/phase comparator, integrator, and phase modulator gains. The high DC gain of the integrator (100 dB) reduces the phase error between f_R and f_T to zero, resulting in near perfect synchronization. The differentiation gain constant provides sufficient damping to eliminate "hunting" and improve the overall dynamic performance and speed regulation to less than 1 ppm.

5.5.2 DC Brushless Motor Control

The velocity control of a brushless DC motor differs from that of hysteresis synchronous in that the brushless motor speed is a function of applied motor voltage, as opposed to
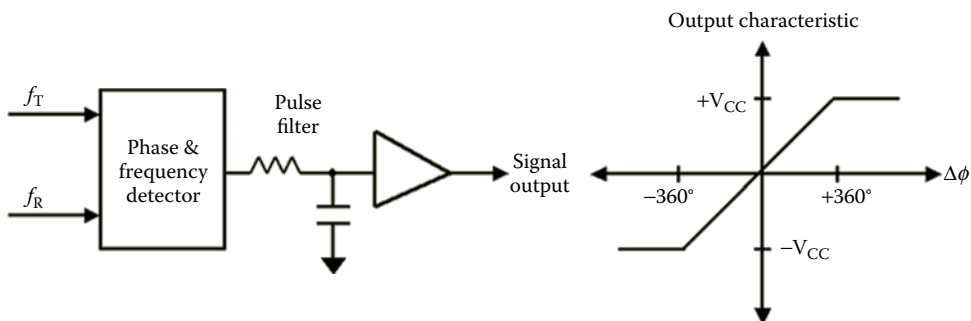
FIGURE 5.23
Phase/frequency detector (comparator) characteristics.
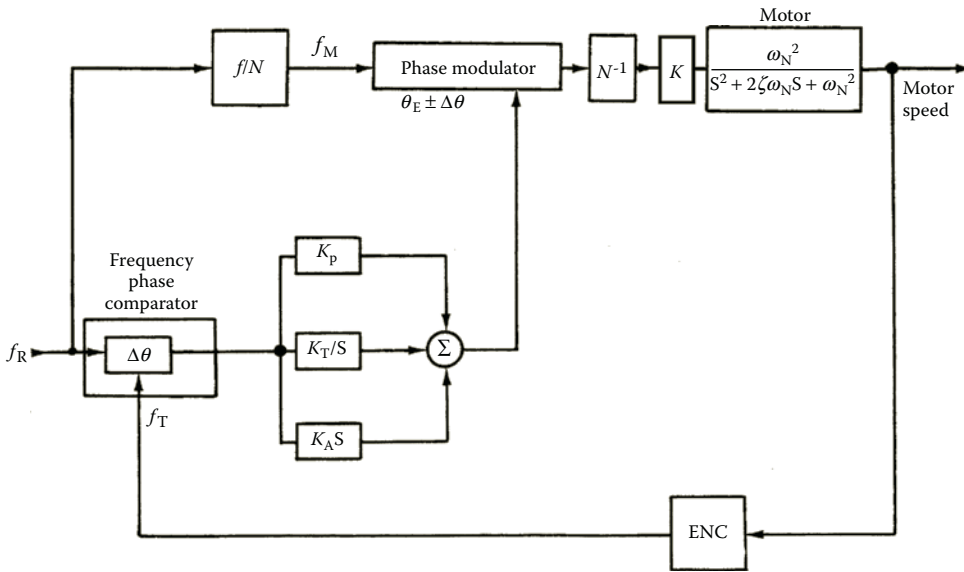
FIGURE 5.24
Control system block diagram, hysteresis motor. (From Marshall, G.F.; Rynkowski, G.A. Eds. *Optical Scanning*; Marcel Dekker, Inc.: New York, 1991. With permission.)

the frequency/phase as the driving function of the latter. The same principles of feedback velocity/position control are utilized in a similar fashion and are depicted in Figure 5.25. The elements within the closed-loop block diagram are essentially the same, with the exception of the motor transfer function and the addition of a DC power amplifier. The DC brushless motor transfer function is shown in detail in Figure 5.26. For simplification purposes, the commutating and pulsewidth modulation circuits have been omitted, but will be covered in a later discussion.

As before, to implement phase-lock control, the shaft frequency and phase position are measured with a shaft-mounted incremental encoder. The encoder pulses are frequency and phase compared with the reference frequency using a frequency/phase comparator, which has the transfer characteristics shown in Figure 5.23. The frequency/phase comparator has a unique transfer characteristic that allows the device to produce an output that is in saturation until the two input frequencies are equal. The saturation level is either positive or negative in value, and is useful in determining whether the motor speed is too high or too low.

The controller will accelerate or decelerate the motor until there is no frequency difference between the reference and the encoder signal. At this point a phase measurement is made by the comparator with every reference pulse cycle and an output voltage is generated that is proportional to the phase error.

The phase error signal is then processed through a PID controller and compensator, similar to that used in the hysteresis motor control system (Figure 5.24).

The detailed DC motor transfer function shown in Figure 5.26 relates the motor angular velocity w_s to the applied terminal voltage V_T. In English units the constants are defined as (across any two leads for a three-phase delta, or three-phase Y motor): R = motor winding resistance (Ohms); L = motor inductance (Henry); I_M = motor current (Amperes); K_T = motor
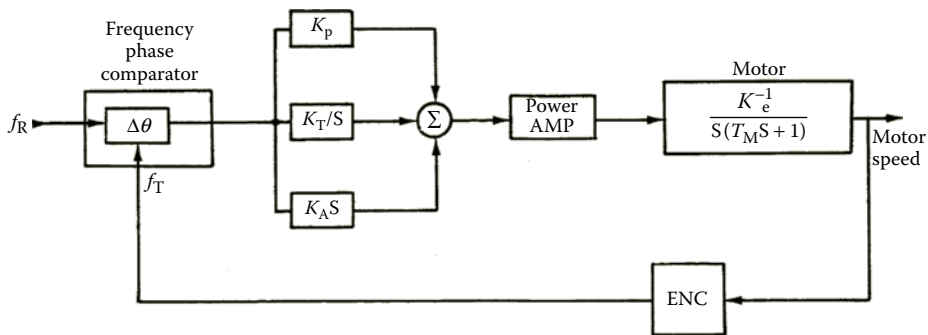
FIGURE 5.25
Control system block diagram, brushless DC motor.
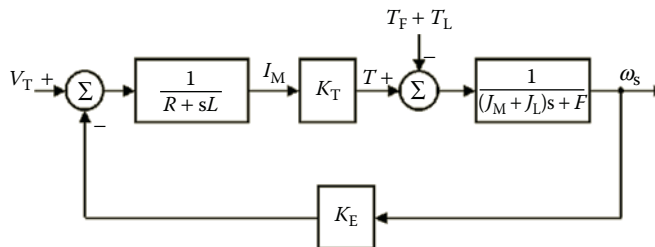
FIGURE 5.26
DC motor transfer function block diagram.

torque sensitivity (oz-in/Amp); K_B = motor BEMF constant (Volts/radian/s); J_M = motor moment of inertia (oz-in/s²); J_L = load moment of inertia (oz-in/s²); $T_F + T_L$ = sum of the friction and load torque (oz-in); and F = motor damping coefficient.

As the control system is turned on, the error signal is power amplified, causing the motor to accelerate to a speed at which f_T exceeds (overshoots) f_R. At this point, the error signal reverses polarity, reducing the motor speed until f_T equals f_R. Ultimately, a point of equilibrium is reached at which time the frequency/phase comparator error voltage is zero, and the integrator output voltage regulates the speed of the motor. Furthermore, the high DC gain of the integrator maintains a zero phase difference between f_T and f_R resulting in edge lock synchronization.

To ensure stable and accurate speed control, and to determine the gain coefficients K_P, K_I, and K_A, the motor and load characteristics should be modeled. Several very useful simulation programs such as *SIMULINK* (The Math Works, Inc., Natick, MA) are available for the systems designer which greatly reduce development time by allowing the rapid testing of various control configurations.

In conclusion, the DC brushless motor, when properly designed for the scanning application at hand, is capable of meeting or exceeding the performance of AC hysteresis motors in all but the highest speed applications. All of the successful scanning products that are discussed at the end of this chapter use brushless DC motor technology.

5.6 APPLICATION EXAMPLES

The following sections describe a few of the many scanner designs and control systems that have been developed over the last few years. The overall industry trend reflects the unrelenting drive of the end use market to improve performance, reduce power consumption and size of the supporting electronics, and to reduce the cost of the scanner subsystem.

Fortunately, the consumer electronics and automotive markets have provided many of the electronic components that have proven to be invaluable in the quest to reduce the size and cost of the scanning system. Progress in miniaturization and power reduction of the scanner control electronics is also largely due to the advancements in brushless DC motor technology. As the cost and complexity of the drive electronics for brushless motors approach those of conventional DC motors, brushless motor technology will likely displace all other motor types in scanning subsystems, as it offers high efficiency along with the high reliability, as previously found only in AC motors.

5.6.1 Military Vehicle Thermal Imager Scanner

Pictured in Figure 5.27 is a small 12-facet polygon scanner that is designed for a military vehicle thermal camera system. This compact ball bearing scanner operates at 600 rpm and serves to generate the field scan function in conjunction with an infrared detector array. The motor and control system is designed to maintain polygon speed regulation to within 15 ppm (0.0015%). In addition, the control system must maintain the specified speed regulation in the presence of base disturbances, which are passed to the scanner as a result of vehicle motion. These challenges have been met in this compact scanner by the use of a high-resolution encoder and lightweight polygon design in conjunction with an agile control system.

The scanner employs a small low-voltage brushless DC motor, which is optimized for low cogging torque and smooth operation at low speed. Motor commutation is derived
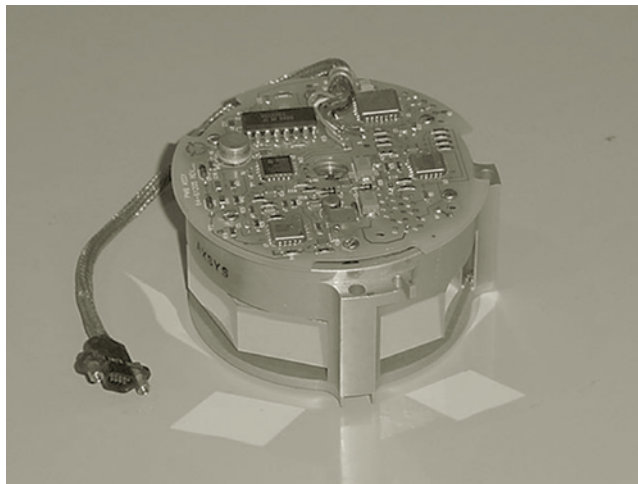
FIGURE 5.27
Military vehicle thermal imager scanner.

from three dedicated commutation tracks on the optical encoder disk, which also includes a high resolution 3000 count tachometer track as well as an index.

At the center of the control system is a single-chip motor driver, which decodes the commutation information provided by the encoder and produces the correct three-phase motor current waveforms necessary for proper operation of the motor. Motor driver ICs of this type are commonplace devices found in computer disk drive and CD player applications. The motor driver produces a current through the motor windings that is proportional to a command voltage at its input and also incorporates a brake feature that is used to decelerate the motor for better control of the polygon speed under the influence of disturbances.

Tight speed control is accomplished by the use of a phase-lock loop regulation method as described in detail in the previous sections. At the operating speed of 600 rpm, the 3000 line optical encoder produces a tachometer frequency of 30 KHz, which is compared against an externally generated reference frequency in the phase/frequency comparator circuit. Any resulting phase error is amplified and filtered, and then fed to the motor driver, which increases or reduces the motor current to maintain the speed and minimize the phase error. If the motor control voltage falls below a predetermined level indicating that the scanner is operating above the reference speed, then the control system applies dynamic braking to the motor, which quickly decelerates the motor and the polygon.

A block diagram of the controller/driver is shown in Figure 5.28.

5.6.2 Battery-Powered Thermal Imager Scanner

The polygon scanning system pictured in Figure 5.29 was designed for a compact, low-cost, military thermal sighting device primarily intended for small arms applications. Many unique and difficult requirements are imposed by the system specifications on this relatively low-cost device. A wide operating temperature range, precise scanning speed, low cross-scan error, and low power consumption must simultaneously be met. In addition, the scanner must meet the demanding performance specifications for a high-resolution
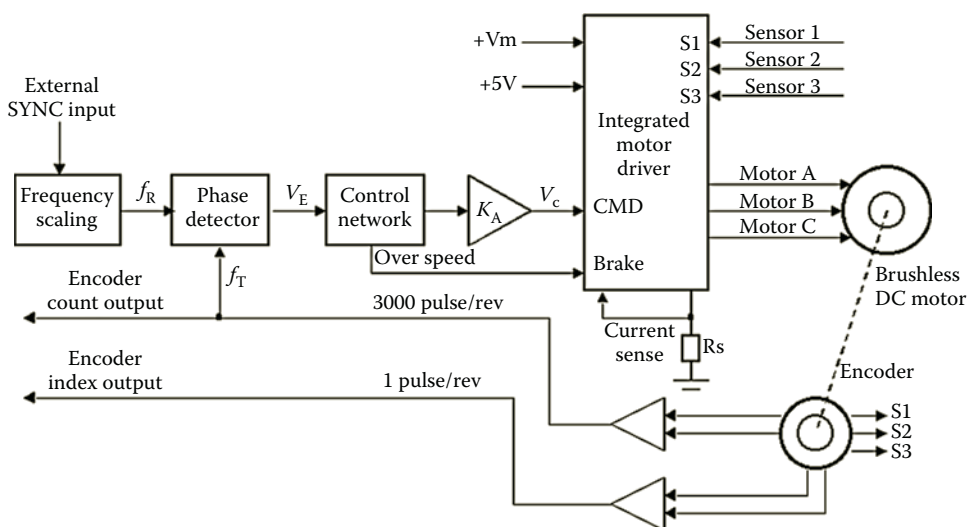
FIGURE 5.28
Block diagram, military vehicle scanner controller/driver.
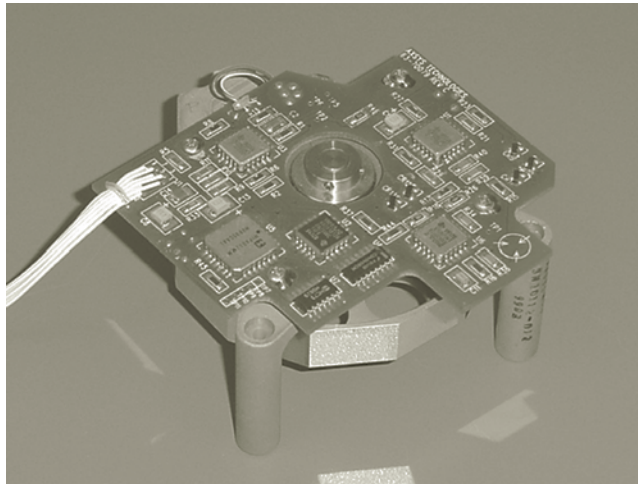
FIGURE 5.29
Battery-powered thermal imager scanner.

TABLE 5.5

Battery-Powered Scanner Characteristics

Operating temperature	−40 to +75°C
Polygon speed	600 rpm
Speed jitter (one rev)	15 ppm or less (0.0015%)
Encoder outputs	1500 ppr and index
Input voltage	+10 VDC and +5 VDC
Peak shock level	500 g (0.5 ms)

imaging system while being subjected to severe levels of shock and vibration. Particular attention was given to ensure the mechanical stability of the polygon under severe environmental conditions. Some of the scanner requirements are as shown in Table 5.5.

In order to maximize battery life, the total power consumption of the scanner was limited by the specification to less than 0.4 W while operating in synchronism at the low-temperature extreme.

A compact low-speed brushless DC motor similar to that in Figure 5.22 was designed specifically to address the low power and low cogging torque requirements necessary for this application. A new, single chip motor driver, which employs PWM control, was selected to improve efficiency in the drive and help meet the power consumption requirement. The driver PWM efficiency benefit becomes more apparent as the motor load increases and the drive must supply more current to maintain speed regulation. With a linear motor driver such as the one used in the military vehicle scanner, additional power is lost in the form of heat dissipated in the driver power section.

Effective speed regulation is achieved by the implementation of the same control scheme used in the military vehicle scanner shown in Figure 5.28. The optical encoder tachometer track line density was reduced to 1500 counts per revolution to meet the interface requirements of the imaging system. At 600 rpm the encoder tachometer track produces a 15-KHz pulse frequency, which is at an adequately high rate to maintain precise speed control of the polygon.

The phase-lock loop control system was optimized to maintain phase lock of the encoder tachometer with the synchronization reference under the influence of base motion disturbances and thereby prevent the loss of the image produced by the camera system during movement or under vibration. The need to reject base motion disturbances while providing precise speed control poses a unique challenge in the design of the control system. The scanner total rotational inertia must be minimized to allow the motor torque to accelerate and decelerate the polygon in order to overcome disturbance-induced speed changes. On the other hand, the rotational inertia within the scanner acts to reduce the effect of bearing and motor torque fluctuations, which tend to degrade speed stability. A compromise design was reached that adequately addresses both needs and meets the performance targets set for the scanner.

In order to reduce cost and update the electronic design of the control system, a new controller/driver circuit board was developed, which is shown in Figure 5.30. Significant cost savings were realized as older, ceramic packaged military grade integrated circuits were replaced with industrial quality devices. Performance and environmental specifications for the scanner were met with the plastic-packaged industrial ICs, and the part obsolescence issues were solved that appear with ever greater frequency in the electronic components world.

5.6.3 High-Speed Single-Faceted Scanner

The successful scanner design shown in Figure 5.31 was developed for the publishing and printing industry for use in the image setting machines that are the output devices leading to the manufacture of printing plates. The single-faceted mirror, or monogon, rotates at high speed and scans an intensity modulated laser spot along a sheet of film not unlike that shown in the film recorder system of Figure 5.9. The exceptions are that the film sheet

FIGURE 5.30
Low-cost scanner controller.

FIGURE 5.31
High-speed scanner and controller/driver.

usually lies on the inside surface of a cylindrical drum and the F–θ lens is omitted. Also, the film sheet is stationary and the scanner rides the length of the drum on precision linear bearings driven by a ball-screw mechanism.

For improved speed stability and cross-scan accuracy, as well as greatly improved bearing life, the scanner rotating elements are supported by self-pumping conical air bearings. This type of bearing provides excellent high-speed stability and low friction with the benefit of virtually unlimited operating life. The self-pumping action of the air bearing design is a major factor in reducing the cost of the scanner because an external air supply is not required. Some versions of this scanner operate at 60,000 rpm and utilize the highspeed motor design shown in Figure 5.10. The low-cost optical encoder shown in Figure 5.9 provides up to 2000 pulses per revolution for effective speed control and also sends mirror position information to the imaging system. The success of this scanner design is partly due to the development of a low-cost single card controller, which is described in greater detail in the next section. The scanner and control system block diagram is shown in Figure 5.33.

5.6.4 Versatile Single Board Controller and Driver

The need for a versatile, compact, and inexpensive motor driver and speed controller led to the development of a successful circuit capable of delivering up to 100 W of power to a three-phase brushless motor scanner. All of the functions necessary to achieve precise scanner speed control have been integrated into one unit. The reference frequency generator, phase-lock loop controller, and motor driver are combined on a single low-cost circuit card, which measures 4 × 8 in. A great deal of flexibility has been incorporated into this low-cost controller design so that many scanning applications can be readily accommodated without the need for circuit modifications. Various encoder resolutions and reference frequencies may be accommodated by setting jumpers that reconfigure the digital logic in order to present compatible frequencies to the phase comparator circuit.
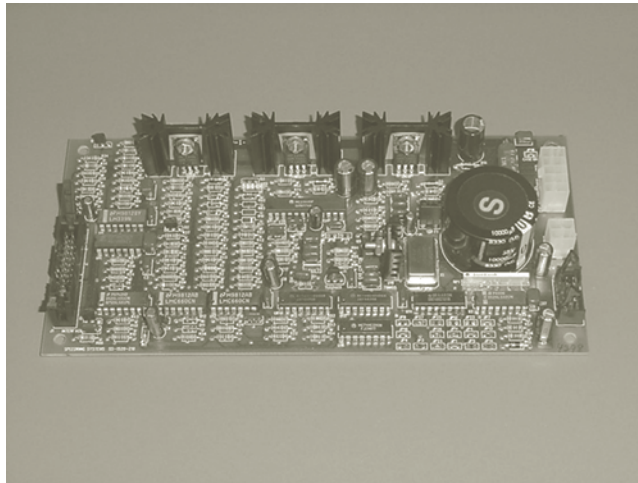
FIGURE 5.32
Single board scanner controller and motor driver.

This single board controller has been utilized to drive single-faceted as well as polygon scanners operating between 3000 and 81,000 rpm for a variety of laser scanning applications, and has demonstrated excellent speed regulation capabilities. The rotational speed jitter in many of the air bearing scanners driven by this controller was measured to be only a few parts per million within one rotation. This successful design has been incorporated into thousands of scanning systems sold to the printing and publishing industry worldwide. The controller is shown in Figure 5.32.

As shown in Figure 5.33, the circuit functions can be divided into the following main categories:

1. Reference frequency generator and external sync processing circuit
2. Phase detector and PWM synchronization circuit
3. Control loop PID circuit
4. Brushless motor controller and FET power section

The controller reference frequency generator and external sync processing circuits function to provide a precise and stable frequency reference to the phase comparator. An on-board quartz reference oscillator and programmable divider provide for the selection of up to 16 preset operating speeds for the scanner. The speed may also be continuously varied within a wide band with the application of an external reference frequency from the imaging system controller. In this way, fine adjustments to the scanning speed may be made to trim the system optical parameters.

The phase detector and frequency comparator circuits produce the speed error voltage, which is then amplified and sent to the servo compensation network. The phase detector exhibits high gain since the output is in saturation until the two frequencies f_R and f_T are exactly the same, as discussed previously and shown in Figure 5.23. This feature is responsible for the precise nature of phase-lock loop speed control systems.
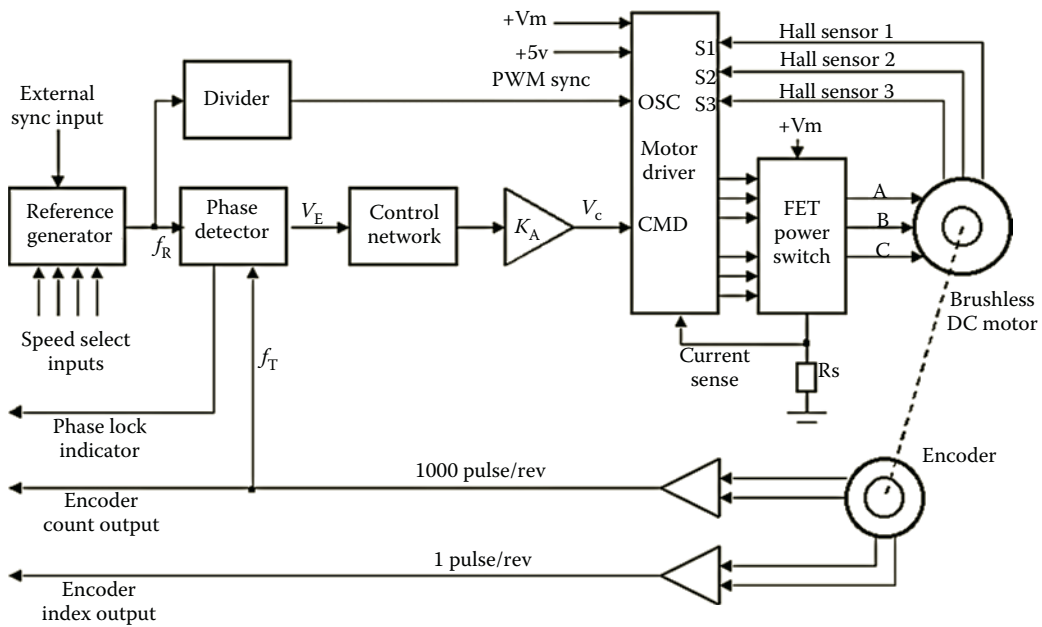
FIGURE 5.33
High-speed scanner and controller functional diagram.

Another useful innovation developed in the quest to provide the best possible speed regulation is the synchronization of the PWM oscillator with the phase detector reference frequency. The synchronization of these two frequencies ensures that the noise generated by the beat frequencies do not interfere with the scanner speed control circuits. When synchronized, the two frequencies produce a sum or difference (beat frequency), which will be constant. Stationary beat frequencies may be filtered or appear only as DC offsets, which may be subtracted from the speed control signal. The PID servo controller is similar to the arrangement shown in Figure 5.25 and described in earlier sections. Because most of the scanning systems targeted for this controller are for stable and controlled environment applications, the servo control loop is optimized for speed control regulation rather than response time. In these applications, high rotating inertia in the scanner rotor and polygon is a benefit to speed regulation.

The motor driver and power output section consists of a monolithic (single chip) controller and discrete FET power switches. With a modest amount of forced air cooling, the FET power section is capable of delivering up to 4 A continuously and 6 A for several seconds to deliver higher motor starting current. The motor current is regulated by the driver IC using PWM control, which is effective in delivering power to the motor with minimum heat generation in the controller.

For some high-speed polygonal scanning applications it may not be possible to include an optical encoder within the scanner housing. In this instance, the polygon facet frequency may be used as the speed feedback sensor as depicted in Figure 5.34. The optical pulse frequency should be at or above 1 KHz in order to provide precise speed regulation in this configuration.
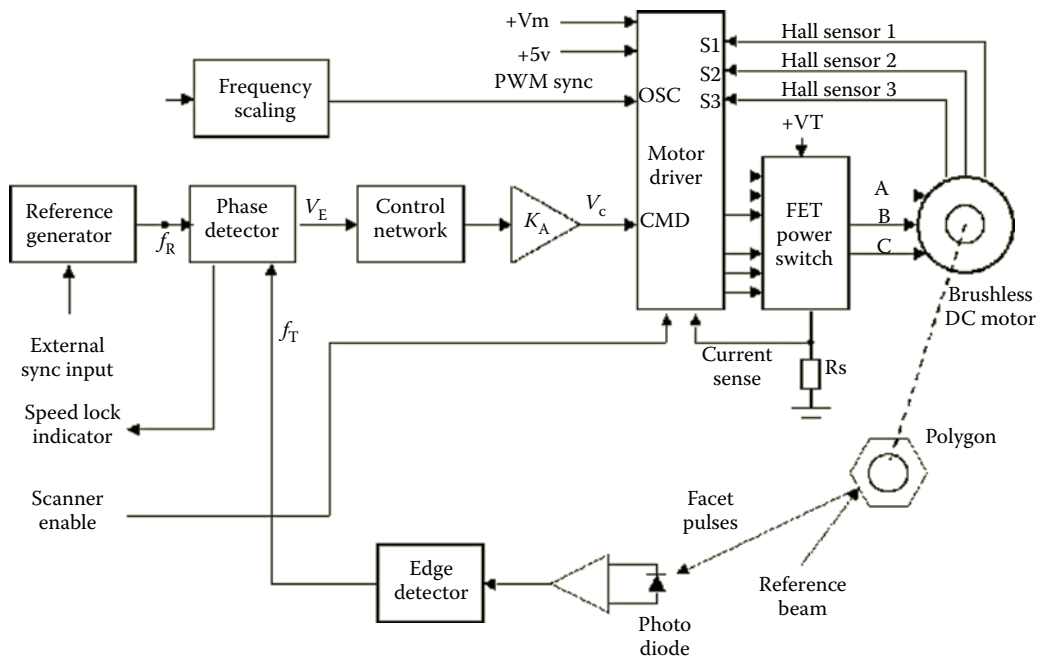
FIGURE 5.34
Scanner controller configured for high-speed polygon application.

5.7 CONCLUSIONS

The availability of low-cost brushless DC motors and drivers has made a significant impact on rotary scanner designs over the past 15 years. Advances in the motor driver area have been especially important in that the size and cost of these devices have been reduced remarkably. The power efficiency of brushless DC motors as well as the motor driver allows for high mechanical power output with minimal temperature rise. For many polygon-scanning applications, it is now practical to integrate the drive and control functions directly on the scanner or motor body.

ACKNOWLEDGMENTS

REFERENCES

1. Marshall, G.F.; Rynkowski, G.A. Eds. *Optical Scanning*; Marcel Dekker, Inc.: New York, 1991.
2. Speedring Systems Group. "Ultra precise bearings for high speed use," 102–1; "Gas bearing design considerations," 102–2; "Rotating mirror scanners," 101–1,101–2, 101–3. Technical Bulletins:Rochester Hills, MI.
3. Roters, H. C. "The hysteresis motor—advances which permit economical fractional horsepower ratings," AIEE Technical Paper 47–218, 1947.
4. Lloyd, T.C. *Electric Motors and Their Applications;* Wiley: New York, 1969.
5. Axsys Technologies Motion Control Products Division, San Diego, CA. *Brushless Motor Sourcebook;* Axsys: San Diego, CA, 1998.

6

Bearings for Rotary Scanners

Chris Gerrard

Westwind Air Bearings Division, a GSI Group company
Poole, Dorset, United Kingdom

CONTENTS

6.1 INTRODUCTION

Although rotary scanners can take a variety of forms today, the basic concept of smoothly rotating a reflective, or holographic, optic remains the same. The optic must be rotated around a defined axis with a high degree of repeatability, and within a specified speed stability. These requirements will define, in the broadest sense, the type of bearings to be selected within the scanner assembly. Other considerations will include package price, maximum speed, thermal and environmental issues, and lifetime.

Owing to the design interactions of the different components within the scanner assembly, discrete parts can rarely be designed in isolation from one another. It is necessary to understand how the dynamics of the shaft interact with the bearing system, together with the effects of additional parts such as the motor, encoder, and optic.

The object of this chapter is to examine many of the compromises and trade-offs necessary to specify, rather than design, the correct bearing/shaft system, for the machine designer with limited experience of such systems. Owing to recent advances in the design of gas bearings and the ever increasing demands on performance, this chapter will focus more on this technology rather than on ball bearing designs, although all the alternatives will be discussed. A more detailed analysis of ball bearing design criteria can be reviewed.[1]

6.2 BEARING TYPES FOR ROTARY SCANNERS

When designing a new rotating product, the traditional first choice for most designers will be some form of rolling element bearing. Readily available, easy to incorporate, and usually relatively inexpensive, it appears the ideal solution for a rotary scanner, and indeed many successful designs were used in early internal drum and flatbed image setters as well as laser printers, plotters, faxes, and photocopiers.

However, with the advent of higher resolution and productivity, machine designers have had to find alternative solutions to the traditional ball bearing assembly. The main types of bearings that can be considered will now be examined briefly.

6.2.1 Gas-Lubricated Bearings

Regarded by most now as the industry standard for high-quality scanning devices, the use of gas-lubricated, and more specifically air-lubricated, bearings has become widespread across the image setting and laser printing industries. These bearings take the form primarily of self-acting, or aerodynamic, design, generating their own internal air film between the shaft and bearing, but there are also larger designs using externally pressurized bearings requiring some form of compressor. Each type has its own advantages and these will be examined in detail later in this chapter.

6.2.2 Oil-Lubricated Bearings

The hydrodynamic oil bearing, being self-acting, could be utilized in a rotary scanner, but its main disadvantage of static oil leakage limits it to special applications only.

6.2.3 Magnetic Bearings

The use of active magnetic bearings where the shaft is supported by a strong magnetic field, is becoming a practical alternative due to enormous improvements in the electronic controls to maintain the position of the shaft within the magnetic fields. A hybrid combination of passive magnetic bearings and air-lubricated bearings is also now in use.

6.2.4 Ball Bearings

For certain less demanding applications, angular contact ball bearings are still the ideal choice, especially with the recent advances in hybrid bearings utilizing ceramic ball bearing technology, and improved grease lubricants. These products can be found in certain desktop laser printers, fax machines, and most barcode scanner systems.

6.3 BEARING SELECTION

Before examining each bearing type in some technical detail, it is proposed to compare the relative benefits and disadvantages of each technology to allow the designer to focus quickly on the correct selection and move on to the relevant section of this chapter. Figure 6.1 shows a simple comparison chart of the most likely bearing systems to be considered for a rotary scanner based on rotational accuracy, maximum speed capability, relative price, and lifetime.

From Figure 6.1, it is clear that for most high-accuracy scanner applications, air bearing technology is the most suitable, while for lower specification products, the ball bearing design would be the most cost effective. However, a more detailed comparison may be necessary if the scanner is to be operated under special conditions, which could demand other bearing systems. Table 6.1 gives a more detailed comparison, and compares the merits of self-acting against aerostatic air bearings.
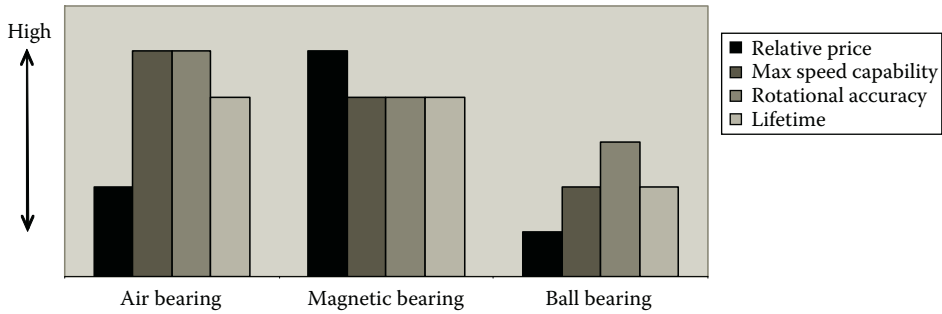
FIGURE 6.1
General comparison of bearing systems.

TABLE 6.1

Detailed Comparison of Bearing Systems

Parameter	Air bearing		Oil bearing hydrodynamic	Mag bearing active	Ball bearing ang. contact
	Self-acting	**Aerostatic**			
Accuracy of rotation	Excellent	Excellent	Good	Good	Fair
Speed: <1000 rpm	Poor	Excellent	Excellent	Good	Fair
1000–30,000 rpm	Excellent	Excellent	Fair	Good	Good
>30,000 rpm	Excellent	Excellent	Poor	Excellent	Poor
Low vibration	Excellent	Excellent	Excellent	Excellent	Fair
Shock resistance	Fair	Good	Excellent	Good	Good
Frequent stop/starts	Good	Excellent	Excellent	Excellent	Good
Low starting torque	Fair	Excellent	Good	Excellent	Good
Long lifetime (>20,000 h)	Good	Excellent	Excellent	Excellent	Poor
Wide temperature range	Good	Excellent	Fair	Excellent	Fair
Contamination to surroundings	Excellent	Good	Poor	Excellent	Poor
Resistance to dust ingress	Fair	Excellent	Good	Good	Good
High axial/radial loads	Fair	Good	Excellent	Good	Excellent
High axial/radial stiffness	Fair	Excellent	Excellent	Good	Good
Small space envelope	Good	Fair	Good	Fair	Excellent
Low heat generation	Good	Good	Poor	Excellent	Good
Run in partial vacuum	Poor	Fair	Fair	Excellent	Fair
Low running costs	Excellent	Fair	Good	Fair	Good

6.4 GAS BEARINGS

This section of the chapter examines in some detail the two main types of gas bearings; namely self acting, aerodynamic and pressure-fed aerostatic designs. Typical mechanical construction of both types will also be investigated.

6.4.1 Background

The concept of using a gas as a lubricant is a logical derivative of the study of hydrodynamic fluid film bearings. Analytical work on the characteristics of a gas-lubricated

bearing can be traced back as far as 1897 with the work of Kingsbury.[2] This was followed up in 1913 by Harrison,[3] who developed an approximate theory governing the performance of a gas-lubricated bearing, which allowed for the effects of compressibility.

At this early stage, the theory made clear that extreme accuracy would be necessary in the manufacture of a gas bearing. For this reason the concept lay dormant for the next 40 years. During the 1950s, many new fields of research were developing, most notably that of atomic energy. Nuclear reactors were being created, and the study of the radioactive environment necessitated the circulation of gas through the atomic pile. The demands on the circulators were considerable, with some power requirements being in excess of 100 HP. The original circulators designed for the purpose used conventional lubricants. Unfortunately, however, it was soon discovered that the radioactive environment caused the bearing lubricants to solidify, resulting in bearing seizure.

The failure of the gas circulators seriously hampered atomic research, and sparked an extensive search for a solution. It became clear that the only lubricant available for the circulators was the radioactive gas itself, and necessity became the mother of invention, as so often in the past.

Early work on research into gas bearings was carried out at Harwell, and the task was soon passed on to a number of major manufacturers of aero engines, these being the only companies at the time having facilities to achieve the level of accuracy required. Early results were encouraging, but following bearing seizures in many research establishments, it became clear that a bearing instability termed half-speed whirl presented a major obstacle, which had to be surmounted before desired speeds could be achieved. A number of solutions were finally devised, and circulators were built capable of handling radioactive gases at temperatures of up to 500 °C at pressures of 350 psig. One of the largest circulators was constructed by Societe Rateau for the Dragon reactor project. The pump ran at 12,000 rpm at 120HP circulating helium at 289 psig at 350 °C.

A further need for gas lubrication during this period was in the field of inertial navigation, where the replacement of miniature ball bearings resulted in a remarkable advance in accuracy of the instrument.

In the early stages of the above developments, theoretical performance characteristics gave little more than a guide, and the demands stimulated intensive theoretical and experimental research. Of the many vital theoretical contributions made, perhaps Raimondi should be singled out for his informative paper in 1961 entitled "A numerical solution for the gas lubricated full journal bearing of finite length."[4] By the use of computer-generated design charts in this paper, it proved possible to obtain excellent agreement between theory and practice for the aerodynamic bearing. Unfortunately, at this stage, half-speed whirl defied accurate prediction, and solutions relied heavily on practical experience.

In parallel with studies on aerodynamic bearings, work was also proceeding on the theoretical performance of pressure-fed bearings. This type of bearing tended to be more amenable to prediction, and again many valuable contributions were made to assist engineers in their efforts to create practical gas-lubricated bearings.

One of the earliest practical applications of the pressure-fed air bearing was in the realm of dentistry. In the 1960s a dental drill produced by Westwind Air Bearings proved very successful in the field, operating at 500,000 rpm with minimal vibration. Other applications included precision grinding and drilling spindles for the machine tool industry.

It is only within recent times that the aerodynamic bearing has come to the fore once again, this time in the field of laser scanning. The characteristics of the aerodynamic bearing are ideally suited to this particular application, which demands high-speed rotation with very low levels of vibration, and zero contamination of the environment. Figure 6.2

FIGURE 6.2
High-speed internal drum aerodynamic scanners and associated parts.

shows typical aerodynamic scanners and items used in the internal drum laser scanning market.

6.4.2 Fundamentals

There are certain fundamental characteristics of gases that explain why gas bearings are particularly suitable for high-speed rotary scanner designs.

6.4.2.1 Low Heat Generation

Compared with even the lightest instrument bearing oil, dynamic viscosity of the common gases used in gas bearing spindles are several orders of magnitude lower (Table 6.2). The main benefit of low viscosity can be seen from the equation below showing the power loss in a journal bearing.[5]

$$P_{\text{Loss}} = \frac{pmD^3Lw^2}{4c} \tag{6.1}$$

where μ = viscosity of the fluid/gas, D = shaft diameter, L = length of journal, ω = angular velocity of shaft, and c = radial clearance between the shaft and journal.

It can be readily seen how the power consumed is proportional to the lubricant viscosity, and this allows the gas bearing to be run at much higher speeds for the same shaft diameter as in an oil bearing. Figure 6.3 shows typical journal heat generation figures for several common air bearing shaft sizes. Also, the shaft diameter has a critical effect on the heat generation, due to the cubic function in Equation 6.1.

Similarly, the power loss in a thrust bearing is given by Equation 6.2:[5]

$$P_{\text{Loss}} = \frac{pmw^2}{2h}(b^4 - a^4) \tag{6.2}$$

where b = outside radius of the thrust bearing, a = inside radius of the thrust bearing, and h = axial clearance between the shaft and bearing surface.

6.4.2.2 Wide Temperature Range

Another important factor is that the variation of viscosity with temperature is small for all the gases in Table 6.2. This allows gas bearings a very wide operating temperature range, and it is the mechanical properties of the bearing and shaft materials that will usually limit the maximum temperature of operation, not the bearing itself. This limit could be the differential thermal expansion between the shaft and journal changing the clearance to an unacceptable amount or even the maximum thermal conductivity of the bearing material to transmit heat out of the bearing system.

6.4.2.3 Noncontamination of Environment

With aerodynamic bearings, the gas surrounding the bearing is used as the lubricant, usually air for common rotary scanners, and no contamination of the gas will occur (provided
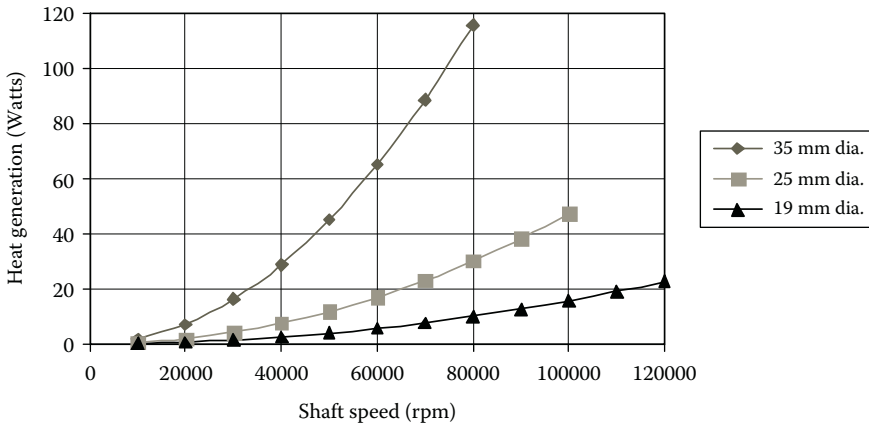
FIGURE 6.3
Air bearing journal heat generation versus shaft speed for various shaft diameters (c = 12.7 microns, L/D = 1).

TABLE 6.2

Viscosity in cP, Oil versus Several Gases

Gas/fluid	Temperature	
	27°C	**100°C**
Instrument oil	70	5.5
Argon	0.022	0.027
Air/helium	0.018	0.021
Nitrogen	0.017	0.021

the bearing materials do not chemically react with the gas). With aerostatic bearings, the pressurized gas supply (again usually air in normal scanners) will purge out of the bearing, mixing harmlessly with the environment. This has the added benefit of preventing the ingress of dust or other particles that could eventually cause damage to the shaft/bearing assembly by blocking the gap between the shaft and bearing.

6.4.2.4 Repeatability of Smoothness

As there is no physical contact between the shaft journal and the bearing during rotation, the axis of rotation, or the orbit of the shaft, will not degrade over the lifetime of the spindle, ensuring repeatable optic performance. The smoothness of rotation within one revolution of the shaft ensures minimal cross-scan errors off the optic.

6.4.2.5 Accuracy of Rotation

The shaft journal will find the average centerline of the bearing, as it is surrounded by the gaseous lubricant, which will conform to any local irregularities created during the manufacturing process, and as a general rule the shaft orbit will be an order of magnitude better than the measured roundness of the bearing in which it is revolving.

6.4.2.6 Noise and Vibration

Particularly with reference to aerodynamic bearings, audible noise is negligible from the bearing system. The main source of noise is generated by the windage of the optic. The damping properties of the gas film help ensure that transmission of any shaft vibration through to the bearing is reduced.

6.4.3 Aerostatic Bearings

A constant supply of pressurized gas must be supplied to both the radial and axial bearing gaps to support the shaft load with aerostatic bearings. Although the lift-off, or float pressure, will be at a very low pressure, to achieve useful loads and stiffness a pressure in the order of 3–6 bar is normally used.

This requires the use of an external compressor, which is a big disadvantage in many rotary scanner applications as the rest of the scanning system will not usually require compressed air, and additional noise, vibration, and cost associated with the compressor could be prohibitive. However, particularly when rotating large, overhung optics, the benefits of high radial and axial loads may justify the use of aerostatic bearings, particularly if very low speed operation is required (such as a few hundred revolutions per minute).

Another benefit with aerostatic bearings is that the shaft can be rotated in either direction with identical performance, something not normally possible with aerodynamic bearings. This feature could be of use where one aerostatic spindle is to be used in a variety of scanning products, which could well run in different directions. Finally, although the exhausting of gas through the end of the bearings creates additional noise, the self-cleaning effect could be useful to prevent particles created in the scanning machine (particularly paper and carbon) from being deposited inside the scanner.

The general design principles will now be examined, followed by some general notes on construction.

6.4.3.1 Aerostatic Journal Bearing

The general principle of operation can be explained by reference to Figure 6.4. The bearing consists of an annular cylinder containing two sets of orifices, or jets, one row towards each end of the bearing.

The jets are supplied with gas at pressure P_s and the gas exhausts at P_a to atmosphere. With no load on the shaft (and ignoring its own mass) the downstream pressure in the gap between the shaft and the bearing is equal all round any circumference as shown in the cross section of the bearing through one of the jet planes. The associated pressure profile diagram along the bearing shows how the discharge pressure P_d slowly drops as the gas flows towards the ends of the bearing until it exhausts at the atmospheric pressure P_a. In other words, there is a constant flow of gas between the jet plane and the end of the bearing, while the area between the jet planes remains at a constant pressure.

When a radial load is applied to the shaft, it will be displaced in the direction of the force, reducing the gap between the shaft and the bearing. The localized gas flow will reduce,
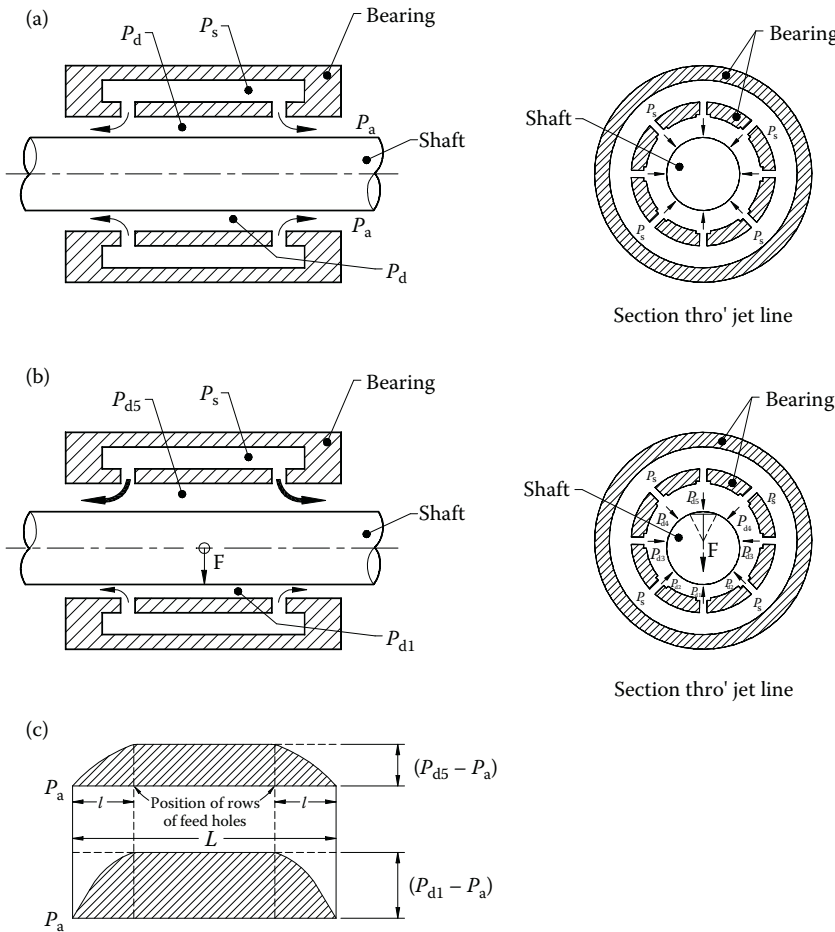
FIGURE 6.4
Aerostatic journal bearing operation: (a) with no load, (b) with load and (c) pressure profile with load.

causing an increase in pressure ($P_{d1} \boxtimes P_a$), with a similar reduction in pressure ($P_{d5} \boxtimes P_a$) due to an increase in flow on the other side of the shaft. This resultant pressure difference across the shaft will cause it to resist the applied load, preventing surface contact between the two parts. When the load is removed, the pressure distribution will return the shaft to the central position again.

In practice, due to the relatively small number of jets per row, typically 8 to 12, dispersion effects reduce the effective pressure zone between the rows of jets. This will reduce the load capacity slightly, as will circumferential flow around the bearing from the high pressure to the low pressure zone. Although there are other methods of feeding the gas into the bearings, such as slot feeds or the use of porous materials, the discrete jet orifice method has become the favorite for this market.

6.4.3.1.1 Load Capacity

The standard equation for expressing the radial load capacity of an aerostatic journal bearing is:

$$\text{Load } W = C_L (P_s - P_a) L \times D$$

where P_s = supply pressure, P_a = ambient pressure, L = bearing length, D = bearing diameter, and C_L = dimensionless load coefficient. The load coefficient C_L is affected by several different parameters, including the eccentricity ratio, the downstream pressure P_d, the number of jets (the dispersion effect), and the jet position in relation to the end of the bearing. There are ways of estimating these effects as shown by Shires.[6]

To be able to estimate the actual maximum load capacity, the designer must decide how close to the bearing surface the shaft can be moved before local irregularities in the bearing or shaft cause actual contact; that is, balance, roundness of the shaft, ovality of the bearing, and squareness of the shaft to the bearing all have an effect on this decision. This displacement is usually referred to as the eccentricity ratio ε of the shaft in the bearing and a typical maximum figure while rotating would be 0.5. That is, the shaft has been displaced by half the total radial clearance. Figure 6.5 shows typical load capacities for shaft diameters used in rotary scanners, with two different bearing length-to-diameter ratios: using $\varepsilon = 0.5$, jet position = $0.25 \times L$, $P_s = 5.5$ bar g.
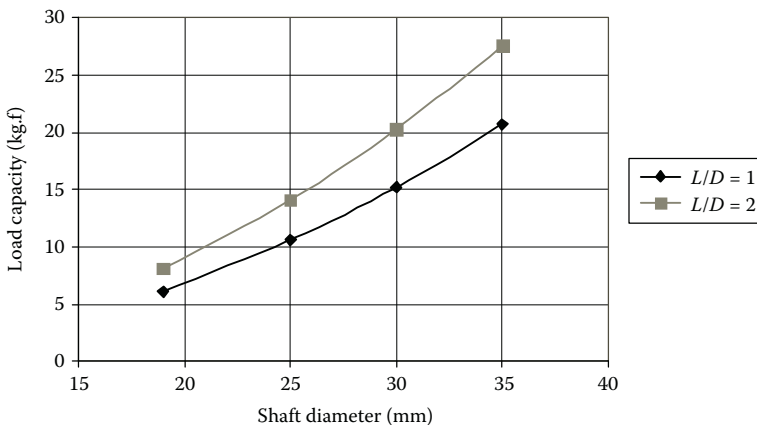
FIGURE 6.5
Radial load capacity versus shaft diameter ($P_s = 5.5$ barg, $\varepsilon = 0.5$).

6.4.3.1.2 Radial Stiffness

Radial bearing stiffness is constant at low eccentricity ratios, and can readily be derived from the following equation.

$$\text{Stiffness } K = \frac{W_e}{e \times C_0}$$

where W_e is the load capacity at $\varepsilon = 0.1$, $\varepsilon = 0.1$, and C_0 = radial clearance between the shaft and bearing. Figure 6.6 shows the effect of clearance variation on radial stiffness for various shaft sizes.

6.4.3.1.3 Heat Generation

At first glance the designer would want to keep the radial clearance as small as possible to ensure maximum stiffness, but the trade-off is that bearing heat generation is inversely proportional to the clearance as shown in Equation 6.1, and a compromise has to be reached between these two factors. Figure 6.7 shows bearing heat generation plotted against clearance for the common sizes of shaft.
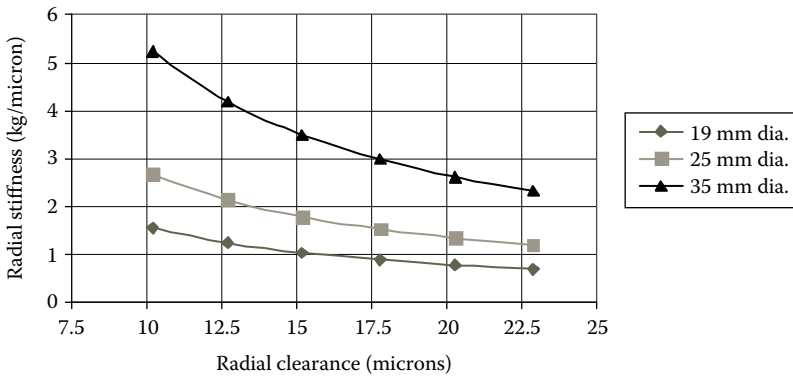
FIGURE 6.6
Radial stiffness versus clearance for various shaft diameters ($\varepsilon = 0.1$, $L/D = 1$, $P_s = 5.5$ bar g).
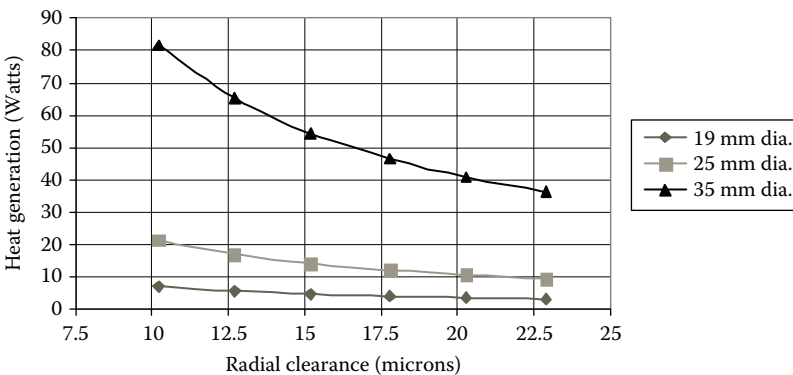
FIGURE 6.7
Journal bearing heat generation versus radial clearance (60,000 rpm, $L/D = 1$).

Dependent upon the construction of the spindle, there will be a critical limit on the heat generation per square centimeter of bearing surface above which liquid cooling will be necessary to maintain the correct clearance between the shaft and bearing (due to thermal expansion).

6.4.3.1.4 Bearing Gas Flow

Before the air flow can be calculated, another design factor that must be considered is the shape of the jet orifice itself. There are two common forms of discrete jet: the plain jet and the pocketed jet. Figure 6.8 shows a simplified form of both types.

With the plain jet the smallest flow area is controlled by the bearing radial clearance *c* and hence the area used for flow calculations is the surface area of a hollow cylinder with the length equal to the radial clearance *c.*

$$A = \pi dc$$

However, with the pocketed jet (or simple orifice) the smallest flow area is controlled by the jet diameter itself, hence the area for flow calculations is the cross-sectional area of the jet itself.

$$A = \frac{pd^2}{4}$$

Obviously, if very large clearances are used, which is rare, even a plain jet will run as a simple orifice.

For better control of flow, pocketed jets are preferable and yield higher stiffness due to reduced dispersion effects, but plain jets provide greater damping reducing the likelihood of instability. This instability or resonance can occur if the pocket volume is too large for the bearing design and a self-induced pneumatic hammer can be heard.

Figure 6.9 shows typical air flow for a 25-mm shaft for both pocketed jets and plain jets with a diameter of 0.16 mm.

6.4.3.2 Aerostatic Thrust Bearing

An aerostatic thrust, or axial bearing system consists of two opposed circular thrust plates sandwiching the shaft axial runner with the gap between the surfaces being controlled by
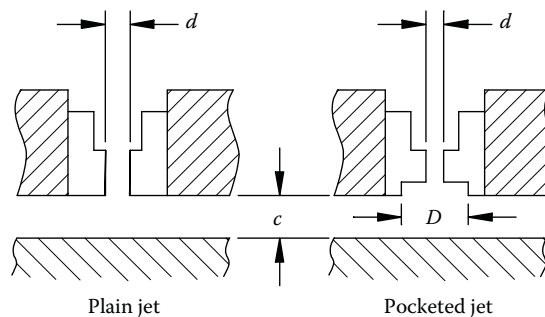
FIGURE 6.8
Common types of jet orifice.
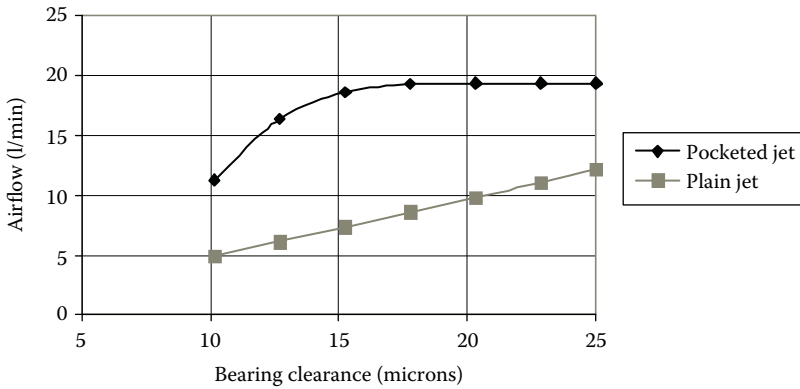
FIGURE 6.9
Air flow versus bearing clearance for a 25-mm diameter bearing ($L/D = 1$, $P_s = 5.5$ bar g, 16 jets, $d = 0.16$ mm dia).
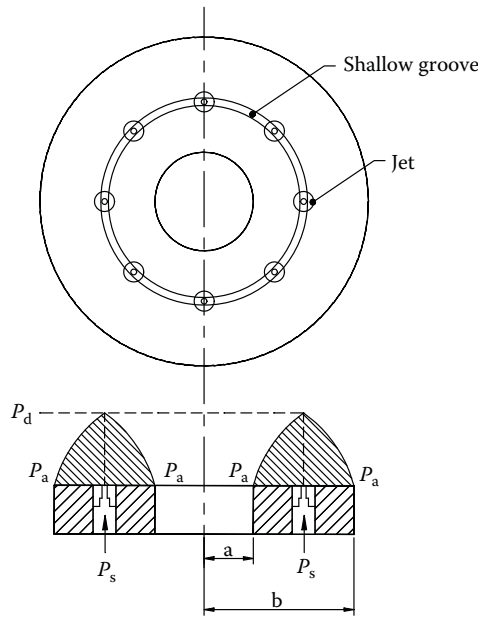
FIGURE 6.10
Aerostatic thrust bearing.

a spacer, slightly thicker than the shaft runner, which is located around the outside diameter of the shaft.

Figure 6.10 shows a single thrust plate fitted with an annular row of discrete jets, linked by a narrow groove. The purpose of the groove is to create a pressure ring around the jet pitch circle diameter (PCD) for optimum performance, particularly when the shaft runner is almost at touchdown condition on the thrust face. The associated pressure profile is also shown.

For stability, two thrust plates are used in opposition, trapping the shaft runner between. In a similar way to the journal bearing mechanism, when a load is applied to the shaft

axially the shaft will move towards one face of the axial bearing and the flow through the jets will drop, causing an increase in pressure over that bearing face. This will create an opposing force on the shaft runner, preventing it from moving closer to the bearing surface. Meanwhile, the opposite happens on the other bearing face, reducing the force on the other shaft runner face.

This mechanism can be seen more clearly with reference to Figure 6.11a, which shows the load capacity lines of both faces and where they cross will become the equilibrium position of the shaft runner. Other forms of axial bearing such as the center fed or journal fed are possible but the annular ring of jets is the most suitable for this market.

6.4.3.2.1 Load Capacity

The load capacity from one plate can be expressed by the following equation.[5]

$$W = \frac{(P_d - P_a)p(b - a)^2}{\ln(b/a)}$$
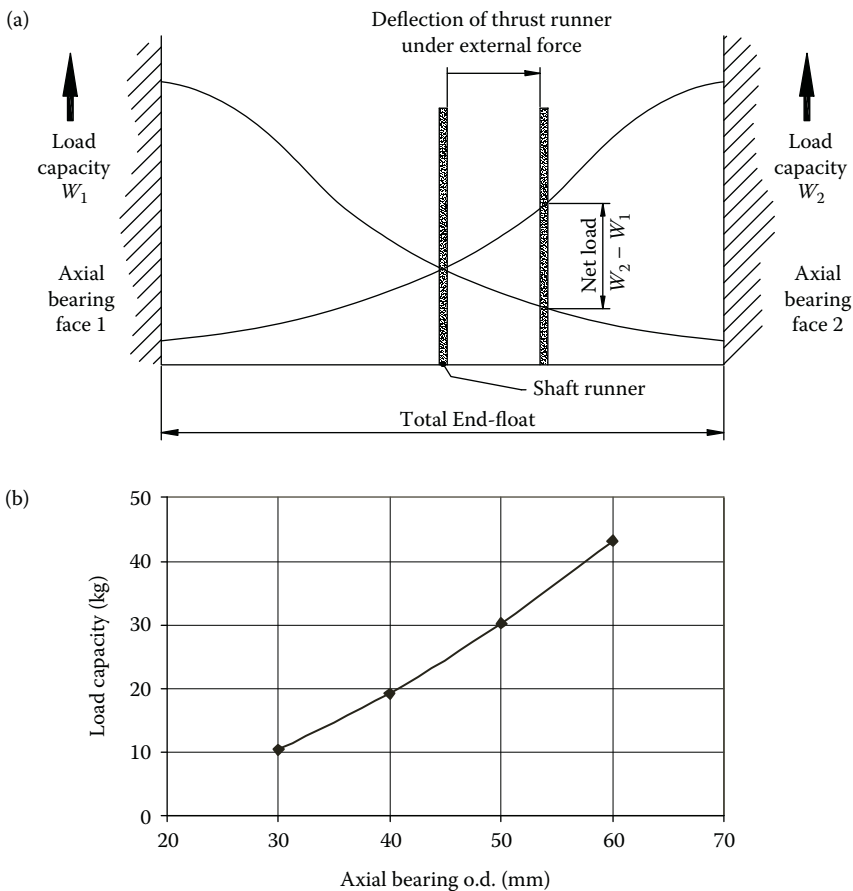
FIGURE 6.11
(a) Axial bearing diagram. (b) Axial load capacity versus bearing outside diameter (b/a = 1.6, P_s = 5.5 bar g, 8 jets, d = 0.27 mm dia).

where P_d = downstream pressure, P_a = ambient pressure, b = outside radius of plate, and a = inside radius of plate. Estimation of the downstream pressure P_d is based on a number of factors, which are beyond the scope of this chapter (see Reference 5). Again, with reference to Figure 6.11a, it can be seen that the maximum load capacity of an opposed thrust plate assembly is the summation of ($W_2 \boxtimes W_1$), at the point where the runner is approaching contact with thrust face 2.

Figure 6.11b shows some typical load values used in axial bearings for scanner applications, with a constant ratio between the outer radius and the inner radius of 1.6.

6.4.3.2.2 Axial Stiffness

Axial stiffness is calculated using the equation

$$K = \frac{W_2 - W_1}{d}$$

where W_2 = load capacity of face 2 for the position of (equilibrium position $-\delta$), W_1 = load capacity of face 2 for the position of (equilibrium position $+\delta$), and δ = distance moved for applied external load. Note, for maximum stiffness, K is normally calculated for a δ of <10% of the total endfloat.

It can be seen from Figure 6.12, a plot of axial stiffness against axial clearance for various bearing diameters, that the centerline (or equilibrium position) stiffness is highly dependent upon the clearance between the shaft runner and the bearing.

To achieve optimum stiffness, the number of jets, the radius of the jets, and the dimension of the groove must all be considered. For minimum air flow, and highest stiffness, a small clearance is desirable, but again heat generation will be highest at this clearance.

6.4.3.2.3 Heat Generation

Using Equation 6.2, the effect of small clearances on axial bearing heat generation is shown in Figure 6.13.

6.4.3.3 Aerostatic Scanner Construction

A typical aerostatic bearing polygon scanner is shown in Figure 6.14. The polygon is attached to a removable threaded mount in the front of the shaft and to help counterbalance
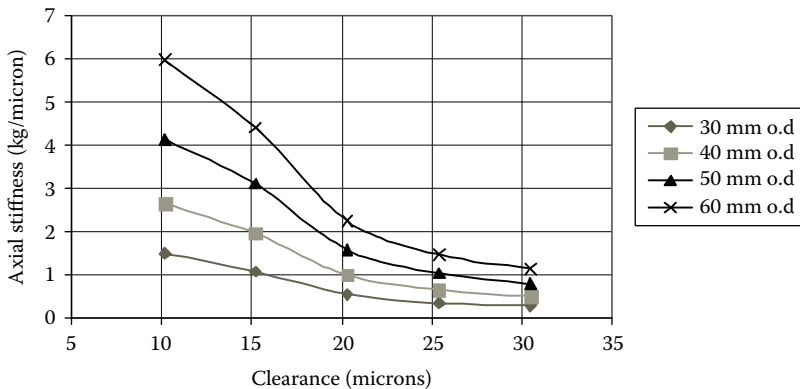
FIGURE 6.12
Axial stiffness versus clearance for various axial bearing outside diameters (a/b = 1.6, P_s = 5.5 bar g).
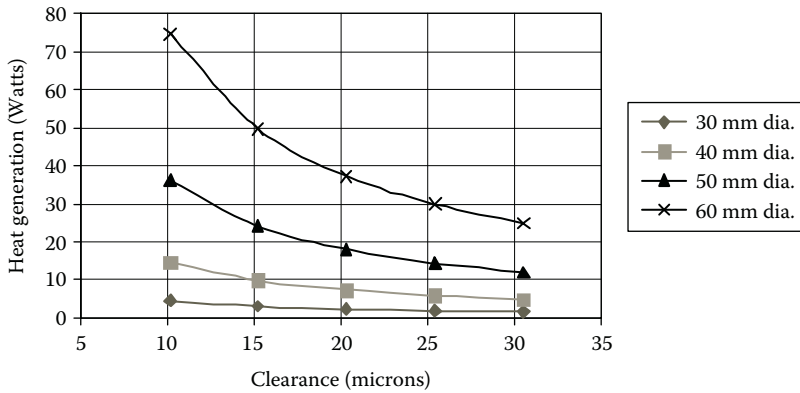
FIGURE 6.13
Heat generation versus clearance for various axial bearing outside diameters (60,000 rpm, b/a = 1.6).

Speed	: 5000–7500 rpm
Shaft dia	: 32 mm
Air supply pressure	: 5 Bar g
Air consumption	: 15 l/min
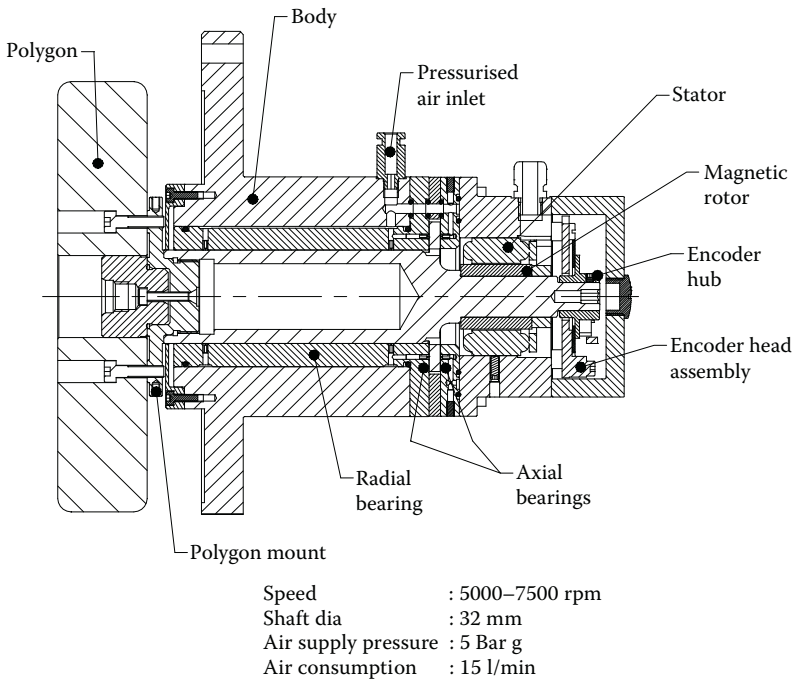
FIGURE 6.14
Aerostatic bearing polygon scanner with speed 5000–7500 rpm, shaft diameter 32 mm, P_s = 5 bar g, and air consumption 15L/min.

the mass of the polygon the axial bearing system is located at the rear of the shaft, adjacent to the brushless DC motor. To minimize air flow, a long single radial bearing design has been used with only two rows of jets. For maximum radial stiffness and load capacity, a four-jet row, twin bearing system could be used at the expense of higher-air flow. An optical encoder system is fitted at the rear of shaft to guarantee high-accuracy speed control. The bearing materials would typically be bronze or gunmetal, while the shaft itself would be stainless steel.

6.4.4 Aerodynamic Bearings

In recent years, the use of aerodynamic or self-acting bearings has become more widespread, replacing ball bearing or aerostatic bearing assemblies in this market. In general, machining tolerances are much smaller for aerodynamic bearings and it is only with the recent advent of higher precision computer numerical control (CNC) machines that it has become more cost effective to introduce this technology for high-volume manufacturing.

In its simplest form, the aerodynamic bearing consists of a plain circular tube in which the shaft rotates as shown in Figure 6.15. If a load W is imposed onto the shaft as shown, causing the shaft to move off center by an amount εh_0, the pressure will rise in the reduced gap due to viscous shear of the gas, creating a "wedge" similar to the mechanism occurring in a hydrodynamic oil bearing. This high-pressure zone allows the shaft to float so it can rotate without contact in the bearing.

Owing to the low viscosity of gases, the clearance between the shaft journal must be very small, typically a few microns for this effect to be useful. Obviously, when the shaft is stationary, there is no viscous shear, or supporting pressure, and the shaft journal and bearing will be in contact. To avoid damage to both surfaces when starting to rotate the shaft a high-torque motor is required to accelerate the shaft to floating speed in a very short time. For shaft diameters utilized in scanner assemblies, this speed is typically several hundred rpm.

Obviously gases are compressible, and this will reduce the effect of the pressure wedge compared with a liquid, and the resulting load calculations can be complex. However, it can be shown that the compressibility number Λ can be used as a guidance of bearing performance.[6]

$$\frac{\Lambda}{6} = \frac{mw\,[r]^2}{P_a\,[c]^2} \tag{6.3}$$

where μ = viscosity, ω = shaft angular speed, P_a = ambient pressure, r = shaft radius, and c = bearing clearance.
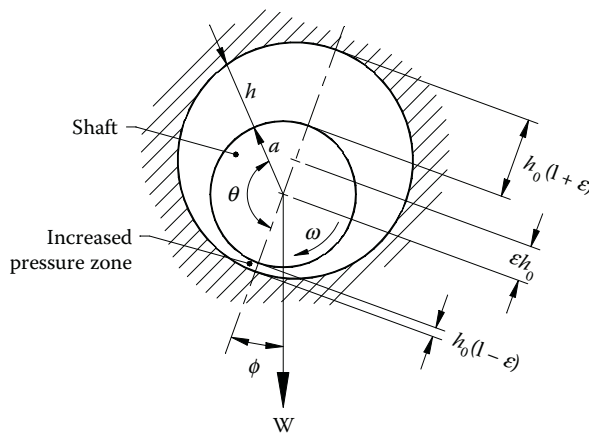
FIGURE 6.15
Aerodynamic journal bearing.
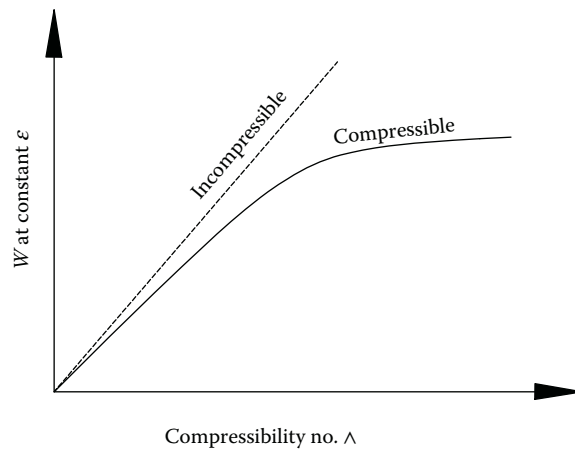
FIGURE 6.16
Variation of load capacity with compressibility number.

Figure 6.16 shows how the load capacity at constant eccentricity ε for compressible and incompressible fluids varies with increasing compressibility number Λ. It can be seen that for gases at high compressibility numbers, the load capacity becomes independent of this number. Practically speaking, this means that there comes a point when the radial load capacity (and stiffness) become independent of the speed of the shaft.

Also from Figure 6.15, it can be seen that when load W is applied to the shaft, the closest approach between the journal and bearing is not in direct opposition to W, but at an attitude angle ϕ. This angle can theoretically vary between zero and 90° and is mainly dependent on the compressibility factor.

Load Capacity   Estimating the load capacity of an aerodynamic bearing is more complex than for the aerostatic version, due to the compressibility effects. However, Raimondi[4] managed to compute numerical solutions and create design charts using the dimensionless group:

$$\text{load ratio } \frac{P}{P_a} = \frac{W}{2rLP_a}$$

where P = bearing pressure, P_a = ambient pressure, W = load capacity, r = shaft radius, and L = bearing length, and the compressibility number Λ (Equation 6.3) for various eccentricity ratios. Figure 6.17 shows the Raimondi graph[4] for a bearing L/D ratio of 2, probably the minimum number for practical aerodynamic bearings. Having calculated the compressibility number, and decided what is the maximum practical eccentricity ratio, the load capacity can be derived.

Unfortunately, from the explanation regarding load capacity, it is obvious that if no load is applied to the shaft in an aerodynamic bearing, then there is no wedge action and without the associated pressure effect, the shaft is essentially unstable. To overcome this instability, the designer must incorporate some kind of surface form into the bearing of the shaft that creates a wedge effect without the use of a load. Obviously, the shaft mass itself creates a small eccentricity if the shaft is running horizontally, but often it is not. Also with

the correct surface form design, the effect of load at constant eccentricity becoming independent of speed at high compressibility numbers can be dramatically reduced.

There are two main surface forms used in aerodynamic scanner bearings; spiral grooves and lobing.

6.4.4.1 Spiral Groove Bearings

A series of shallow spiral grooves can be machined into either surface, but usually in the shaft journal, which are open to the atmosphere at one end. Gas is drawn into these grooves by viscous shear during rotation and creates a pressurized zone towards the closed end (Figure 6.18). The important journal groove geometric parameters are: groove angle, α; no. of grooves, N; groove depth ratio, h_0/h; groove length ratio, L_g/L; and groove width ratio, $W_1/(W_1 + W_2)$.

The compressibility number from Equation 6.3 can still be used and from using the appropriate graphs[7] the optimized geometry can be deduced.
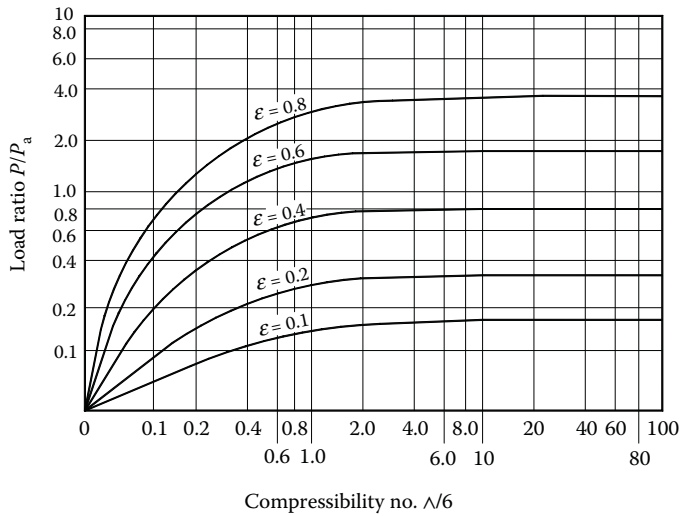
FIGURE 6.17
Load capacity versus compressibility number for a bearing of $L/D = 2$.
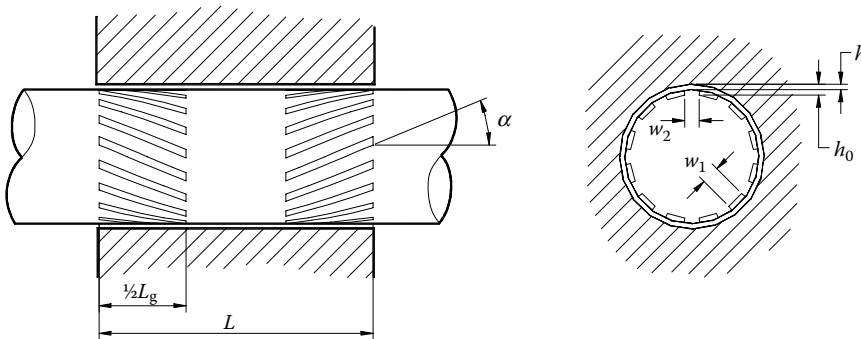
FIGURE 6.18
Spiral grooved shaft.

The same concept can be applied to the axial annular bearings with a series of shallow grooves machined into one of the surfaces, which spiral in towards the center, again creating a pressurized zone at the closed inner ends. Figure 6.19 shows a typical example with the pressure profile. The important thrust groove geometric parameters are: radius ratio, $(r_b - r_i)/(r_b - r_a)$; width ratio, W_1/W_2; depth ratio, h_0/h; and groove angle, θ. Practical values of these parameters have been established by Whitley and Williams[8] and from these an estimation of the load capacity and stiffness can be made.

There are three basic bearing configurations that can be utilized with spiral groove technology: separate parallel radial and axial bearings, conical bearings, or spherical bearings, as Figure 6.20 shows.

Both the conic and the spherical bearing designs have the advantage of only one set of grooves, which makes for a compact design, especially if the motor can be placed between the bearings, but creating the bearing surfaces precisely is quite a production challenge, particularly for the hemispherical type.

6.4.4.2 Lobed Bearings/Shaft

Another form of aerodynamic geometry is the lobed bearing or shaft, which has an out-of-round surface that generates an axial increased pressure zone, or zones, along the length of the bearing as it rotates. Figure 6.21 shows a typical bearing form, which has three lobes and also a stabilizing groove that creates a small shaft eccentricity for stability (Westwind Patent No. EP0705393[9]).
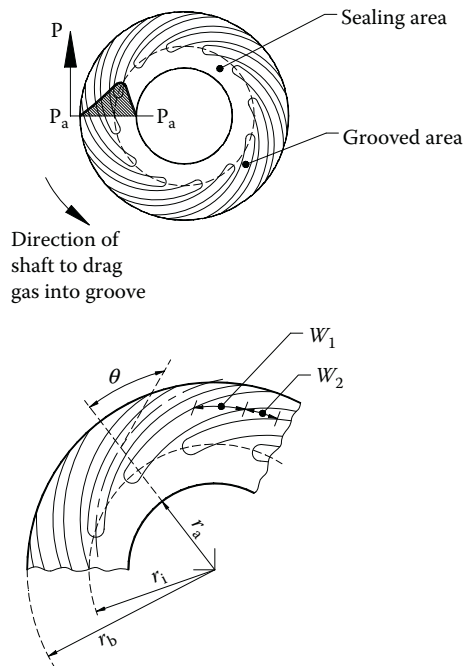
FIGURE 6.19
Axial bearing plate.
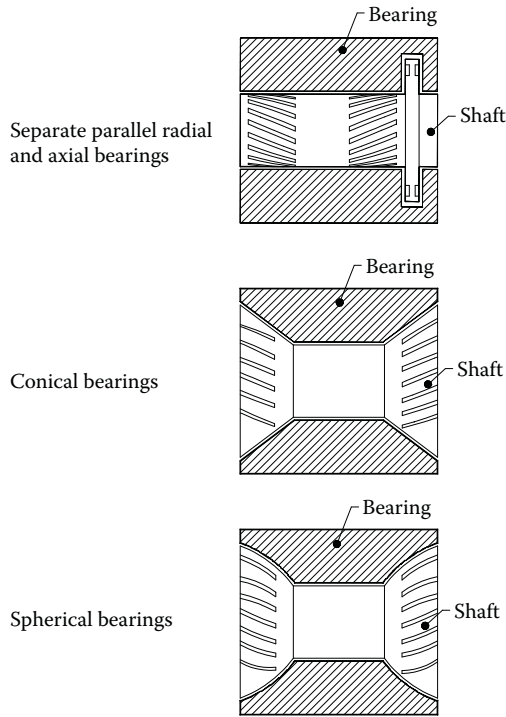
FIGURE 6.20
Spiral groove bearing types.
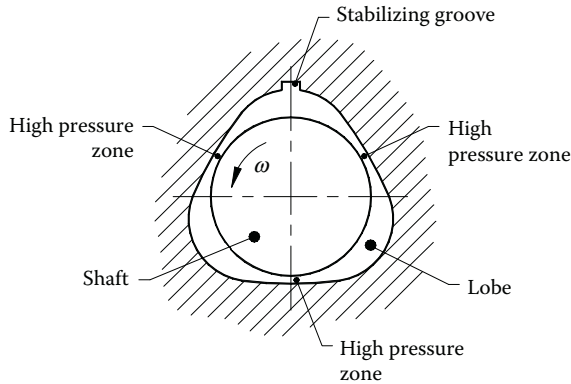
FIGURE 6.21
Lobed bearing.

This design has the advantage of simple manufacture in production volumes with stable performance over a wide speed range, and is ideally suited to supporting overhung optics due to the large bearing centerline separation.

Alternatively, the shaft could be lobed, and with the latest CNC cam grinding machines this form can be machined relatively easily. However, some form of spiral groove axial bearing will still be required on either design.

Speed   : 5000 rpm
Polygon size   : 126 mm a/f
Shaft dia   : 32 mm
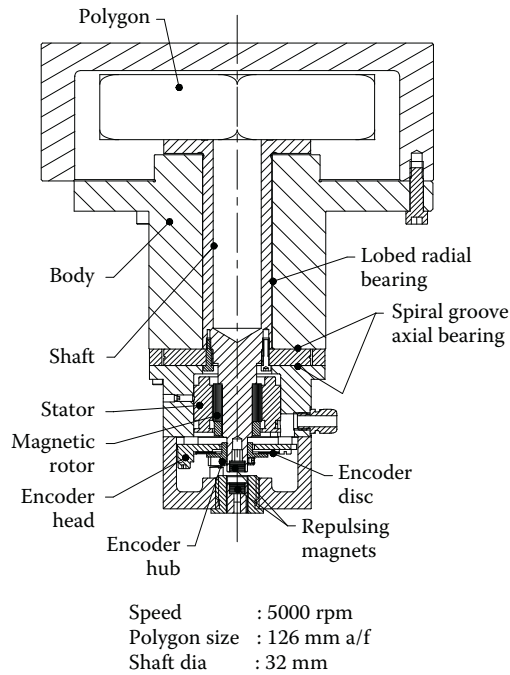
FIGURE 6.22
Aerodynamic bearing polygon scanner.

6.4.4.3 Spindle Construction

A typical overhung polygon scanner unit is shown in Figure 6.22 using lobed radial bearing and spiral groove axial bearing technologies. To help balance out the mass of the polygon on the front of the shaft, the axial bearings, together with the motor and encoder, have been located at the rear. For vertical operation, additional repulsing magnets are contained at the rear of the spindle, one in the shaft and one in the housing, to provide additional upwards force to reduce starting frictional torque on the lower thrust plate.

To allow many thousands of stop/start cycles, bearing surfaces need to be coated with some form of antiscuff, low-friction material, probably containing PTFE (Teflon®). The choice of shaft and bearing materials is important to ensure that the correct bearing clearances are maintained over a wide range of temperatures.

For center-mounted polygon scanners, a conical bearing design with spiral groove technology is probably the most practical due to its compact shape, with the motor mounted between the bearings as shown in Figure 6.23. In this case, the rotor rotates around the static central stator.

For many applications today, a monogon optic design is required, which by its very nature must be mounted onto the front of the shaft. A typical scanner cross section is shown in Figure 6.24 using a lobed radial bearing and spiral groove axial bearing technology. Note that the axial bearing is now at the front of the spindle, close to the optic. This minimizes the forward axial growth of the shaft due to heat generation within the radial bearing system as monogon scanning systems are sensitive to axial movements (which can misplace the output beam).
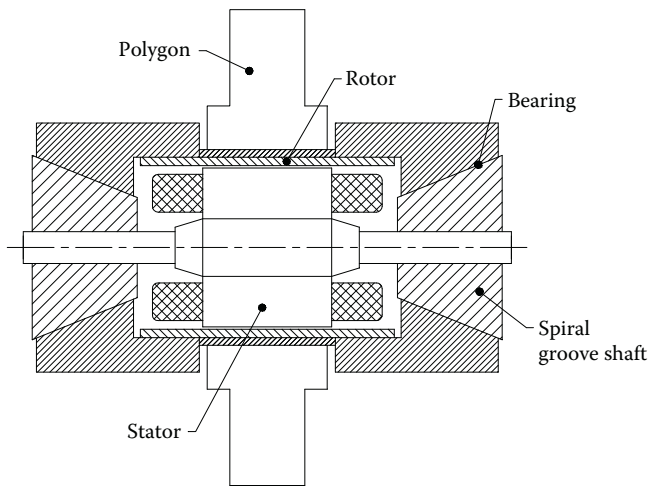
FIGURE 6.23
Scanner with conical gas bearing.
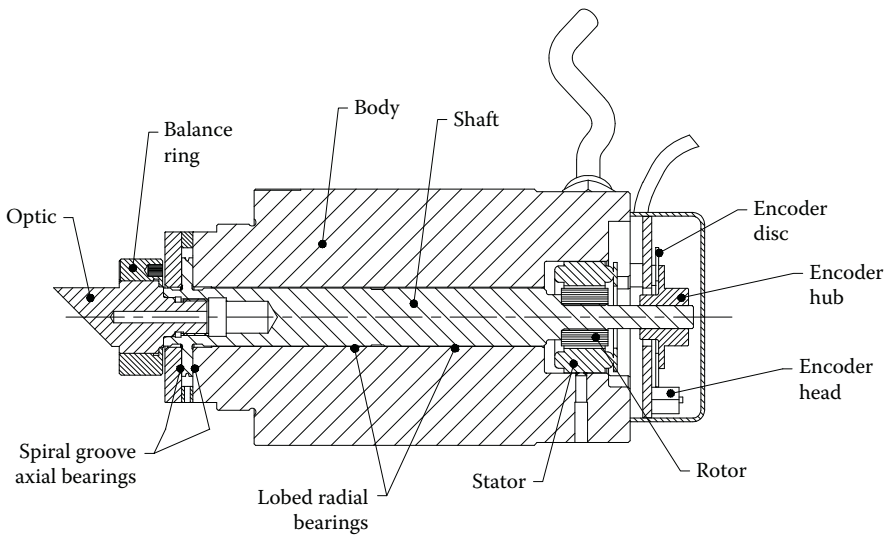
Speed   : 30,000 rpm
Optic dia : 14 mm
Shaft dia : 16 mm

FIGURE 6.24
Aerodynamic bearing monogon scanner.

6.4.5 Hybrid Gas Bearings

There is a special form of radial bearing that combines the technologies of both aerostatic and aerodynamic design. If the spacing between the two jet rows is increased significantly in an aerostatic design, a large increase in stiffness and load capacity will be achieved at high shaft rotational speed. Typically a length-to-diameter ratio of 2 would be considered ideal, with the jets spaced about one-eighth of the bearing lengths from the bearing ends,
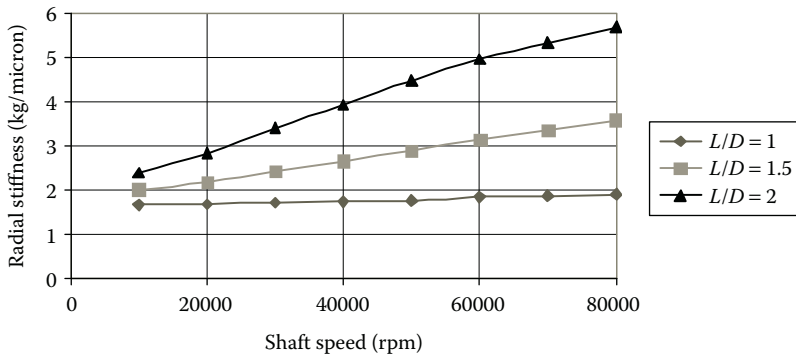
FIGURE 6.25
Hybrid radial bearing stiffness versus shaft speed for a 25-mm shaft dia ($\varepsilon = 0.1$, $P_s = 5.5$ bar g, $c = 12.7$ microns).

but to achieve significant hybrid performance, the bearing clearance must be kept to a minimum.

By utilizing the Raimondi curves, as mentioned previously for aerodynamic performance, the additional load capacity can be calculated and summated into the aerostatic load capacity calculations. From Figure 6.25, the improvement can be seen in the radial bearing stiffness with three different bearing lengths as the speed increase (assuming constant bearing clearance over the speed range). This has the great benefit of increasing the gas film critical speeds and hence the maximum operating speed of the spindle. This will be discussed under Section 4.6 "Bearing and Shaft Dynamics."

6.4.6 Bearing and Shaft Dynamics

Whichever gas bearing system is selected, certain bearing and shaft dynamics have to be considered to allow high-speed operation, and specifically the synchronous whirls, the half-speed whirl, and the shaft natural frequency.

6.4.6.1 *Synchronous Whirls*

Synchronous whirls occur at the shaft rotation speed and can be due to an inherent imbalance in the shaft itself, which will increase quickly with speed as the out-of-balance forces are proportional to the square of the speed. Therefore careful dynamic balancing is necessary to minimize these forces. Typically a balance standard of better than G0.4 (International Standard ISO 1940) should be achieved for acceptable performance, involving a two-plane balancing methodology (see Section 6.4.6.4 for more details on balancing).

However, another phenomenon is encountered in aerostatic bearings, which is due to the natural resonant frequencies of the gas film system. As these frequencies are approached, any out-of-balance force is magnified dramatically due to the almost total lack of damping. However, the shaft can be run through these frequencies and operate very comfortably in a "supercritical" mode, with the shaft now rotating around its mass center, and not its geometric center.

The speeds at which the natural resonant frequencies occur can be calculated as shown below, ω_1 being defined as a cylindrical whirl and ω_2 as a conical whirl mode.

$$w_1^2 = \frac{2k}{m}$$

where k = gas film stiffness, and m = mass of the shaft.

$$w_2^2 = \frac{2kJ^2}{I - I_0}$$

where J = (distance between bearing centers/2), I = transverse moment of inertia of the shaft, and I_0 = polar moment of inertia of the shaft. Obviously, for maximum shaft speed, the designer needs to achieve the highest radial stiffness possible and keep the shaft mass at a minimum.

6.4.6.2 Half-Speed Whirl

This highly destructive phenomenon is encountered in both aerostatic and aerodynamic designs, and is usually the limiting speed parameter. In aerostatic designs, it occurs in practice at a speed somewhat below twice the lowest of the gas film resonant frequencies, typically at about 1.8 times. In aerodynamic designs it is much harder to predict the speed at which it will occur, but in general, spiral groove bearing designs do not suffer from this problem nearly as much. Half-speed whirl occurs when the rotor is orbiting the bearing centerline at a frequency equal to half its rotational speed. The shaft increases its orbit without a further increase in speed, and quickly contacts the bearing surface, causing seizure.

6.4.6.3 Shaft Natural Frequency

Like any other bearing system, the shaft natural frequency must be calculated using the normal processes. It must include the mass effect of additional items like the rotor, encoder disc, and, most importantly, the optic and its mount.

In small shaft scanner designs, the natural frequency is usually well above the operating speed range, maybe by a factor of 2 to 3, but with large shafts this frequency must be allowed for. In general, no shaft should be run beyond 80% of the shaft natural frequency.

The shaft critical speed can be calculated from the general formula for a uniform shaft (simply supported on short bearings):

$$\text{Angular velocity } w_{\text{crit}} = \frac{p^2}{l^2}\sqrt{\frac{EI}{m}}$$

where l = distance between bearings, E = modulus of elasticity, I = movement of inertia, and m = mass of the shaft. From this equation it can be seen that to keep the shaft mass and the bearing separation distance to a minimum is really important for high-speed operation.

6.4.6.4 Shaft Balance

For successful high-speed spindle operation, dynamic balancing of the shaft and associated optic assembly is essential. However precisely the components are manufactured, there will always be some small level of residual unbalance around the axis of rotation

which must be corrected. This can be achieved by the addition, or subtraction, of material at the appropriate position on the shaft or optic assembly:

1. Addition process: Usually this is in the form of adding screws of known mass to selected premachined threads in non-operating portions of the shaft, or the optic assembly. Alternatively, in some automatic balancing processes, a precise amount of fast-curing adhesive can be added to the required position in the shaft.

2. Removal process: This entails the physical removal of material from the shaft, or optic assembly, and can be achieved using a small grinding wheel or drill.

Both methods have advantages dependent upon the application, but with all balancing techniques the process will often be a compromise as it is not always possible to add, or remove, material at the exact point of unbalance as this may well be in an operating part of the shaft (journal, motor rotor, etc.) or in the optic surface itself. In this situation, the nearest convenient plane must be selected, and any couple induced by this must be corrected for.

Highly sensitive balance equipment must be used to detect and display the residual imbalance and Figure 6.26a shows a typical two-plane balancing cradle with an aerodynamic monogon scanner fitted, ready for final balancing on the optic assembly (removal method). Each of the clamp assemblies supporting the scanner contains a vibration transducer, which feeds a signal back to the analysis equipment shown in Figure 6.26b. A shaft position sensor is also fitted to the cradle so that the angular position of the unbalance can be calculated. The equipment will display the level of unbalance in gm mm (mass × radius) and the angular position in degrees from a reference point on the shaft in both planes. The rotational speed used for the balancing trials may well be much lower than the maximum operating speed due to the response capability of the balancing equipment, and the need to find a "sweet spot" between the resonant frequencies of the transducers and clamp assemblies themselves. This could well be in the range of 10,000 to 20,000 rpm.

6.4.7 Shaft Assembly

Although this chapter is primarily concerned with the different bearing types for use with scanners, it is important to realize that all the extra components mounted onto the shaft
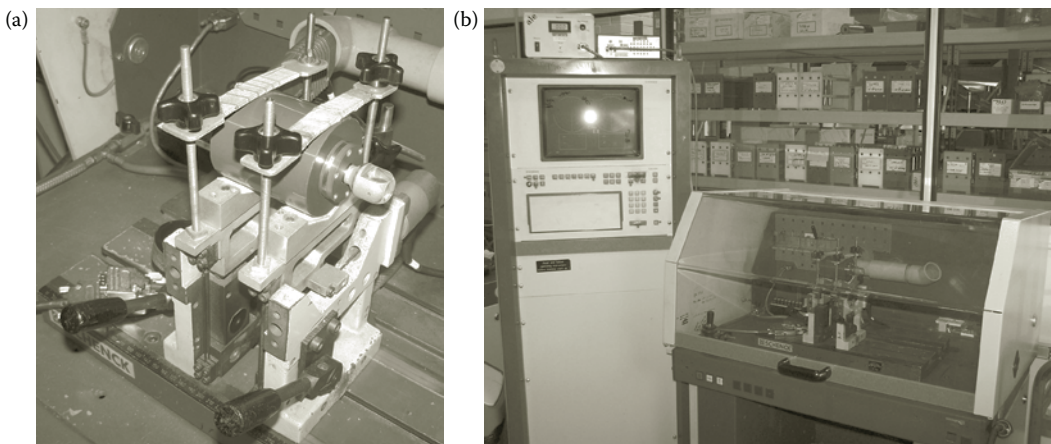
FIGURE 6.26
(a) Two plane balancing cradle; (b) Commercial balancing machine.

will have an impact on the performance of the scanner. As shown in the previous sections, the total mass of the rotor assembly affects both shaft natural frequency and the cylindrical whirl frequency, hence any additional masses built in the rotor must be minimized. Another consideration has to be the stresses that the shaft is imposing on these additional components when rotating at high speed.

6.4.7.1 Optics and Holders

The types of optics that can be fitted to rotary scanners fall into three basic groups: polygons, monogons, and holographic disks. All these optics are selected for specific applications and are discussed in much greater depth elsewhere in this book. However, the bearing design type of the scanner will in general be dictated by this optic selection.

6.4.7.1.1 Polygons

As shown in Section 4.4.3, polygons can be mounted on the front of the spindle (Figure 6.22) or in the center of the bearing system (Figure 6.23). From a bearing stability viewpoint, the center-mounted conic bearing design is better suited for small, high-speed polygons (<100 mm across flats) although the large bore size required to fit over the scanner body may cause optical distortion problems.

The polygon can also suffer from thermal distortion due to the close proximity of the motor rotor. Its compact space envelope is probably its greatest advantage and has widespread use across the scanning market.

Large diameter (>100 mm across flats) and also thick-faceted polygons are better suited to mounting on the front of the spindle. These tend to operate at lower speeds (<30,000 rpm), but require a substantial bearing system to support the overhung mass. There is a point where even a large, aerodynamic spindle is not really suitable due to high starting torque and insufficient load capacity and at this point an aerostatic, or hybrid, bearing system will have to be selected (as shown in Figure 6.14).

Whichever system is selected, due design consideration must be given to the polygon mount design. The method of polygon attachment, be it using screws or bonding, must not affect the optical properties of the facets, either statically or dynamically. The choice of material for both the polygon and its mount must take into account the total additional mass being supported by the bearing system, the rotational stresses induced, and thermal growth characteristics between the hub and mounts. Most scanner polygons and mounts are manufactured from high-grade aluminum.

Finally, trapezoidal polygons, that is, polygons with facets not parallel to the axis of rotation, in general will be mounted on the front of the spindle to allow access to the incoming axial laser beam.

6.4.7.1.2 Monogons

Monogons, or single-faceted reflective optics, tend to be used when the output beam from the optic is to strike a circular, rather than a flat surface, such as in a cylindrical drum image setter, although if the output beam is passed through an *F*-theta lens it can then be used on flat surfaces too. The optics used can be of a simple open reflective surface form or a more complex glass form such as a prism. The input laser beam is usually coincident and parallel to the shaft rotational axis and hence all monogons need to be mounted on the front of the spindle.

Particularly with reference to glass optics, but even with aluminum and beryllium, great care is required in the design of the optics holder. Unlike polygons, which normally have

some form of bore for location, most glass optics are designed for their optical quality, not their ease of mounting. Figure 6.27a and b show two forms of housing into which glass prisms have been bonded. Figure 6.27a is a cube prism manufactured from BK7 optical quality glass with the two non-active sides bonded to the extended cheeks of the housing. As can be seen, there are several square edges to the housing, which will cause noise and windage.

An improved version is shown in Figure 6.27b, where the optic bonded in the housing has had all the non-active exposed corners ground to a spherical shape. This is called a ball prism. The housing is also more aerodynamically shaped for reduced turbulence. In both cases, the optics holder is manufactured in high-strength stainless steel. Although this adds extra weight to the shaft assembly over using aluminum, only steel can survive the very high forces exerted by the cheeks at the speeds used; typically 30,000–60,000 rpm.

Figure 6.27c shows a simple version of the open aluminum mirror. Although the mounting technique is easier, with a thread machined onto the back of the mirror stub, there is now the problem of imbalance to correct for, due to the asymmetric form of the mirror. An aluminum ring has to be added to the assembly as shown with heavy metal pins added in the appropriate position. In addition, the open mirror acts as a form of pump, which not only increases the turbulence around the mirror, but it actively attracts dust particles to the mirror surface.

For higher speed application, an improved version of the angled mirror design can be used as shown in Figure 6.27d. The entire mirror is now enclosed in a sphere-shaped housing with input and output windows to stop the pumping action. The stress analysis of this housing, especially around the output window, is of paramount importance due to the large centrifugal force exerted onto the edges of the housing by the output window. Again, balancing the asymmetric form of the combined optic and housing is critical with heavy metal pins added to the end face of the assembly as necessary.
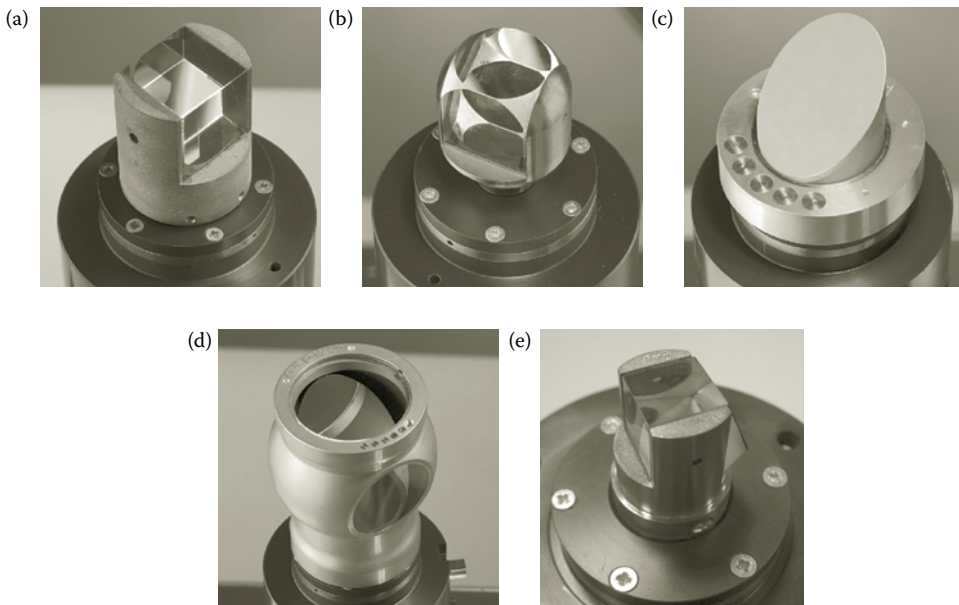
FIGURE 6.27
(a) Cube prism; (b) ball prism; (c) open face mirror; (d) spherical housing containing 45° mirror and (e) pentaprism.

With large clear apertures, up to 30 mm in diameter, optical distortion due to rotational centrifugal forces across the face of the mirror becomes a further challenge to contend with. The use of beryllium, rather than aluminum, reduces the deflection considerably due to its greater modulus of elasticity-to-density ratio, but there is a large cost penalty and processing can be a problem due to the health issues associated with machining beryllium.

Another solution that can be used if the application is designed to operate at only one fixed speed is to bias the surface of the optic. This involves machining a concave surface across the optic face, which becomes flat at the required operating speed. Although the extra process will add additional costs, this might allow the use of an aluminum substrate rather than beryllium, much reducing the overall cost of the optic assembly.

Obviously, the additional mass created by the spherical housing will reduce the top speed of the shaft assembly and Figure 6.28 shows the typical maximum speed for each optic type and size.

One optic not already mentioned in this section is the pentaprism (Figure 6.27e), a special type of prism with two internal reflectance faces. It has the unique ability to correct for a shaft error termed "wobble." This is a random conic motion of the shaft about its longitudinal axis, which occurs much more in ball bearing scanners, but can occur at a low level even in an air bearing. Owing to its relatively large mass, expensive manufacturing processes, and difficult mounting shape, it is rarely used in modern scanner designs.

6.4.7.1.3 Monogon optic bonding

With the recent demands for higher speeds with larger optics, the process of bonding the glass monogon into a metallic holder has become a major issue. Not only for optimum optical performance, due consideration must be given to the safety aspects of the stress in the bond line too. Enormous centrifugal forces are generated within the open ended holders at high speed, particularly in the design in Figure 6.27b, as the cantilever sides try to deflect radially away from the rotational centerline. This stress is imparted into the optic surface via the adhesive film, which must accommodate the resultant deformation without peeling away from either surface. An example of the stress that can be induced in the optic at high speed can be seen in the FEA analysis shown in Figure 6.29. The highest stress can be seen at the root, or the bottom, of the optic, with the lowest around the edges.
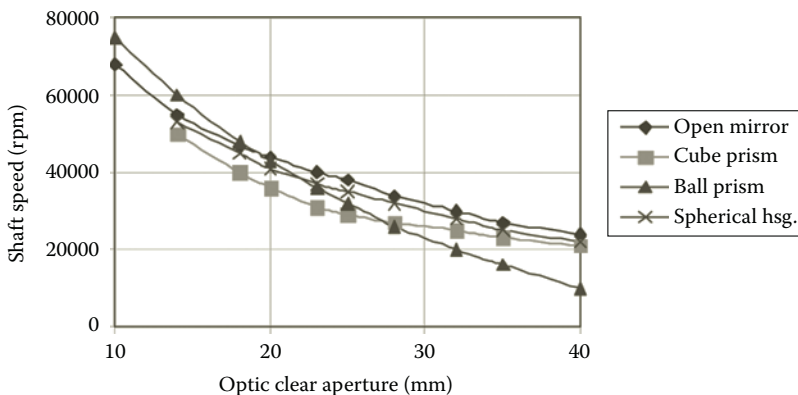
FIGURE 6.28
Speed versus optic size for various monogon types.

Factors to be considered in the bonding process include:

1. Correct adhesive selection for both material surfaces, including reliable material property data from the manufacturer
2. Maximum operating temperature of the assembly during use
3. Surface preparation and cleaning
4. Consistent mixing and spreading of the adhesive
5. Curing time/temperature versus residual stresses in the adhesive
6. Minimizing air entrapment in the adhesive during the process

Visual inspection of the finished optic assembly can often reveal possible process issues. From Figure 6.30, which shows the bonded area between one side of the optic and the cantilever side of the holder, it can be seen that most of the surface has bonded well with only
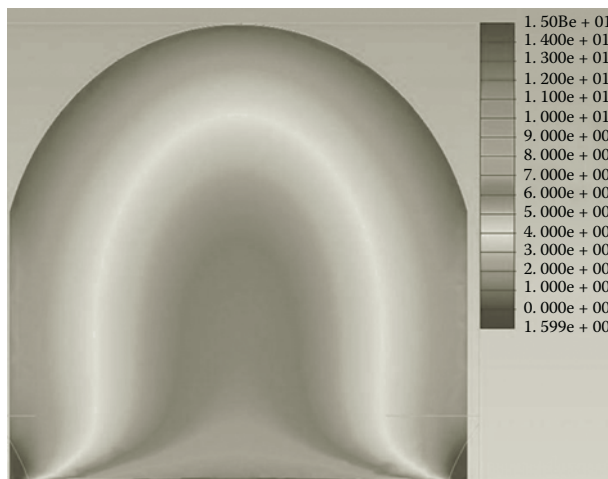
FIGURE 6.29
Finite element analysis of prism.
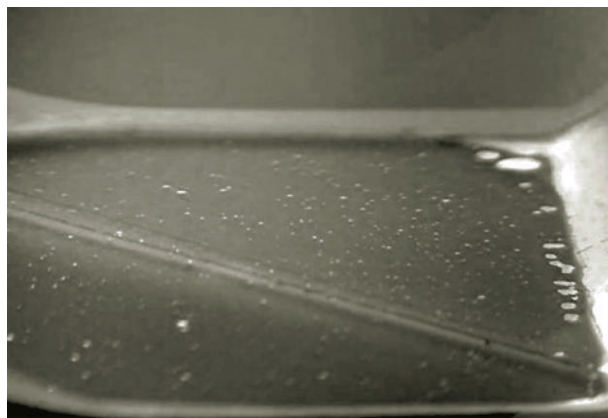
FIGURE 6.30
Adhesive bond area between prism and holder.

very small air bubbles. However, a few large bubbles can be seen at one edge which could lead to premature failure of the bond during rotation.

6.4.7.1.3 Holographic Disc Optics

Certain scanning applications require the use of a holographic disc to be rotated at a relatively low speed. With a large outside diameter usually, and a small bore, this design is again best suited to mounting on the front of the spindle. As the disc is usually made of a glass sandwich, the mass is quite high, so either an aerodynamic or an aerostatic scanner may be used, depending on the minimum and maximum speeds at which the disc must run.

Owing to the principle of the holographic disc, this scanning process is not as optically demanding as rotating monogons and polygons, so many disc scanners in this market are still run on ball bearings.

6.4.7.2 Motors

For most scanning applications, some form of synchronous motor is required to ensure predictable speed control. The two usual designs employed are the hysteresis motor and the brushless DC motor. The lack of brushes and a commutation ring are a definite advantage in gas bearing systems due to there being no wearing parts. In both cases the rotor is simple in mechanical construction, and ideal for high-speed rotation due to its composite nature.

The brushless DC motor consists of a wound laminated stator with integral "Hall effect" devices for commutation wound in. The rotor, in its simplest form, is a cylindrical hollow tube manufactured from sintered samarium–cobalt material. This is then magnetized in a powerful magnetic fixture to the number of poles required. To withstand the high centrifugal forces, this material is usually contained inside another thin steel tube. Some typical brushless DC motors are shown in Figure 6.31.

The rotor assembly can then be bonded directly onto the scanner shaft. In larger motor designs, discrete magnet pairs are used rather than the sintered material, but a steel or carbon fiber containment ring is still required to prevent individual magnets from separating from the shaft at high speed.

The brushless DC motor has become more widely used in recent years and is ideal for aerodynamic scanners, which require a very fast acceleration to reach floating speed

FIGURE 6.31
Brushless DC motor parts.

before surface damage occurs within the bearings. This typically requires a high starting torque, which this motor type can deliver. Also, owing to the use of rare earth magnetic materials, the power density is very high, minimizing the amount of magnetic material and therefore keeping the weight of the rotor down. This is particularly important in overhung rotor designs, which are more difficult to design for high-speed operation than center motorized shafts.

The hysteresis motor has a wound, laminated stator, without "Hall effect" devices, and the rotor consists of a thin cylinder of hardened cobalt steel. This rotor can be shrunk directly onto the scanner shaft. Instead of the torque being created by the permanent magnets in the brushless DC motor, the magnetizing force from the stator induces magnetic fields in the cobalt steel due to hysteresis effect. This type of rotor produces very little heat due to the near absence of rotor eddy currents, and is therefore particularly suited to center motorized polygon spindles where low heat output from the rotor is critical.

6.4.7.3 Encoders

Many high-quality scanning systems require very accurate speed control, typically less than 10 ppm. A typical brushless DC motor can be speed controlled to about 2% using an open loop control system, but if an incremental encoder is fitted, the position of the rotor and shaft can be accurately measured during each revolution and controlled by a phase lock loop controller, yielding results better than 5 ppm with careful optimization.

Some typical encoder disc/hub assemblies and head assemblies are shown in Figure 6.32. The head contains a photo diode emitter and receiver assembly, which is focused onto a fine grating engraved into a precision glass disc. Typical gratings vary from 200 to 1400 lines per revolution. In addition, there is a second track that provides a once per revolution index pulse, often used as a trigger for the start-of-scan process.

The glass disc is mounted on an aluminum hub very accurately, and the disc/hub assembly can then be fitted to the shaft end, usually after the rotor and stator have been fitted. This allows access for final adjustment of the head assembly but from the shaft dynamics
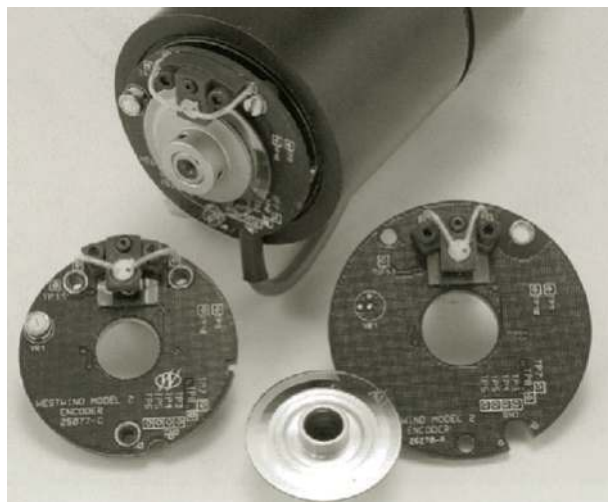
FIGURE 6.32
Optical encoder assembly parts.

viewpoint, this is yet another additional mass that must be allowed for in the design calculations, albeit a relatively light structure.

The strength of the glass disc can also be a limiting factor to the maximum speed of the scanner and for special very high-speed applications, a metal grating disc must be used. However, due to constructional reasons, high line counts with a metal disc are not possible.

For successful operation in a high-accuracy machine, the disc must be centered to a run-out value of <5 microns and would typically exhibit an electronic jitter level of <2 ns.

6.5 BALL BEARINGS

Over the last 50 years or so, the quality of ball bearings has improved immensely and the designs have become well refined. The major manufacturers supply detailed applications design rules together with comprehensive ball bearing characteristic data. Hence it is not intended to revisit this information in detail within this chapter. However, a brief review of their application to rotary scanners is useful to understand the advantages and disadvantages of this bearing technology.

6.5.1 Bearing Design

For precision, high-speed ball bearing scanners, angular contact bearings are normally selected as shown in Figure 6.33. Typically, the contact angle used will be in the range of 12–25°, with the greater the angle, the larger the thrust capacity available.

To ensure an accurate rotational axis, the bearings need to be used in pairs, with an axial preload, which will remove all the play in the shaft/bearing system. The preload can be ground into the bearing assembly by using different length spacers between the outer
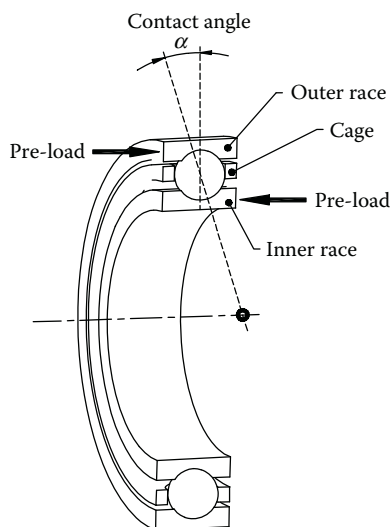
FIGURE 6.33
Angular contact ball bearing.

and inner races, or created by fitting disc springs between the outer face and the bearing housing.

ABEC 5 or ABEC 7 quality bearings are usually specified to ensure that the critical mechanical tolerances are controlled to known standards, particularly parameters such as the radial runout, which will have a big effect on the wobble characteristic of the shaft. Lubrication normally uses some form of light grease, with seals or shields on the bearings to help prevent any leakage. However, evaporation of the lubricant can be a major problem and will affect the life of the bearings.

In recent years one of the main improvements in ball bearing technology has been the introduction of new materials to improve overall performance. Ceramic balls, usually made from silicon nitride, can replace the steel balls in the ring and this design is known as the hybrid bearing. It is widely available from most manufacturers and gives the following advantages:

1. Significantly extended bearing life due to the improved running behavior between the ceramic and steel materials
2. Higher speed rating due to the reduced density of the ceramic balls and hence lower centrifugal forces
3. Lower thermal expansion coefficient of the balls reducing the variation in bearing preload
4. Higher bearing rigidity caused by the higher modulus of elasticity of the ceramics

The improvement in the maximum speed capability of a hybrid bearing, over a standard precision steel ball bearing, for several common inner bore diameters, can be seen in Figure 6.34. These ratings are based on grease lubrication, ignoring preloads and other constructional constraints, and are extracted from typical manufacturers' available data.

A further improvement can be introduced by changing the material of the races themselves, from 440c typically to a finer structure steel, such as a high nitrogen content stainless steel. This allows for cooler operation and higher allowable contact pressure, again increasing the maximum speed rating still further, as shown in Figure 6.34, depicted by the term "hybrid-ultra."
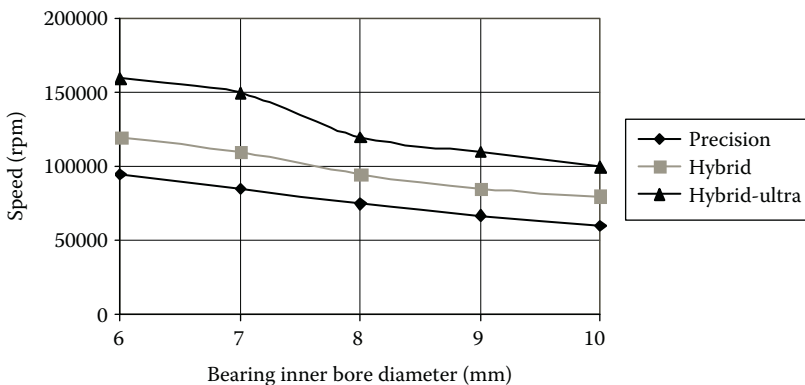
FIGURE 6.34
Speed versus bore diameter for different ball bearing types (grease lubrication).

The improved running life between ceramics and steels is indicated in Figure 6.35, which shows the expected service life for high-speed grease for the two different materials at various DN numbers (mean bearing diameter × speed). This graph can only be taken as an indication of the improvement in grease life, as the individual application, environment, and duty will have an effect on the numbers. The high DN values shown also indicate the recent advancements in synthetic grease technology to allow bearings to run above a DN of 1,000,000 at all without having to resort to oil mist lubrication. Information is based on manufacturers' available data.

6.5.2 Scanner Construction

Owing to the compact design of ball bearings, a variety of bearing configurations can be used in scanner designs. Polygons can be center mounted, overhung, or contained, as shown in Figure 6.36 with both the polygon and the motor assembly sandwiched between the bearings. To keep the surface speed of the balls to a minimum for the longest life,
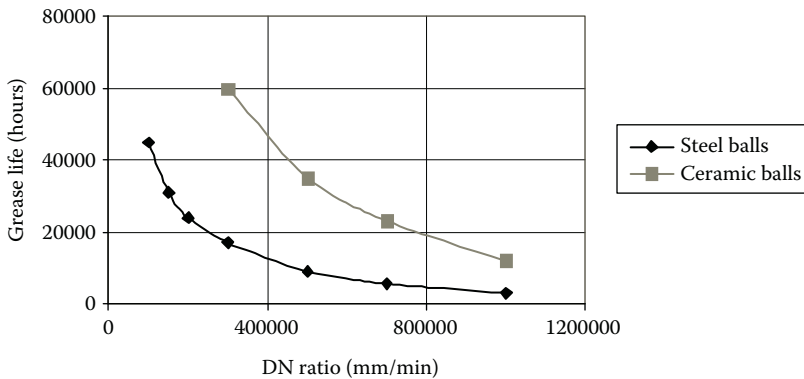
FIGURE 6.35
Grease life versus bearing DN ratio for steel and ceramic balls.
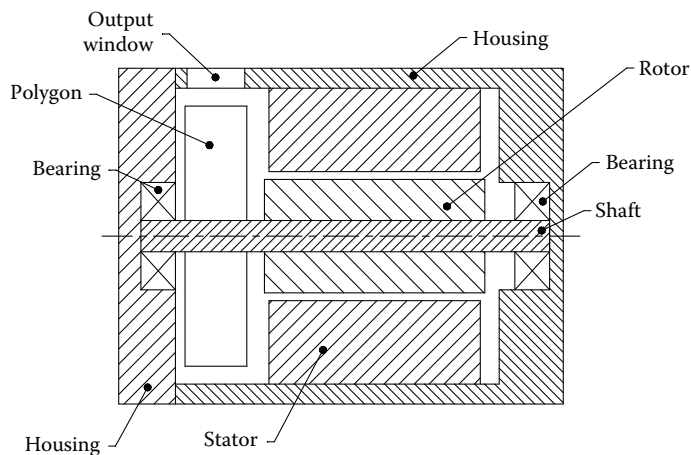
FIGURE 6.36
Scanner with ball bearings.

the shafts for ball bearing designs tend to be small where they fit into the bearing bores, whichever configuration is chosen. This can lead to problems with the shaft critical speeds due to its low stiffness. For very compact designs, for laser printers for example, special motor technology is incorporated such as a "pancake" or radial wound motor.

6.6 MAGNETIC BEARINGS

In recent years, active bearing spindles have been developed commercially for the machine tool industry for high-speed aluminum routing and more recently for turbo-molecular vacuum pumps in the semiconductor processing industry. Prior to this, magnetic bearing technology had been confined to special applications in the aerospace and satellite industry. The success of the commercialization lies mostly in the improvements in the electronic control system for the bearings, requiring powerful, very fast processor chips. With the cost of such chips plummeting in recent years, the largest drawback of the magnetic bearing, that is, the cost of the complete control system, has reduced dramatically.

These systems are now being applied to other lower cost applications within the electronics process industry, and it will not be long before magnetic bearing systems will appear in special applications in the scanning industry, particularly in vacuum conditions.

6.6.1 Bearing Design Principle

The principle of the active magnetic bearing is relatively simple. The journal bearing consists of a laminated core that is fitted onto the shaft, and is surrounded by a wound static stator, which, once energized, holds the shaft in its magnetic center. A displacement transducer close to the bearing constantly monitors the position of the shaft, and if any external force moves the shaft off center, then the control feedback system connected to the transducer adjusts the current in the stator coils to move the shaft back to the magnetic centerline. The feedback system can typically correct the shaft position every 100 µs. Obviously, this system is inherently unstable and in the case of transducer, or even electrical power failure, a set of catch bearings are necessary to prevent the shaft contacting the bearing coils and causing instant damage.

The main technical advantages of this are as follows:

1. Very large clearances between the shaft and the stator coils minimizing heat generation within the bearings and reducing the motor power required to run the spindle
2. Active damping control, which allows the shaft to be driven through shaft criticals that other bearing types could not accommodate
3. Operation in total vacuum conditions without contaminating the environment

6.6.2 Scanner Construction

Figure 6.37 shows a diagrammatic cross section of a magnetic bearing spindle. The design contains two radial and one bidirectional axial magnetic bearings with the high-frequency
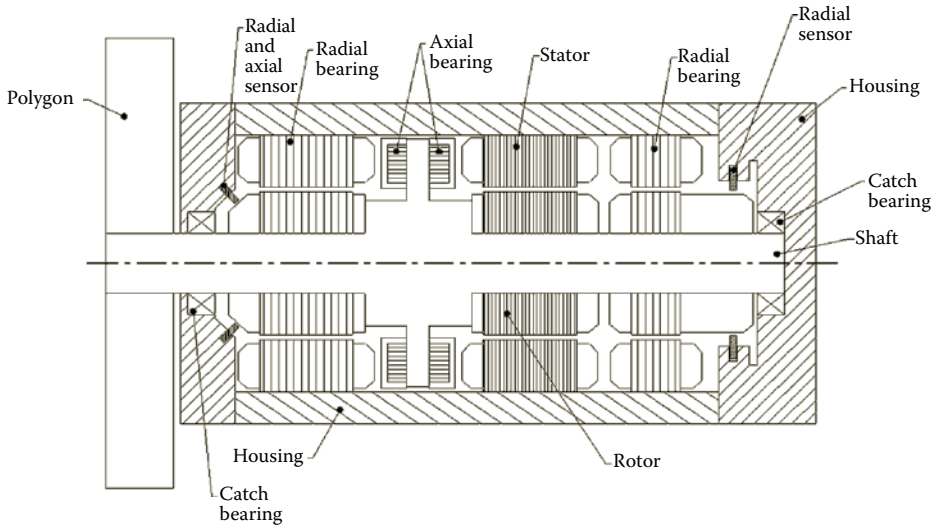
FIGURE 6.37
Scanner with magnetic bearings.

motor between the axial bearing and the rear bearing. In front of the front bearing there are position sensors looking on the shaft at 45° to the axis to provide radial and axial displacement information to the controller, and further radial sensors are placed next to the outer end of the rear bearing.

For overload conditions, or a power failure, two small angular contact "catch" bearings are located one at each end of the shaft, with about 0.1–0.2 mm clearance with the shaft under normal running conditions. Typically, the system controller can maintain the radial runout of the shaft to less than 1 micron.

6.7 OPTICAL SCANNING ERRORS

In any scanning system there will always be errors associated with the rotation of the optic that will lead to the output beam displaying certain distinguishable errors that can be traced back to either bearing-related or optic-related issues.

6.7.1 Bearing-Related Errors

Errors due to the rotation of the shaft can be subdivided into two categories: synchronous (repeatable) and asynchronous (nonrepeatable) motions. With synchronous errors, the shaft is usually performing some kind of regular conic (or wobble) motion around its axis, which could be caused by one of the following issues:

1. Poor shaft assembly balance
2. Running at or close to an axial or radial bearing resonant speed, which will magnify any residual shaft imbalance

3. Manufacturing errors within the bearings or shaft, such as ovality or misalignment

4. Magnetic pulsations from the rotor as it passes the windings

Asynchronous errors are more difficult to locate by their very nature and the shaft may describe quite an irregular motion depending on the cause and the type of bearing.

With gas bearings, the onset of half-speed whirl for whatever reason will produce strange shaft motion before normally leading to bearing failure. This could be due to an excessive bearing clearance caused by thermal effects within the scanner, or an overspeeding of the shaft, or in the case of an aerostatic scanner, a sudden large drop in the supply pressure level. Also, pneumatic hammer occurring in the aerostatic scanner will create asynchronous errors in the shaft.

With ball bearing scanners, asynchronous errors can be caused by manufacturing errors within the ball set beating with the synchronous errors in the track races.

In either bearing system, the ingress of dirt into the bearing will cause intermittent motion errors, which may eventually result in premature failure.

Finally, the motor can cause its own form of asynchronous error, which is commonly called "jitter." It is the result of the feedback system controlling the motion trying to correct for speed variations, caused mainly by optic windage, which will always tend to over- or undershoot the actual target speed. This can be minimized by using a closed-loop system with encoder feedback and reducing the effects of turbulence around the optic.

6.7.2 Optic-Related Errors

Although both polygons and monogons suffer from many of the same errors, their effects on the scanning system need to be examined separately.

6.7.2.1 Polygons

1. *Mounting.* Misalignment of the polygon axis to the spin axis is likely to be the largest part of the total tracking error. Very accurate machining of the polygon bore and the hub on the shaft is required to minimize this tilting effect. *Effect:* Repeating weave pattern.

2. *Manufacture.* Errors include pyramidal (facet-to-datum), facet-to-facet, dividing angle and facet flatness. *Effect:* Repeatable positional errors.

3. *Dynamic distortion.* Loss of geometry due to thermal growth or mechanical stresses. *Effect:* Positional tracking errors varying with speed and time.

6.7.2.2 Monogons

1. *Mounting.* Will only cause a slight permanent change of facet angle, which will occur on every revolution and will not usually affect the scan process. *Effect:* Small, permanent positional change of beam.

2. *Manufacture.* Errors include facet flatness, deflection angle, wavefront distortion, and astigmatism. *Effect:* Spot quality and focus issues with speed and time.

3. *Dynamic distortion.* Change in flatness and astigmatism due to thermal growth or mechanical stresses. *Effect:* Change in spot quality and focus with speed and time.

6.7.3 Error Correction

6.7.3.1 Polygons

Mounting errors can be minimized by machining the polygon hub in situ on the shaft running in its own bearings. This helps to correct for many of the synchronous bearing-related errors as well as the mechanical errors. Alternatively, an adjustable mount can be used to fine tune the tilt of the polygon to bring it on spin axis. The mount can be manufactured from a thermally insulating material to stop thermal effects reaching the polygon.

With synchronous errors, an active correction system can be employed to slightly modify the beam path prior to striking the polygon facet to compensate for the error about to be put into the beam. This can be permanently preprogrammed in, or in more complicated systems, a facet error detector must be incorporated to constantly update the error compensation system.

6.7.3.2 Monogons

Error correction is more limited in monogon optic systems. To correct for dynamic mechanical optic distortion, biased optics can be used that will deform to the correct shape over a small specific running speed range, but this only usually refers to open facet mirrors, not prisms. However, many of the synchronous errors are not so noticeable as they occur on every scan line and in general will not cause banding.

As mentioned earlier in this chapter, the use of a pentaprism can dramatically reduce wobble errors generated in the bearings and is ideally suited for ball bearing scanners where wobble is a major problem in higher accuracy designs.

6.8 SUMMARY

Throughout this chapter, it has been the intention to provide enough theoretical and practical information for the designer to be able to understand, and therefore correctly specify, the bearing system most suitable for the scanning device under consideration. If the conclusion is that a gas bearing is required, then it is anticipated that in most cases, the reader will be intending to buy rather than make the scanning unit due to the complexity of design and manufacture. The graphs contained within this chapter, together with the theory, should provide valuable data regarding the critical parameters such as loads, stiffness, and heat generation. These parameters will have effects on the surrounding parts of the machine and must be taken into account during the machine design process.

Should the designer opt for ball bearing technology, then the option to make rather than buy is more realistic, as both design data and components are readily available. However, the design of the optic and the mount can be the most challenging part of the whole scanner and should not be treated lightly.

ACKNOWLEDGMENTS

REFERENCES

1. Shepherd, J. Bearings for rotary scanners. In Marshall, G. *Optical Scanning;* Marcel Dekker, Inc.: New York, 1991; Chap. 9.
2. Kingsbury, A. Experiments with an air lubricated journal. *J. Am. Soc. Nav. Eng.* 1897, *9,* 267–292.
3. Harrison, W.J. The hydrodynamic theory of lubrication with special reference to air as a lubricant. *Trans. Camb. Phil. Soc.* 1913, *22,* 39.
4. Raimondi, A.A. A numerical solution for the gas lubricated full journal bearing of finite length. *Trans. A. S. L. E.* 1961, *4*(1).
5. Powell, J.W. *The Design of Aerostatic Bearings;* The Machinery Publishing Co.: Brighton, UK, 1970.
6. Grassam, N.S.; Powell, J.W. Eds. *Gas Lubricated Bearings;* Butterworth: London, 1964.
7. Hamrock, B.J. *Fundamentals of Fluid Film Lubrication;* McGraw-Hill, 1994.
8. Whitley, S.; Williams, L.G. The gas lubricated spiral-groove thrust bearing. U. K. A. E. A. I. G. Rep. 28 RD/CA, 1959.
9. Westwind Air Bearings Ltd. An improved bearing. (European EP) Patent No. 0705393, May 1994. Corresponding U.S. Patent No. 5593230.

7

Pre-Objective Polygonal Scanning

Gerald F. Marshall

Consultant in Optics
Niles, Michigan, USA

CONTENTS

7.1 INTRODUCTION

Design equations for regular prismatic polygonal scanning systems have been analyzed and described by Kessler[1] and Beiser.[2,3] Beiser's analytical treatment is comprehensive in that the performance in terms of resolution is the key criterion used for the system designs and analyses. Henceforth, throughout this chapter, the term polygonal scanner shall infer a regular prismatic polygonal scanner.

The prime objective of this chapter is to provide a comprehensive visual understanding of the effects of changing the incident beam width (diameter) *D*, the incident beam offset angle 2*β*, which is the angle the incident beam is offset from the *x*-axis, and the number *N* of mirror facets on a polygonal scanner without regard to performance in terms of resolution. Diagrams, equations, and coordinates bring to light these insights.

Cartesian rectilinear coordinate axes O*x* and O*y* are chosen for the equations of lines, loci, and the coordinates of significant points. The origin coincides with the axis of rotation O of the regular prismatic polygonal scanner. The *x*-axis (O*x*) is parallel to the optical axis of the objective lens.

There are three distinct topics associated with pre-objective polygonal scanning systems that are covered in this chapter by three separate sections, 7.2, 7.3, and 7.4. To assist a reader interested in only one of the topics certain definitions are repeated for continuity of a topic in a section so that the reader does not have to cross-refer back and forth to different sections.

The topics are: equations and coordinates of a polygonal scanning system; instantaneous center-of-scan (ICS); and stationary ghost images outside the image format.

7.1.1 Equations and Coordinates of a Polygonal Scanning System

The midposition orientation of the polygonal scanner facets is such that the reflected collimated incident beam is parallel to the *x*-axis and defines the *scan-axis*, both of which are chosen to be parallel to the optical axis of the objective lens (Figure 7.1).

For a given incident beam offset angle 2*β* of the collimated incident beam, the equations for the scan-axis, the incident beam, the mirror facet plane, and the objective lens optical axis scanner are expressed with respect to the rotation axis O of the polygonal scanner; likewise are the precise coordinates of significant points.[4]

7.1.2 Instantaneous Center-of-Scan (ICS)

Presented is the derivation of the parametric equations for the loci of the ICS for six- and twelve-facet prismatic polygonal scanners.[5] Depicted are figures that show the changes in the loci characteristics for different incident beam offset angles 2*β*.

7.1.3 Stationary Ghost Images Outside the Image Format

Presented is a pictorial display of diagrams illustrating the permissible angular ranges of the incident beam offset angle 2*β* to ensure that the ghost images lie outside the image format of the scanned field image format.[6]

7.2 EQUATIONS AND COORDINATES OF A POLYGONAL SCANNING SYSTEM

The origin of the Cartesian rectilinear coordinate axes can be chosen to be either at the rotation axis O of the polygonal scanner, or at the point of incidence P on a mirror facet;
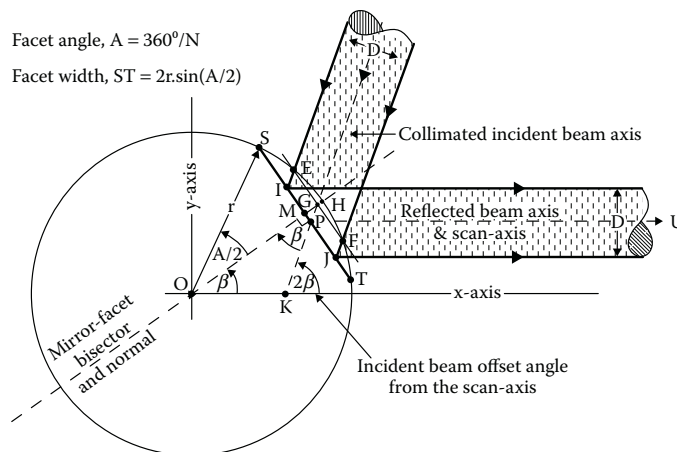
FIGURE 7.1
A single facet ST of a polygonal scanner oriented in the midposition. See Figure 7.2 for greater clarity around the point of incidence P.

each approach has an advantage for giving insights. In this section the rotation axis O has been chosen to be the origin.[4]

Consider Figure 7.1. The diagram depicts a single facet ST of a regular prismatic polygonal scanner with N facets and its circumscribed circle of radius r. The facet ST is oriented so that the collimated incident beam, which is offset at an angle 2β from the x-axis, is reflected parallel to the x-axis. The beam has a finite width D (see Section 7.2.5).

7.2.1 Objective

The goal is to present the precise coordinates of significant points, the distances between these points, and the equations of three axes (incident beam, scan, and objective lens) with respect to the rotation axis O of the polygonal scanner, thereby eliminating manual or computer-aided iterative techniques. Furthermore, it is to provide unexpected interesting insights into the limitations of the optomechanical design layouts of polygonal scanning systems.

7.2.2 Midposition and Scan-Axis

Shown in Figure 7.1 is a single facet ST of a polygonal scanner oriented in a midposition such that a collimated incident beam is reflected parallel to the x-axis. The reflected beam axis PU in this midposition defines the *scan-axis*.

7.2.3 Mirror Facet Angle *A*

From this midposition the reflected beam angularly scans symmetrically about the scan-axis through an angle of $\pm A$. Character symbol A is the facet angle, which is the angle that the facet ST subtends at the rotation axis O of the polygonal scanner.

$$A = \frac{360°}{N} \tag{7.1}$$

7.2.4 Mirror Facet Width

The tangential width of the facet ST of a regular prismatic polygonal scanner is:

$$ST = 2r \sin\left(\frac{A}{2}\right) = 2r \sin\left(\frac{180}{N}\right) \tag{7.2}$$

7.2.5 Beam Width (Diameter) *D*

The Gaussian laser beam width represented by D is the standard "$1/e^2$ beam width" plus a margin of safety chosen by the system designer to minimize the imaged spot defects when a Gaussian beam is one-sidedly truncated by a facet edge as the polygonal scanner rotates. The margin of safety is inextricably linked to the desired *scan duty cycle* (scan efficiency) η of the scanning system for a given 2β, N, and r. For a one-side truncation of the beam by the facet edges as the polygon rotates, optimally D is 40% greater than the $1/e^2$ beam width.[2,3]

The $1/e^2$ beam width that is symbolized by D_{1/e^2} is the beam width (diameter) beyond which the residual laser beam power is $1/e^2$ of the total power of a laser beam that has a

Gaussian distribution. For a Gaussian distribution, it uniquely and directly also corresponds to the laser beam width (diameter) at which the beam irradiance has dropped to $1/e^2$ of the axial peak laser beam irradiance (see Chapter 1).

In Figure 7.1, in which the polygonal scanner is in the midposition:

1. The boundaries of the incident beam width D are designed to cut through the circumscribed circle at E and F such that the arcs SE and FT are equal. This ensures that the useful angular scan of the reflected beam is symmetrical about the scan-axis.

2. In Figures 7.1 and 7.2 the incident beam axis that has a finite width D intersects the mirror-facet bisector OMH at the point G, as it proceeds to impinge on the mirror-facet as the point P. G is aslo the midpoint of the chord EF. For an infinitesimal beam width G coincides with the point H. As the beam increases G approaches M as does P.

7.2.6 Scan Duty Cycle (Scan Efficiency)

The *scan duty cycle η* of a polygonal scanner is the ratio of the useful scan angle during which the beam width D is unvignetted by the edges of the facets, to the full scan angle $±A$ of a beam with an infinitesimal beam width. One assumes that the tangential width of the footprint of the incident beam is less than the tangential width of the facet, when the polygonal scanner is in its midposition (see Chapter 2 of this volume). Henceforth, widths refer to tangential widths.

It can be shown[4] that

$$h = 1 - \frac{\arcsin[D/(2r \cos b)]}{180/N} \tag{7.3}$$

Knowing or selecting suitable values for r, N, $β$, and D will determine $η$. Alternatively, choosing suitable values for r, N, $β$, and $η$ will determine the required incident beam width

FIGURE 7.2
A geometrical diagram for determining the coordinates of G and P, namely, X_G, Y_G and X_P, Y_P.

D as follows. Transposing Equation 7.3 gives

$$D = (2r \cos b)\sin[(180/N)(1-h)] \qquad (7.4)$$

More simply, if *W* represents the tangential facet width, this expression approximates to[2,3]

$$D_{approx} = W(\cos b)(1-h) \qquad (7.5)$$

Dividing Equation 7.5 by Equation 7.4 gives

$$D_{approx}/D = \frac{[\sin(180/N)](1-h)}{\sin[(180/N)(1-h)]} \qquad (7.6)$$

The closeness of D_{approx} to *D* is illustrated in Table 7.1.
  Let the polygonal scanner be in its midposition, then:

1. If the incident beam has an infinitesimal width (*D* = 0), the circumscribed circle of the polygonal scanner intersects the axis of the beam at the point H, which coincides with the top of the facet sag MH. The scan duty cycle is 100% (η = 1), ignoring the inevitable facet edge manufacturing roll-off (Figure 7.2).

2. If the incident beam has a finite width *D* with a footprint that just covers the facet's tangential width (*D* = 2*r* sin (*A*/2)), the beam axis is directed at M, at the base of the sag MH. The scan duty cycle is 0% (η = 0). Simultaneously, point P coincides with M, which is the midpoint of the mirror-facet chord ST.

3. For all finite incident beam widths *D* with a footprint width that is within the facet width (*D* < 2*r* sin (*A*/2)), the beam axis passes through the midpoint G of the chord EF, and which lies on the facet sag MH, to impinge on the mirror-facet at the point P. The scan duty cycle is finite (1 > η > 0) (Equation 7.3, Figures 7.1 and 7.2).

7.2.7 Sag Dimensions

It can be shown[4] and with reference to the geometry in Figure 7.2 that when sag MH = *m*

$$m = r\left[1 - \cos\left(\frac{A}{2}\right)\right] = r\left[1 - \cos\left(\frac{180}{N}\right)\right] \qquad (7.7)$$

TABLE 7.1

Ratio [D_{approx}/D] for Scan Duty Cycle η versus Number of Facets *N*

	η				
	0.00	0.25	0.50	0.75	1.00
N = 3	1.00	0.92	0.87	0.84	0.83
N = 6	1.00	0.98	0.97	0.96	0.95
N = 12	1.00	0.99	0.99	0.99	0.97
N = 18	1.00	1.00	1.00	1.00	1.00
N = 24	1.00	1.00	1.00	1.00	1.00

If the sag GH = g, then in terms of r, D, and β

$$g = r\left(1 - \sqrt{1 - \left[\frac{D}{(2r \cos b)}\right]^2}\right) \qquad (7.8)$$

Or from Equations 7.7 and 7.8

$$(m - g) = r\left(-\cos\left(\frac{180}{N}\right)\right) + \sqrt{\left[1 - \left(\frac{D}{2r \cos b}\right)^2\right]} \qquad (7.9)$$

$$(r - g) = \sqrt{\left[1 - \left(\frac{D}{2r \cos b}\right)^2\right]} \qquad (7.10)$$

Likewise, in terms of r, N, and η

$$g = r\left(1 - \sqrt{1 - \left(\frac{D}{2r \cos b}\right)^2}\right) \qquad (7.11)$$

$$(m - g) = r\left(-\cos\left(\frac{180}{N}\right)\right) + \sqrt{\left[1 - \left(\frac{D}{2r \cos b}\right)^2\right]} \qquad (7.12)$$

$$(r - g) = \sqrt{\left[1 - \left\{\sin\left[\left(\frac{180}{N}\right)(1 - h)\right]\right\}^2\right]} \qquad (7.13)$$

(see Figures 7.2 and 7.4).

7.2.8 Coordinates of G

From the geometry in Figure 7.2:

$$X_G = (r - g)\cos b \qquad (7.14)$$

and

$$Y_G = (r - g)\sin b \qquad (7.15)$$

Substituting for $(r - g)$ from Equation 7.10, and expressing X_G, Y_G in terms of r, D, and β give

$$X_G = r\left(\sqrt{1 - \left(\frac{D}{2r \cos b}\right)^2}\right)\cos b \qquad (7.16)$$

and

$$Y_G = r\left(\sqrt{1 - \left\{\frac{D}{2r\cos b}\right\}^2}\right)\sin b \qquad (7.17)$$

Likewise, substituting for $(r - g)$ from Equation 7.10, and expressing X_G, Y_G in terms of r, N, and η give

$$X_G = r\left(\sqrt{1 - \left\{\sin\left[\left(\frac{180}{N}\right)(1 - h)\right]\right\}^2}\right)\cos b \qquad (7.18)$$

and

$$Y_G = r\left(\sqrt{\left[-\left\{\sin\left[\left(\frac{180}{N}\right)(1 - h)\right]\right\}^2\right]}\right)\sin b \qquad (7.19)$$

7.2.9 Coordinates of P

Again from the geometry in Figure 7.2:

$$X_P = X_G - (m - g)\left[\frac{\cos 2b}{\cos b}\right] \qquad (7.20)$$

and

$$Y_P = Y_G - 2(m - g)\sin b \qquad (7.21)$$

Substituting for X_G and Y_G from Equations 7.14 and 7.15 gives

$$X_P = (r - g)\cos b - (m - g)\left[\frac{\cos 2b}{\cos b}\right] \qquad (7.22)$$

and

$$Y_P = (r - g)\sin(b - 2)(m - g)\sin b \qquad (7.23)$$

By substituting for $(m - g)$ and $(r - g)$ from Equations 7.9 and 7.10 into Equations 7.22 and 7.23 one obtains the coordinates of X_P, Y_P expressed in terms of r, D, and β.

$$X_P = \left(\frac{r}{\cos b}\right)\left[\cos\left(\frac{180}{N}\right)\cos 2b + \sin 2b\sqrt{\left(1 - \left[\frac{D}{2r\cos b}\right]^2\right)}\right] \qquad (7.24)$$

and

$$Y_P = (r \sin b) \left[2 \cos\left(\frac{180}{N}\right) - \sqrt{\left(1 - \left[\frac{D}{(\cos b)}\right]^2\right)} \right] \tag{7.25}$$

By substituting for $(m - g)$ and $(r - g)$ from Equations 7.12 and 7.13 into Equations 7.22 and 7.23 one obtains the coordinates of X_P, Y_P expressed in terms r, N, and η.

$$X_P = \left(\frac{r}{\cos b}\right) \left[\cos\left(\frac{180}{N}\right) \cos 2b + \sin 2b \sqrt{\left(1 - \left\{\sin\left[\left(\frac{180}{N}\right)(1 - h)\right]\right\}^2\right)} \right] \tag{7.26}$$

$$Y_P = (r \sin b) \left[2 \cos\left(\frac{180}{N}\right) - \sqrt{\left(1 - \left\{\sin\left[\left(\frac{180}{N}\right)(1 - h)\right]\right\}^2\right)} \right] \tag{7.27}$$

7.2.10 Optical Axis of the Objective Lens

The objective lens optical axis, which is parallel to both the *x*-axis and the scan-axis, is directed through the point G to ensure that the scanning beam width *D* scans symmetrically across the aperture of the objective lens (Figures 7.1, 7.2, and 7.3).

The separation between the objective lens optical axis and the scan-axis is given by

$$WG = 2(m - g)\sin b \tag{7.28}$$

FIGURE 7.3
The six facets of the hexagonal scanner head oriented in the midposition at which the reflected incident beam axis lies along the scan-axis. The boundaries of the beam, which has a width *D,* are omitted to avoid overcrowding the diagram.

If the incident beam has an infinitesimal width, G coincides with H ($g = 0$), and the separation WG between the objective lens optical axis and the scan-axis is a maximum

$$WG_{max} = 2m(\sin b) \tag{7.29}$$

Substituting for m from Equation 7.7 leads to

$$WG_{max} = 2r\left[1 - \cos\left(\frac{180}{N}\right)\right]\sin b \tag{7.30}$$

As G and P simultaneously approach M the objective lens optical axis and the scan-axis move toward each other at the same rate in the y-axis direction until they both coincide at M.

When the incident beam has a finite width D and the beam width of the footprint just covers the facet chord ST, G and P coincide with M, $m = g$, $(m - g) = 0$.

Substituting $(m - g) = 0$ into Equation 7.28 gives

$$WG_{min} = 0 \tag{7.31}$$

Thus, the objective lens axis is coincident with the scan-axis.

7.2.11 Equations

Except for the incident beam, the scan-axis and objective lens optical axis are parallel to the x-axis and, therefore, have equations independent of x.

7.2.11.1 Scan-Axis PU

The equation to the scan-axis corresponds to Y_P, given in Equation 7.23, namely

$$Y_P = (r - g)\sin b - 2(m - g)\sin b \tag{7.32}$$

See Equations 7.25 and 7.27.

In a reverse sense Y_P also represents the offset distance of the rotation axis O from the scan-axis PU for a given offset angle β of the incident beam (see Sections 7.2.12 and 7.4.9).

7.2.11.2 Objective Lens Optical Axis

The equation to the objective lens optical axis corresponds to Y_G given in Equation 7.15, namely

$$Y_G = (r - g)\sin b \tag{7.33}$$

From Equation 7.17, expressing Y_G in terms of r, D, and β gives

$$Y_G = r\left(\sqrt{1 - \left[\frac{D}{(2r\cos b)}\right]^2}\right)\sin b \tag{7.34}$$

From Equation 7.19, expressing Y_G in terms of r, N and η gives

$$Y_G = r\left(\sqrt{\left[1 - \left\{\sin\left[\left(\frac{180}{N}\right)(1-h)\right]\right\}^2\right]}\right)\sin b \tag{7.35}$$

7.2.11.3 Incident Beam Axis Through GP

$$y = (\tan 2b)x - (r - g)[(\tan 2b)(\cos b) + \sin b] \tag{7.36}$$

where from Equation 7.10 $(r - g)$ is expressed in terms of r, D, and β, namely

$$(r - g) = \sqrt{\left[1 - \left(\frac{D}{2r \cos b}\right)^2\right]} \tag{7.37}$$

Alternatively, where from Equation 7.13 $(r - g)$ is expressed in terms of r, N, and η, namely

$$(r - g) = \left[\sqrt{1 - \left\{\sin\left[\left(\frac{180}{N}\right)(1-h)\right]\right\}^2}\right] \tag{7.38}$$

7.2.11.4 Mirror Facet Bisector and Normal

The linear equation to the bisector of the mirror facet and normal has a slope of $\tan \beta$ with no intercept and passes through the rotation axis O, which is the origin of the coordinate system. Thus

$$y = x(\tan b) \tag{7.39}$$

7.2.12 Insights from an Alternative Analytical Approach

An alternative perspective for the analysis is to set the Cartesian rectilinear coordinate axes to be Px and Py with the origin at the point of incidence P on the facet for when the polygonal scanner is in a midposition (Figure 7.4). In this approach the scan-axis is collinear with the abscissa, the x-axis (Px), while the ordinate is the y-axis (Py). See the first paragraph at the beginning of Section 7.2.

The immediate advantage is that equations for the scan-axis, the incident beam axis, and the facet plane all pass through the origin P. The goal of this second approach is to determine the coordinates (X_O, Y_O) of the rotation axis O of the polygonal scanner with respect to the point of incidence P and the scan-axis Px.

The approach presents a diagrammatic visualization of the existence of a finite areal zone, with respect to the fixed point P, of a set of loci for the rotation axis O_N of a polygonal scanner that results from changes in the number N of facets ($3 \leq N < \infty$), the laser beam width D, and the scan duty cycle η, ($0 \leq \eta \leq 1$), for a given circumscribed circle of radius r of the polygonal scanner (Figure 7.4).[4]
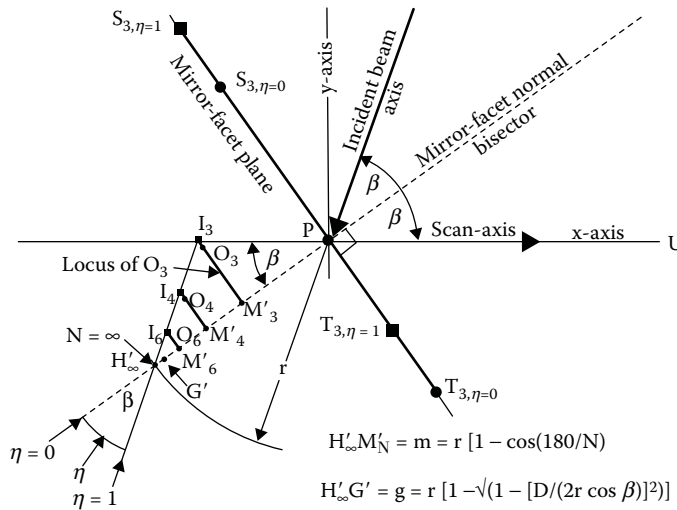
FIGURE 7.4
Displayed are the loci of the rotation axes O_N relative to the point of incidence P on a mirror-facet plane of a polygonal scanner oriented in a midposition. I_3 is the rotation axis O_3 for a three mirror-faceted polygon N = 3 for an infinitesimal beam width, such that $\eta = 1$. As the beam width increases O_3 moves along $I_3 M_3'$ toward M_3' where $\eta = 0$. Likewise for I_4 and I_6; as O_4 and O_6 move along $I_4 M_4'$ and $I_6 M_6'$, respectively.

All these coordinates and equations can be obtained from those already given in Sections 7.2.8 to 7.2.11 by the transformation of the origin at O to an origin at P.

7.2.13 Features of Figure 7.4

In Figure 7.4, the loci for $N = 5$ and for $N > 6$ are omitted to avoid overcrowding the diagram. Certain character symbols are primed because of a direct, but not obvious, relationship to those corresponding unprimed symbols in Figure 7.2.

The set of loci for the position of the rotation axis O_N are confined to the series of parallel base lines $I_N M_N$ of a nest of right triangles $H_\infty' M_N' I_N$ within the triangle $H_\infty' M_3' I_3$. These base lines are parallel to the facet plane ST (Table 7.2).

$$I_N M_N = m \tan b = r \left[1 - \cos\left(\frac{180}{N}\right) \right] \tan b \qquad (7.40)$$

The base lines $I_n M_n$ are spaced at ever diminishing distances toward the apex H_∞' of the triangle as the number N of facets increases. The spacing between every sixth base line is given by

$$\left[m_N - m_{N+6} \right] = r \left[\cos\left(\frac{180}{N+6}\right) - \cos\left(\frac{180}{N}\right) \right] \qquad (7.41)$$

Simultaneously at a lesser rate, the facet widths $S_N T_N$ shorten as the number N of facets increases:

$$S_N T_N = 2r \sin\left(\frac{180}{N}\right) \qquad (7.42)$$

TABLE 7.2

Facet Width W_N and Locus Length $I_N M_N$ versus Number of Facets N for r = 50 mm

	N			
	6	12	18	24
$W_N = S_N T_N$ (mm)	50.00	25.9	17.4	13.1
$I_N M_N$ (mm)	3.87	0.46	0.13	0.06
$[S_N T_N]/[I_N M_N]$	12.9	56.7	130	232
$[m_N \boxtimes m_{N+6}]$ (mm)	5.00	0.94	0.33	0.15

From Equations 7.42 and 7.40 the ratio of the facet width $S_N T_N$ to length of the locus $I_N M_N$ is expressed by

$$\frac{S_N T_N}{I_N M_N} = 2[\cotan(90/N)]^2 \tag{7.43}$$

for which the incident beam offset angle $2\beta = +A$.

The position of the rotation axis on a locus $I_N M_N$ depends on the scan duty cycle $(0 < \eta < 1)$. A fan of straight lines emanating from H'_∞ toward $I_N M_N$ represents a set of values for constant scan duty cycle η. The rotation axis O_n lies at the intersection of one of these fan lines of constant η with a base line $I_N M_N$.

A set of straight lines parallel to $H'_\infty I_3$ represents a set of values for constant beam width D. Similarly, the rotation axis O_N lies at the intersection of one of these parallel lines of constant D with a base line $I_N M_N$. The rotation axes O_N may not lie beyond M_N where the incident beam width footprint matches the facet width.

All facet widths $S_N T_N$ lie between the points $S_{3,h=1}$ and $T_{3,h=0}$ according to the values of N and η. The positional range of facet $S_N T_N$ directly corresponds to the range of the locus O_N, that is, the length of the baseline $I_N M_N$. Uniquely, the rotation axis O lies on the scan-axis only when $N = 3$ and $D = 0$, that is, an incident beam of infinitesimal width.

7.2.14 Conclusion

The visualization of the effects of changing the controlling parameters N, β, D, η, and r of an optical scanning system helps in its design, while, in particular, the explicit coordinates and equations eliminate manual or computer-aided iterative techniques.

7.3 INSTANTANEOUS CENTER-OF-SCAN

Reflective scanning devices, resonant, galvanometric, and polygonal, have plane mirrors that oscillate or rotate about an axis. The rotation of the reflecting mirror deflects an incident light beam. When (1) the axis of rotation O is coincident with the mirror surface, and (2) the incident beam is directed at the axis of rotation, the ICS is a single stationary point on, and at, the axis of rotation O for all angular positions of the mirror. These

two conditions are difficult to achieve and are rarely met; as a result, the ICS moves with respect to the rotation axis O, and, therefore, is a locus (Figure 7.5).[5]

7.3.1 Objective

This section explores and illustrates the characteristic form of the ICS locus for polygonal scanners with respect to the incident beam offset angle 2β; that is, the angle between the incident beam and the scan-axis. The analysis, study, and depiction of the ICS loci for several incident beam offset angles for regular prismatic polygonal scanners of six and 12 facets give a visual appreciation of the asymmetry in the optical path lengths of the deflected beam as it sweeps through the full scan angle $\pm A$. These characteristics provide interesting insights for consideration when undertaking the design of a polygonal scanning system.

7.3.2 Locus of the Instantaneous Center-of-Scan

In Figure 7.5a, the center-of-scan of the reflected beam is a stationary point on the rotation axis O. This is because two conditions are met: (1) the axis of rotation lies in the reflecting surface of a mirror (reflector); and (2) the incident beam is directed at the axis of rotation O of the mirror.

In Figure 7.5b, the center-of-scan is not a stationary point, because, although the axis of rotation lies in the reflecting surface of a mirror, the incident beam is directed to one side of the rotation axis O. However, there is an ICS at point $C \equiv (\alpha, \gamma)$ that has a locus.
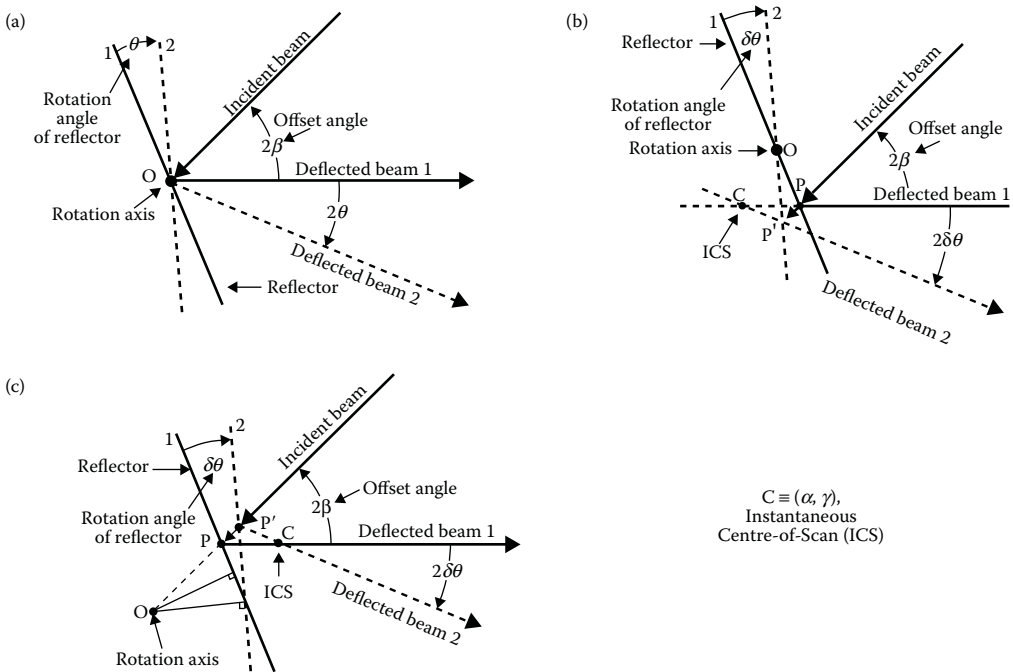
FIGURE 7.5
(a) The axis of rotation lies in the reflecting surface of a mirror (reflector) and the incident beam is directed at the axis of rotation O of the mirror, (b) The axis of rotation lies in the reflecting surface of a mirror but the incident beam is directed to one side of rotation axis O. (c) The incident beam is directed at the axis of rotation but the rotation axis O is displaced from the reflecting surface of the mirror.

Again, in Figure 7.5c, the center-of-scan is not a stationary point, because, although the incident beam is directed at the axis of rotation, the rotation axis O is displaced from the reflecting surface of the mirror. Similarly, there is an ICS at point $C \equiv (\alpha, \gamma)$ that has a locus.

7.3.3 Midposition and Scan-axis

Figure 7.6 depicts a cross section of a hexagonal polygonal scanner set in a midposition with an incident beam offset angle 2β of 70°. The midposition is defined by two requirements: (1) the polygonal scanner is oriented such that the reflected incident beam from one of the facets is parallel to the *x*-axis (this reflected incident beam defines the scan-axis); and (2) the rotation axis O is offset from the scan-axis to a position such that, as the polygonal scanner rotates, the reflected incident beam angularly scans symmetrically $\pm A$ about the scan-axis.

7.3.4 Derivation of the Instantaneous Center-of-Scan Coordinates

Consider a regular prismatic polygonal scanner with N facets and a circumscribed circle of radius r (Figure 7.6). Cartesian rectilinear coordinate axes O*x* and O*y* are chosen for the equations of lines, loci, and the coordinates of significant points. The origin O coincides with the axis of rotation of the polygonal scanner. The *x*-axis (O*x*) is parallel to the optical axis of the objective lens. The facet angle A, that is, the angle that the facet subtends at the axis of rotation O, is given by 360°/N. For simplicity it is assumed that the beam width (diameter) is infinitesimal, such that a single ray represents the incident beam. ICS loci for finite beam widths D are considered in Section 7.3.10.

FIGURE 7.6
A scaled cross section of a six-facet polygonal scanner in the midposition from which the incident beam at an offset angle of 2β is reflected parallel to both the objective lens optical axis and the *x*-axis. This reflected incident beam defines the scan-axis.

Consider now an incident beam directed at the facet of a polygonal scanner in a mid-position at an offset angle 2β (70°) (Figure 7.6). Point H on the incident beam is where the circumscribed circle of the polygonal scanner and a facet bisector OM scanner intersects the incident beam.

Figure 7.7 depicts the position of one facet of the polygon after it has been rotated counterclockwise through an angle θ, and the resultant position and direction of the reflected incident beam that has been deflected through an angle 2θ.

The linear equation of the reflected beam passes through the ICS coordinates (α, γ) and is represented by

$$(y - g) = [\tan (2g)](x - a) \tag{7.44}$$

The incident beam linear equation expressed in the intercept form is

$$\frac{x}{(r/2)/\cos b} + \frac{y}{-(r/2)[\tan(2b)]/\cos b} = 1 \tag{7.45}$$
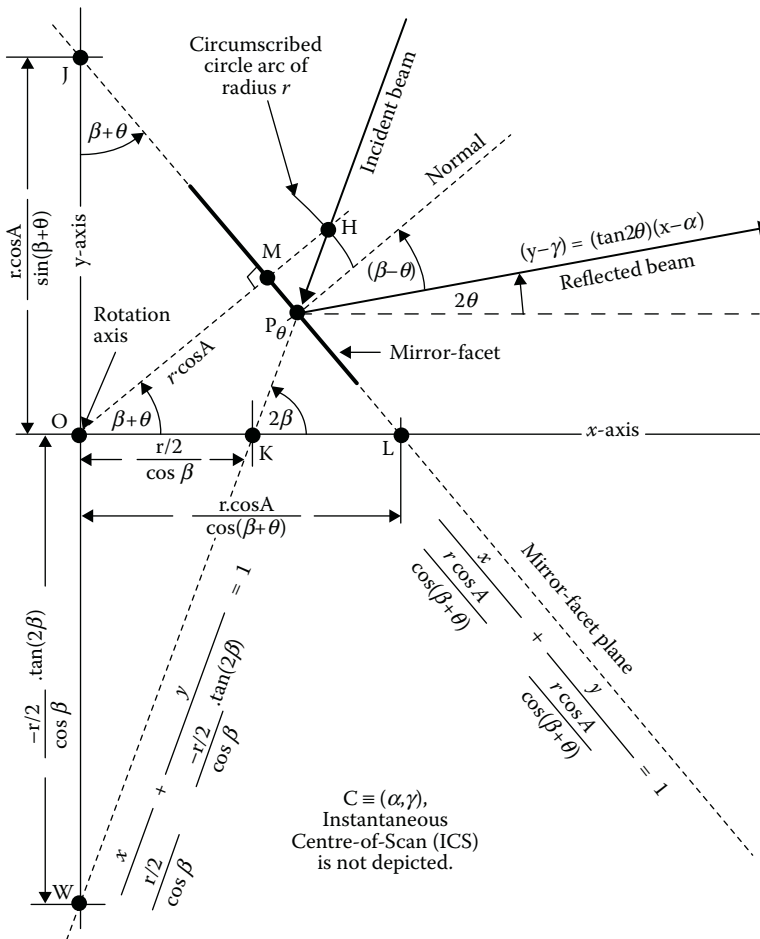
FIGURE 7.7
A trigonometrical diagram depicts the three key analytical equations of one facet of the polygonal scanner illustrated in Figure 7.6 in which the diagram has been rotated counterclockwise through an angle θ.

The linear equation for the line of intersection of the facet plane and the plane of incidence expressed in the intercept form is

$$\frac{x}{r \cos A/\cos(b+q)} + \frac{y}{r \cos A/\sin(b+q)} = 1 \tag{7.46}$$

From the three Equations 7.44, 7.45, and 7.46 the coordinates of α and γ may be determined. The technique is to differentiate Equations 7.44 and 7.45 with respect to θ remembering that α and γ are not variables, but constants at any instant.

Thus the derivative of Equation 7.44 is

$$(x-a) = \frac{(\cos 2q)^2(y' - x' \tan 2q)}{2} \tag{7.47}$$

And the derivative of Equation 7.45 is

$$y' = x' \tan 2b \tag{7.48}$$

7.3.5 Solutions

Solving for $(x - \alpha)$ and $(y - \gamma)$: eliminating y' between Equations 7.47 and 7.48 gives

$$(x-a) = \frac{x'(\cos 2q)^2(\tan 2b - \tan 2q)}{2} \tag{7.49}$$

Substituting for $(x - \alpha)$ from Equation 7.44 into Equation 7.49 leads to

$$(y-g) = \frac{x'(\sin 2q)(\cos 2q)(\tan 2b - \tan 2q)}{2} \tag{7.50}$$

Inspection of Equations 7.49 and 7.50 shows the need to solve and substitute for x and y, and x' with expressions containing only r, A, β, and θ.

Solving for x and y, and x'

Note that simultaneous Equations 7.45 and 7.46 do not contain the ICS coordinates α and γ, thus solving for x and y gives the following parametric equations in terms of r, θ, A, and β:

$$x = \frac{r[\cos A/\tan 2b + \sin(b+q)/2 \cos b]}{[\sin(b+q) + \cos(b+q)/\tan 2b]} \tag{7.51}$$

Likewise

$$y = \frac{r[\cos A - \sin(b+q)/2 \cos b]}{[\sin(b+q) + \cos(b+q)/\tan 2b]} \tag{7.52}$$

Equations 7.51 and 7.52 also represent the locus of the point of incidence $P, (X_{P_q}, Y_{P_q})$, as θ varies and as the reflected incident beam scans. P inherently lies along the segment HP of the incident beam (Figures 7.6 and 7.7).

7.3.6 Spreadsheet Program

The derivative x' is obtained by differentiating Equation 7.51 with respect to θ. An explicit expression is possible but unnecessary when using a computer spreadsheet program. Tabulating data against θ, obtained by using a spreadsheet program, are the values of $(x - \alpha)$ and $(y - \gamma)$ from Equations 7.49 and 7.50; the values of x and y from Equations 7.51 and 7.52, and the values of the derivative x'; thence the coordinates α and γ are deduced and plotted (Figures 7.8 through 7.12).
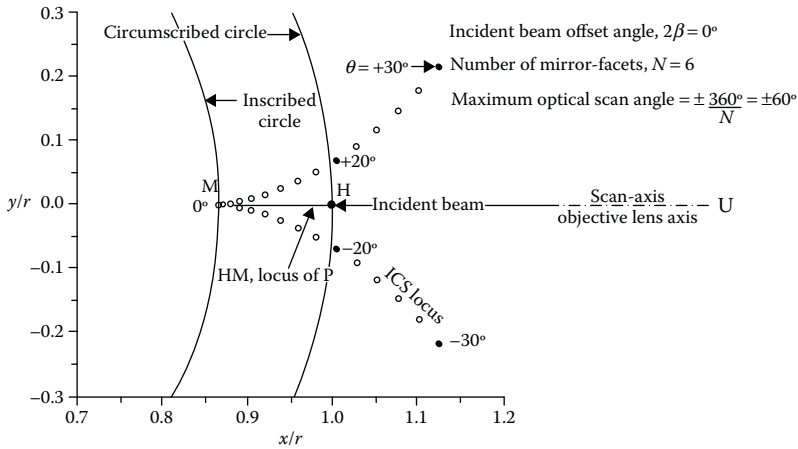
FIGURE 7.8
The ICS locus for an incident beam at an offset angle 2β of $0°$ for a six-facet polygonal scanner.
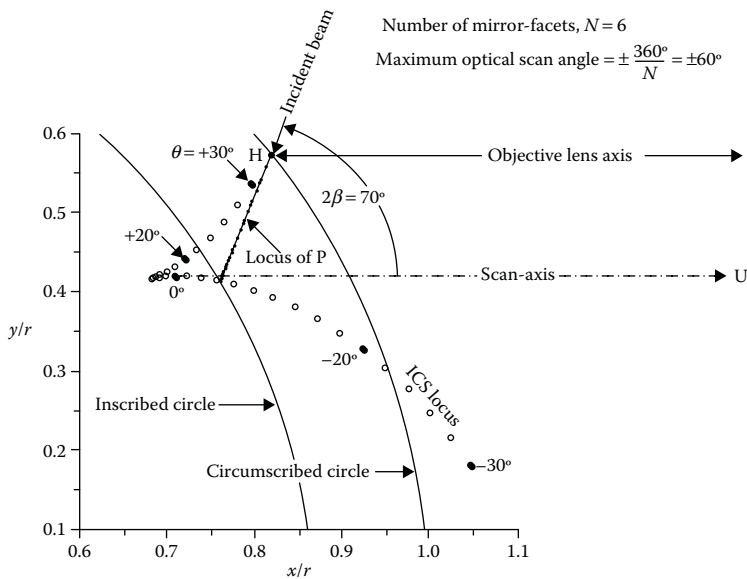
FIGURE 7.9
The ICS locus for an acute incident beam at an offset angle 2β of $70°$ for a six-facet polygonal scanner.
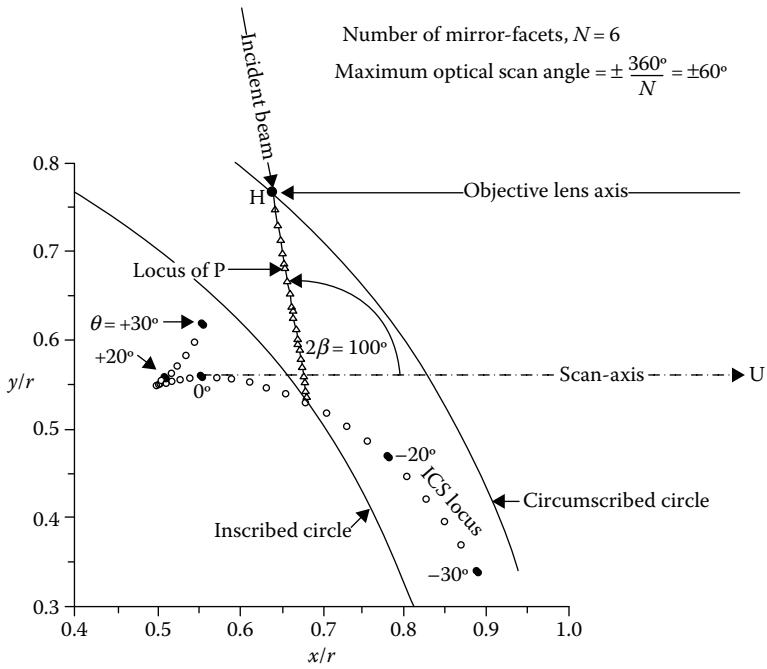
FIGURE 7.10
The ICS locus for an obtuse incident beam at an offset angle 2β of 100° for a six-facet polygonal scanner.
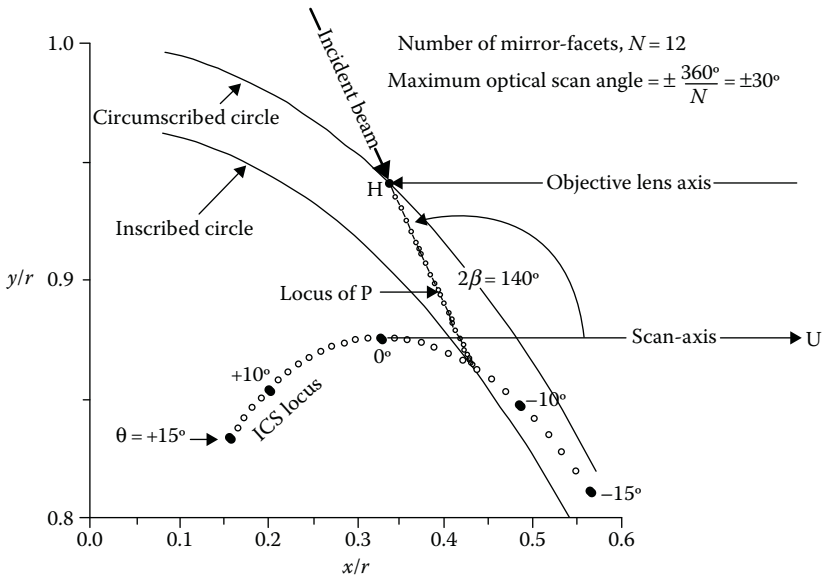
FIGURE 7.11
The ICS locus for an obtuse incident beam at an offset angle 2β of 140° for a 12-facet polygonal scanner. A tangent to the ICS locus represents the position and direction of the reflected incident beam.
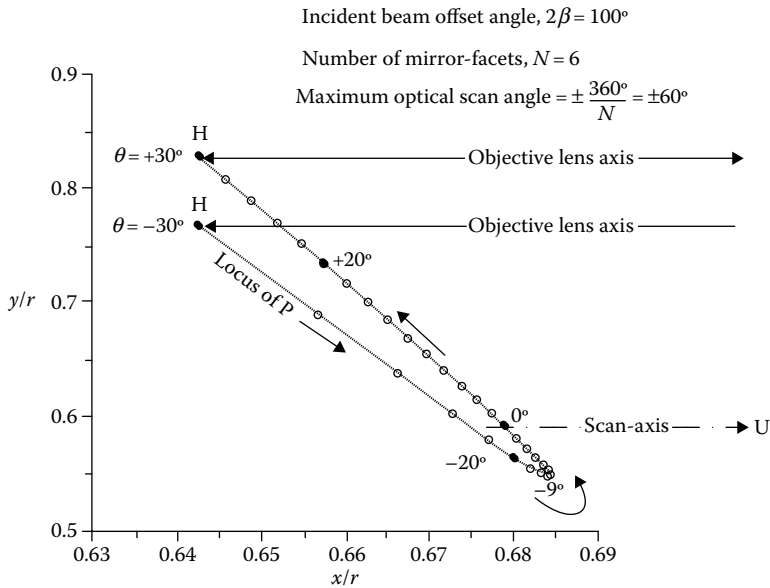
FIGURE 7.12
The locus of P for an incident beam at an offset angle 2β of 100° of Figure 7.10. For visibility and clarity the ordinate data and scale of the abscissa have been adjusted.

7.3.7 Instantaneous Center-of-Scan

Figures 7.8 to 7.11 display the ICS loci for four incident beam β offset angles, namely, 0°, 70°, 100°, and 140°. The data plots on the ICS loci correspond to the mechanical rotation angle θ of the polygonal scanner at two-degree intervals from its midposition.

It should be noted that a tangent at any point on the ICS locus is the position and direction of the reflected incident beam for a rotation angle θ. When the facet edges, S and T, on the circumscribed circle pass through the fixed point H on the incident beam, so also on the ICS locus do the tangents that represent the reflected incident beam at the full optical scan angles $\pm A$ ($\theta = \pm A/2$) (Figures 7.6, 7.7, and 7.8).

The ICS locus shown in Figure 7.8 displays the expected symmetry for an unlikely incident beam offset angle 2β of zero degrees. The peak of the ICS cusp characteristic touches the inscribed circle of the polygonal scanner and the locus extends beyond the circumscribed circle.

Figure 7.9 shows the asymmetry of the ICS locus for a realistic incident beam offset angle 2β of 70°. The peak of the ICS cusp characteristic lies within the inscribed circle of the polygonal scanner, while one extremity lies beyond the circumscribed circle and the other lies between the two circles. The tangent on the ICS locus at the data point $\theta = 0°$ corresponds to the scan-axis.

Figure 7.10 shows the asymmetry of the ICS locus for an incident beam offset angle 2β of 100°. The peak of the ICS cusp characteristic lies within the inscribed circle of the polygonal scanner, as does one extremity $\theta = +30°$, while the other extremity $\theta = -30°$, lies between the inscribed and the circumscribed circles. The tangent on the ICS locus at the data point $\theta = 0°$ corresponds to the scan-axis.

Figure 7.11 shows a more extreme asymmetry of the ICS locus for an incident beam offset angle 2β of 140° for a 12-facet polygonal scanner, $N = 12$. The peak of the ICS cusp

has disappeared because, in part, the range of the full mechanical scan angle θ has been reduced from ±30° to ±15° by virtue of the increased number of facets from six to twelve. The ICS locus extremities range from $\theta = +15°$ within the inscribed circle of the polygonal scanner to $\theta = -15°$ between the inscribed and the circumscribed circles. The tangent on the ICS locus at the data point $\theta = 0°$ corresponds to the scan-axis.

7.3.8 Locus of P

Axiomatically the locus of the point of incidence P lies along the incident beam line segment HP. Point P runs back and forth from the fixed point H on the incident beam, at which the circumference of the circumscribed circle of the polygonal scanner intersects. As the reflected incident beam scans through the full angular range $±A$ the locus of P overlaps itself.

The locus of P is inherently a straight line along the incident beam but doubles back on itself from and to the fixed point H. To provide visibility of this locus the data ordinate y/r values of Figure 7.10 have been mathematically and linearly stretched in Figure 7.12. For clarity the scale of the x/r axis is significantly magnified about tenfold. The markedly different spacing between data plots at two-degree intervals is indicative of a rapid acceleration and slow deceleration as the point of incidence P traverses the facet.

7.3.9 Offset Angle Limits

The incident beam is not likely to lie within the scan angle when the plane of incidence is normal to the axis of rotation; therefore, the smallest offset angle $[2\beta]_{min}$ for a beam with an infinitesimal diameter will always be equal to, or greater than, the semi full optical scan angle $+A$.

The maximum offset angle $[2\beta]_{max}$ of the incident beam with an infinitesimal diameter occurs when the incident beam is at grazing incidence for the semi full optical scan angle $-A$. Therefore, the upper limit to the offset angle will always be equal to, or less than (180°–A). Thus 2β lies in the range

$$\frac{360°}{N} \le 2b \le 180° \left(1 - \frac{2}{N}\right), \quad N \ge 4 \tag{7.53}$$

For real incident beams with a finite width (diameter) the minimum limit 360°/N will increase and the maximum limit 180°(l – 2/N) will decrease (see scan duty cycle η in Sections 7.2.6 and 7.4.8).

The expressions for the limits of the offset angles provide a useful guideline in the design of a scanning system. Although one will endeavor to design the incident beam to have an offset angle 2β to lie close to the semi full scan angle $+A$ of the reflected incident beam, there are occasions, for reasons of packaging, where this is impossible.

7.3.10 Finite Beam Width *D*

For simplicity the width of the beam has been assumed to be infinitesimal (Section 7.3.4) such that the objective lens optical axis is directed through the fixed point H on the incident beam where it is intersected by the circumscribed circle of the polygonal scanner in the midposition. If the incident beam has a finite width D, the radius r in Equation 7.45 would be replaced by $r(1 – g/r)$ because the incident beam shifts to the left to pass through

the point G. The dimensional symbol "g" is that shown in Figure 7.2 (Section 7.2). The r in Equation 7.46 remains unchanged.

For a finite beam width D the basic cusp-shape characteristic of the ICS loci that is shown in Figures 7.8 to 7.11 remains the same, but, with the exception of Figure 7.8 because of symmetry, it will be slightly displaced in an upward direction parallel to the facet plane by an amount ($g \tan \beta$) (Figure 7.2). The scan-axis is raised and the objective lens axis is lowered, each by an amount ($g \sin \beta$) with respect to the coordinate axes x/r and y/r. These displacement amounts are derived from Figure 7.2, Equations 7.21 and 7.15 in Sections 7.2.9 and 7.2.8, respectively.

7.3.11 Commentary

The ICS curves with offset angles greater than zero display interesting asymmetrical cusp-shape characteristics that offer potential insights into pupil movement that give rise to asymmetric aberrations in the image plane of pre-objective scanning systems.

7.3.12 Conclusion

An analysis of real beams of finite width (diameter) is fully expected to produce the same basic ICS characteristics as shown in the above documentation. The bottom line is that the ICS locus is of interest because it can give insight to the asymmetric wandering of the entrance pupil for the optical system lens designer who optimizes a design by minimizing the aberrations in the image plane regardless of the ICS locus.

7.4 STATIONARY GHOST IMAGES OUTSIDE THE IMAGE FORMAT

Ghost images are caused by both specular and scattered reflected rays from optical surfaces and are always unwanted, especially within the image format of the scanned field image plane. Various design innovations have been invented to minimize the effects or the presence of ghost images in the image format. Notable are those given in References 7 and 8 in which a limited angular range for the incident beam offset angle (2β) from the scan-axis is given so that stationary ghost images are formed outside the image format of the scanned field image plane of regular prismatic polygonal scanning systems.[6]

7.4.1 Objective

This section explores and illustrates the formation of stationary ghost images that are produced only by the scattered light rays from the scanned field image plane itself. The goal is to determine the angular ranges and limits of the incident beam offset angles 2β (beyond that given in References 7 and 8 mentioned above), with visual insights that ensure that the stationary ghost images lie outside the image format.

7.4.2 Stationary Ghost Images

Ghost images in the image plane from nonmoving optical components may be expected, but, at first thought, not from a rotating optical component such as a polygonal scanner, and if so, certainly not stationary. However, the rotating polygonal scanner itself synchronously derotates (descans) these unwanted diffusely reflected rays from the image

plane itself, and they are then specularly rereflected at the mirror facets. If these secondary specularly reflected rays are transmitted through the optics of the pre-objective optical scanning system, stationary ghost images will be formed in the image plane.

7.4.3 Facet Angle *A*

The facet angle *A* is the angle that the facet subtends at the rotation axis O:

$$A = \frac{360}{N} \tag{7.54}$$

where *N* represents the number of facets. For this section let *N* = 10. Then,

$$A = 36°, \quad \text{and} \quad 2A = 72° \tag{7.55}$$

7.4.4 Facet-to-Facet Tangential Angle

The mirror facet-to-facet tangential angle is the angle between successive facet normals in a plane perpendicular to the rotation axis. This angle is also denoted by the symbol *A*, because for a regular prismatic polygonal scanner the facet angle and the facet-to-facet tangential angle are geometrically identical.

7.4.5 Scan-Axis

The scan-axis is the axis about which the beam angularly scans symmetrically, ±*A* (see Sections 7.2.2 and 7.3.3).

7.4.6 Offset Angle 2*β*

The incident beam offset angle 2*β* is the angle that the incident beam makes with the scan-axis.

7.4.7 Midposition

The midposition of a scanner is that orientation of the polygonal scanner for which a facet reflects the incident beam collinearly with the scan-axis, which is parallel to the objective lens optical axis (see Sections 7.2.2 and 7.3.3, and Figure 7.1).

7.4.8 Scan Duty Cycle (Scan Efficiency) *η*

The maximum potential scan duty cycle *η* of a polygonal scanner is the ratio of the useful scan angle, during which the beam width *D* is effectively unvignetted by the edges of the facets, to the full scan angle ±*A* of a beam with an infinitesimal width (*D* = 0). We shall assume that the footprint of the beam's tangential width is less than the facet's tangential width in the midposition of a polygonal scanner.

$$h = 1 - \frac{\arcsin[D/(2r\cos b)]}{180/N} \tag{7.56}$$

in which r represents the radius of the circumscribed circle of the polygonal scanner (see Section 7.2.6). It can be seen from Equation 7.56 that for a given beam of finite width D the scan duty cycle η decreases with an increase in the offset angle 2β, or with an increase in the number of facets N.

7.4.9 Rotation Axis Offset Distance

The rotation axis offset distance is that distance Y_P of the rotation axis from the scan-axis when the polygonal scanner is set in its midposition (Figures 7.2 and 7.3).

The rotation axis offset distance Y_P depends on the number of facets N, the incident beam offset angle 2β, and beam width D. Replicating Equation 7.25 with a negative sign from Section 7.2 gives

$$Y_P = -r\sin b \left[2\cos\left(\frac{180}{N}\right)\right] - \sqrt{\left[1 - \left(\frac{D}{2r\cos b}\right)^2\right]} \tag{7.57}$$

in which r again represents the radius of the circumscribed circle of the polygonal scanner.

Alternatively, replicating Equation 7.27 embodying in the scan duty cycle η from Section 7.2 leads to

$$Y_P = -r\sin b \left[2\cos\left(\frac{180}{N}\right)\right] - \sqrt{\left(1 - \left\{\sin\left[\left(\frac{180}{N}\right)(1-h)\right]\right\}^2\right)} \tag{7.58}$$

For an infinitesimal beam width ($D = 0$) or a 100% scan duty cycle ($\eta = 1$), Equations 7.57 and 7.58 both reduce to

$$Y_P = -r\sin b \left[2\cos\left(\frac{180}{N}\right) - 1\right], \quad N \geq 3 \tag{7.59}$$

It can be seen from Equation 7.59, Table 7.3, and supported by Figures 7.13 to 7.16, that as the incident beam offset angle 2β increases, and/or the number of facets N increases, so also does the rotation axis offset distance increase Y_P.

When ($N = 3$), Equation 7.59 leads to $Y_P = 0$, which means the rotation axis lies on the scan-axis (Figure 7.4 and Section 7.2.13).

TABLE 7.3

Incident Beam Offset Angle 2β versus Maximum Potential Scan Duty Cycle η

Incident Beam Offset Angle, 2β	Maximum Potential Scan Duty Cycle, η	Rotation Axis Offset Distance, Y_P, Distance from the Scan Axis	Figure
27°	93.5%	−0.211r	7.13
52°	92.9%	−0.395r	7.14
92°	90.8%	−0.649r	7.15
124°	86.4%	−0.797r	7.16
164°	54.1%	−0.904r	—

7.4.10 Choosing an Incident Beam Offset Angle 2β

Ideally, for symmetry of design, the incident beam should be directed along the scan-axis, but this would obstruct the reflected scanning beam. Therefore, if the image format field angle is 2ω, then the incident beam offset angle 2β must at least be slightly greater than the semi-image format field angle ω to avoid this physical interference (Figure 7.13). That is,

$$2b > w \tag{7.60}$$

Using Equations 7.56 and 7.57 with $N = 10$, $r = 25$ mm, and $D = 1$ mm, leads to Table 7.3.

7.4.11 Ghost Beams gh and Images GH

Pencils of scattered light rays from the incident scanning spot on the scanned surface plane are returned through the objective lens. The objective lens recollimates them as they proceed back to the polygonal scanner's facets, at which they are specularly reflected to produce what are known as ghost beams. In Figures 7.13 through 7.16 these ghost beams are symbolized by the letters gh, with a subscript that identifies the facet whence they came.

Only if the ghost beams gh are reflected from a facet of the polygonal scanner at angles numerically much less than 90°, that is, toward the objective lens, is there a chance that they can traverse back through the objective lens, which will focus them onto the image plane to form stationary point ghost images GH. A subscript to GH refers to the respective facet whence the ghost beam gh came.

If these ghost beams are reflected from a facet of the polygonal scanner at angles numerically greater than 90°, that is, away from the objective lens, there is no chance of them traversing back through the objective lens to produce stationary ghost images GH.

7.4.12 Ghost Beam Field Angles ϕ

The field angle φ of all ghost beams gh, whether they produce stationary ghost images in the image format plane or not, is always at an angle that is a multiple of 2A away from, and on either side of, the incident beam offset angle 2β. This is because the mirror facet-to-facet tangential angle of a regular prismatic polygonal scanner is A. Thus

$$f = 2b \pm n(2A), \quad |2b \pm n(2A)| < 90° \tag{7.61}$$

in which n is an integer.

In Section 7.4.3, $N = 10$, therefore 24 = 72°; hence all ghost beams in Figures 7.13 through 7.16 occur at intervals of 72° from the incident beam offset angle 2β.

When $n = 0$, Equation 7.61 represents the retroreflective ghost beam field angle ϕ_1 that is collinear with the incident beam. Ghost beam gh_1 is not relevant in this discussion and, to avoid confusion, it is not depicted in the figures.

If one increases the incident beam offset angle 2β by say 25° counterclockwise, and repositions the polygonal scanner to a midposition, all the field angles ϕ of the ghost beams will have also rotated by 25° counterclockwise, while the polygonal scanner will have only rotated 12.5° counterclockwise; and vice versa, if clockwise.

7.4.13 Incident Beam Location

Figures 7.13 through 7.16, which shall be discussed in turn, depict the first significant four incidence beam offset angular positions given in Table 7.3, namely, 27°, 52°, 92°, and 124°.

Figures 7.13 through 7.16 depict the respective orientations of the polygonal scanner in the midposition and the rotation axis offset distances Y_P, such that the reflected incident beam is collinear with the scan-axis to focus it to the central point C in the image field format. Pencils of light rays, gh_2, gh_3, gh_4, and gh_{10}, scattered from point C, are shown passing back through the objective lens to meet the facets of the polygonal scanner, whence they are again reflected. The subscripts correspond to the facet number.

In Figure 7.13 one such pencil, gh_2, passes back through the objective lens to produce the point image GH_2 below the image field format. There is one pencil gh_1, which is reflected from facet S_1 that is not displayed so as not to overcrowd the diagram. Pencil gh_1 is the retroreflective pencil that returns collinearly with the path of the incident beam.

As predicted by Equation 7.61 the angle between successive pencils of rays of ghost beams gh reflected from the five facets S_{10}, S_1, S_2, S_3, and S_4 is $2A$ (Figures 7.13 to 7.16).

One should notice that in Figures 7.13 to 7.16 the vertices of the fan depicting the full scan angle $\pm A$ and the image field format scan angle $\pm \omega$ do not coincide, nor do they touch the surface of the facet. They lie at two distinct locations. The first lies on the incident beam axis at its intersection with the circumscribed circle of the polygonal scanner; the second lies below, within the circumscribed circle and above the scan-axis. The difference is best observed in Figure 7.16 (see also Section 7.3).

7.4.14 Image Format Scan Duty Cycle η_ω

The image format scan duty cycle η_ω, is the ratio of the image field format angle 2ω to the full scan angle $2A$ of the polygonal scanner. It must not be confused with the maximum potential scan duty cycle (scan efficiency) η. The image format scan duty cycle η_ω depends directly on the image format angle 2ω.

$$h_w = \frac{2w}{2A} = \frac{w}{A} \tag{7.62}$$

In Figures 7.13 to 7.16, $2\omega = \pm 20°$ (40°). Substituting for ω and A into Equation 7.62 leads to

$$h_w = \frac{20°}{36°} = 55.6\% \tag{7.63}$$

Since the above image field format scan duty cycle η_ω of 55.6% is greater than the maximum potential scan duty cycle η of 54.1% presented in Table 7.3 for an incident beam offset angle 2β of 164°, this offset angle is not relevant and no figure is provided. The image field format scan duty cycle η_ω must be less than the scan duty cycle η.

7.4.15 Incident Beam Offset Angle 27°

The incident beam offset angle of 27° in Figure 7.13 is comfortably outside the semi-image format angle ω of +20° to avoid physical obstruction of the scanning beam, but at an angle less than the half scan angle $A = +36°$, of the ten-facet polygonal scanner.

Ghost image for $2\beta = 27°$
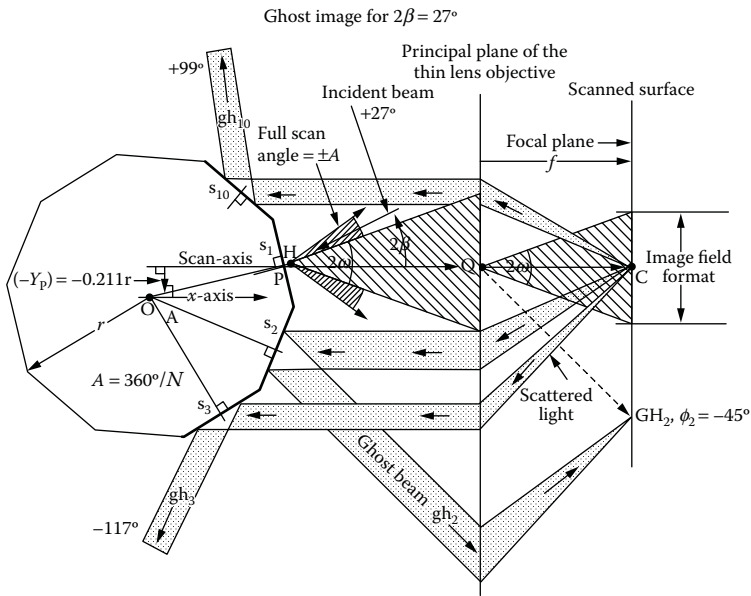
FIGURE 7.13
Formation of stationary ghost image GH_2 is produced by a pencil of scattered rays originating at C and rere-flected from facet S_2 at a field angle $\phi_2 = -45°$.

If the ghost beam gh_2 traverses the objective lens, there is only one stationary ghost image, GH_2, at a field angle of $\phi_2 = -45°$; and it lies outside and 25° below the image field format.

From Equation 7.61 the field angles of ghost beams gh_{10} and gh_3 from facet S_{10} and S_3, are $\phi_{10} = +99°$ and $\phi_3 = -117°$, respectively. These ghost beams are harmless ($|f| > 90°$).

7.4.16 Incident Beam Offset Angle 52°

In Figure 7.13 the incident beam offset angle is 27°. Let the incident beam offset angle 2β with its accompanying ghost beams gh and ghost images GH be rotated counterclockwise through a positive angle of +25°. If the ghost beam gh_2 traverses the objective lens, the ghost image GH_2, $\phi_2 = -45°$, of Figure 7.13 will move up to lie on the lower edge ($-\omega = -20°$) of the image format field, $\phi_2 = (-45° + 25°) = -20°$. The incident beam offset angle increases to $2b = (+27° + 25°) = +52°$ (Figure 7.14).

From Equation 7.61 the field angles of ghost beams gh_{10} and gh_3 from facets s_{10} and s_3 become $\phi_{10} = +124°$ and $\phi_3 = -92°$, respectively. These ghost beams are harmless ($|f| > 90°$).

7.4.17 Incident Beam Offset Angle 92°

In Figure 7.14 the incident beam offset angle is 52°. Let the incident beam offset angle 2β with its accompanying ghost beams gh and ghost images GH be rotated counterclockwise through a positive angle of +40°. If the ghost beam gh_2 traverses the objective lens, the ghost image GH_2, ($\phi_2 = -\omega = -20°$), of Figure 7.13 will move up to lie on the upper edge
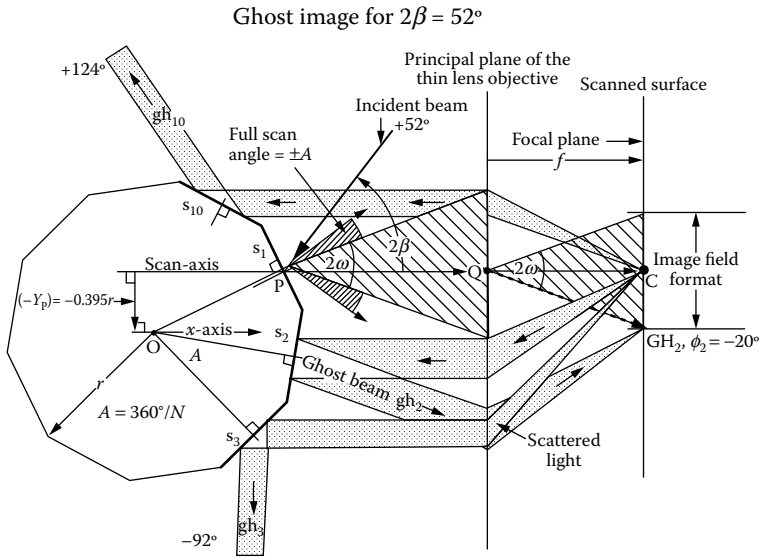
FIGURE 7.14
The stationary ghost image GH_2 at the lower edge of the image field format at a field angle $\phi_2 = -20°$ is produced by a pencil of scattered rays originating at C and rereflected from facet S_2.
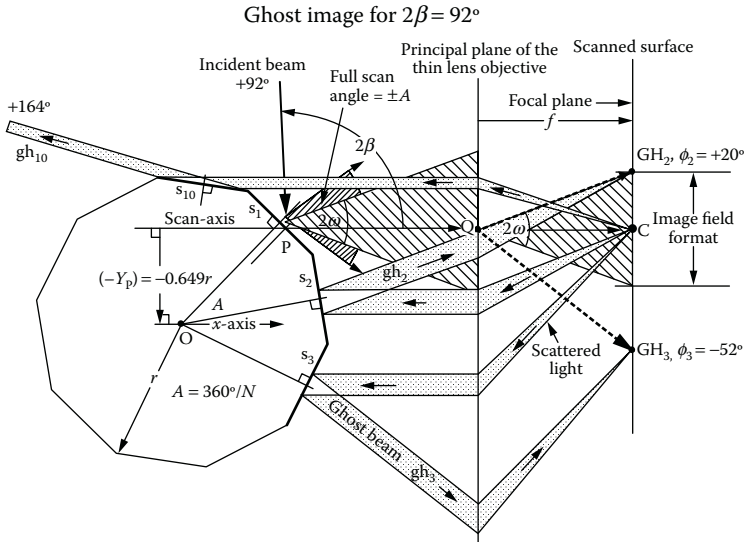
FIGURE 7.15
If the ghost beams gh_2 and gh_3 traverse the objective lens, there is a stationary ghost image GH_2 at the upper edge of the image field format $\phi_2 = +20°$ and a stationary ghost image GH_3 below it, $\phi_3 = -52°$.

$(+\omega = +20°)$ of the image format field, $\phi_2 = (-\omega + 40°) = (-20° + 40°) = +20°$. The incident beam offset angle increases to $2b = (+52° + 40°) = 92°$ (Figure 7.15).

For incident beam offset angles 27° and 52° there is only one stationary ghost image in the image format, namely GH_2. As ghost image GH_2 moves to the upper edge of the image format, a second ghost image GH_3 appears in the image format well below at a field angle, $\phi_3 = -52°$. Note that $(\phi_2 - \phi_3) = 2A = 72°$, as expected.

From Equation 7.61 the field angle of the remaining ghost beam gh_{10} from facet s_{10} is $\phi_{10} = +164°$, and is harmless ($|f| > 90°$).

7.4.18 Incident Beam Offset Angle 124°

In Figure 7.15 the incident beam offset angle is 92°. Let the incident beam offset angle 2β with its accompanying ghost beams gh and ghost images GH be rotated counterclockwise through a positive angle of +32°. If the ghost beam gh_3 traverses the objective lens, the ghost image GH_3, $\phi_3 = ⊠52°$, of Figure 7.15 will move up to lie on the lower edge ($⊠\omega = ⊠20°$) of the image format field, $\phi_3 = (⊠52° + 32°) = ⊠20°$. The incident beam offset angle increases to $2b = (+92° + 32°) = +124°$ (Figure 7.16).

For incident beam offset angles 52° and 92° there are two stationary ghost images, namely GH_2 and GH_3. As ghost image GH_3 moves to the lower edge of the image format, ghost image GH_2 moves up well above the image format at a field angle, $\phi_2 = +52°$. Again note that $(f_2 - f_3) = 2A = 72°$, as should be expected. From Equation 7.61 the field angle of the remaining ghost beam gh_4 from facet s_4 is $\phi_4 = ⊠92°$, and is harmless ($|f| > 90°$).

A simple calculation of adding 72° to the field angle $\phi_{10} = 164°$ of ghost beam gh_{10} in Figure 7.15 produces a reflex angle of 236°, thus predicting that the ghost beam gh_{10} can no longer exist.

A close inspection of Figures 7.13 to 7.16 shows the center-of-scan of the total angular scan $2A$ of the scanner progressively becomes displaced from the center-of-scan of the image format scan angle 2ω with an increase in the incident beam offset angle β. This is valid because the ICS is a locus (see Section 7.3 and Reference 5).

Ghost image for $2\beta = 124°$
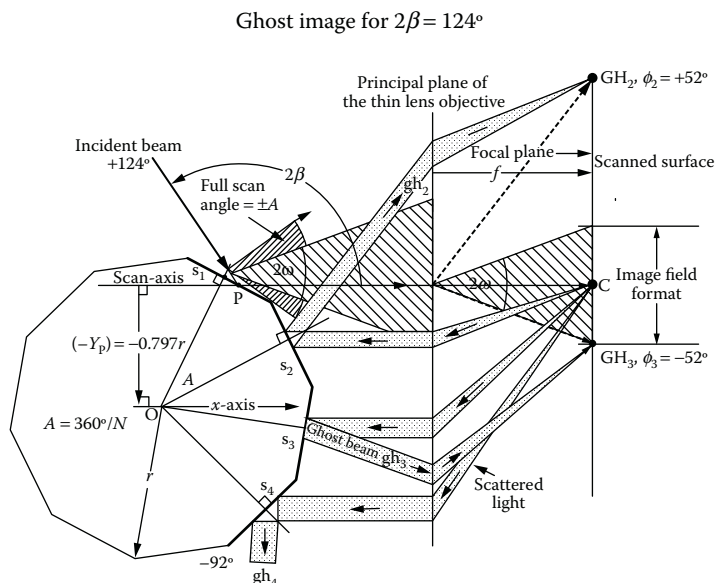
FIGURE 7.16
If the ghost beams gh_2 and gh_3 traverse the objective lens, there is a stationary ghost image GH_3 at the lower edge of the image field format $\phi_3 = ⊠\omega = ⊠20°$ and a stationary ghost image GH_2 above, $\phi_2 = +52°$.

7.4.19 Ghost Images Inside the Image Format

A study of Figures 7.14 and 7.15 shows that, if ghost image GH_2 is set on the lower and upper edges of the image format, the required incident beam offset angles are given by

$$2b = 2A + w \qquad (7.64)$$

Thus, from Equation 7.61 the range of 2β for ghost images to exist inside the image format is expressed by

$$n(2A) - w < 2b < n(2A) + w \qquad (7.65)$$

in which n is zero or a positive integer, $A \geq \omega$, and $2\beta < 180°$.
Substituting $\omega = 20°$ and $2A = 72°$ leads to

$$\text{when } n = 0 \quad -20° < 2b < +20° \qquad (7.66)$$

$$n = 1 \quad +52° < 2b < +92° \qquad (7.67)$$

$$n = 2 \quad +124° < 2b < +164° \qquad (7.68)$$

Each has a range of 40°, which, not surprisingly, equates to 2ω.
A figure showing $2\beta = 164°$ is not relevant or depicted, because it has a scan duty cycle η less than the required image format duty cycle η_ω (Table 7.3).

7.4.20 Ghost Images Outside the Image Format

A study of expressions (7.66), (7.67), and (7.68) shows that when the incident beam offset angle lies between +20° and +52° no ghost image will appear in the image format. Likewise, when the incident beam offset angle lies between +92° and +124°, each has a range of 32°.[2]
Thus, to ensure ghost images lie outside the image format the condition is as follows:

$$n(2A) + w < 2b < (n+1)(2A) - w \qquad (7.69)$$

in which n is zero or a positive integer, $A \geq \omega$, and $2\beta < 180°$.
Let ρ represent the angular range of 2β for ghost images outside the image format, then

$$r = 2A - 2w = 2(A - w) = 2(180/N - w) \qquad (7.70)$$

and is independent of n.

7.4.21 Number of Facets

Subject to $A \geq \omega$, as the number of facets N increases, so also does the number of ghost beams gh and, therefore, there is a greater possibility of multiple ghost images GH in the scanned image plane. A critical case, in this example, occurs when $N = 18$. Then $A = +\omega = +20°$.

Substituting these values for A and ω into the inequalities (69) and (70) leads to Figure 7.17.

$$\text{For } n = 0, \quad 20° < 2b < 20° \tag{7.71}$$

$$n = 1, \quad 60° < 2b < 60° \tag{7.72}$$

$$n = 2, \quad 100° < 2b < 100° \tag{7.73}$$

and the range

$$r = 0°$$

Hence, the positioning tolerance for the incident beam offset angle 2β is zero. For an adequate positioning tolerance for the incident beam

$$A > w \tag{7.74}$$

Substituting $A = 360°/N$ leads to the general condition

$$\frac{360}{N} > w \tag{7.75}$$
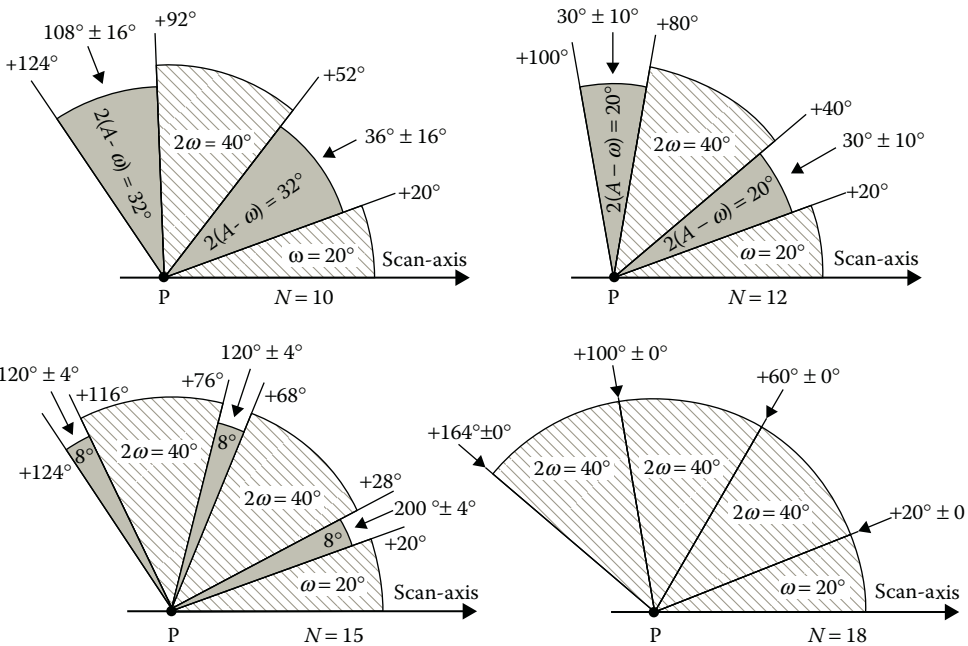
FIGURE 7.17
To ensure that stationary ghost images GH are outside the image format, the angular ranges ρ of the incident beam offset angle 2β are shown at 32°, 20°, 8°, and 0°. The illustration also shows the angular ranges ρ for 10-facet, 12-facet, 15-facet, and 18-facet polygonal scanners.

In Figure 7.17, P represents the point of incidence on the scanner facet. The image format field angle shown is $2\omega = 40°$. For the 18-facet polygon the angular range is zero, but theoretically available, and would simultaneously produce ghost images GH_2 and $GH_{18} = $ at the upper, $\phi_2 = +\omega$, and lower, $\phi_{18} = \boxtimes\omega$, edges of the image format, respectively, when the incident beam offset angle is $2b = +w, +3w, +5w, +7w$, and so on, subject to $2\beta < 180°$.

7.4.22 Diameters of Scanner and Objective Lens

No mention has been made with respect to the diameters of the objective lens, the scanner, nor the apertures near the scanner, or performance. These topics are out of the scope of this section, but all are important issues.[2]

However, the smaller the diameter of the scanner relative to the objective lens diameter, the greater the chance of a ghost beam returning to produce a ghost image in the scanned image plane. Likewise, the closer the scanner is to the objective lens, the greater the chance of a ghost beam returning to produce a ghost image in the scanned image plane.

7.4.23 Commentary

There is more than one angular zone for the incident beam offset angle to avoid ghost images appearing within the image format. These zones have acceptable scan duty cycles η, depending on beam width D, the diameter $2r$ of the polygonal scanner and the number of facets N (Figure 7.17, Table 7.3).[2]

7.4.24 Conclusion

It behooves one to consider the possibility and the whereabouts of stationary ghost images in the image format plane during the initial optical system design stage.

REFERENCES

1. Kessler, D.; DeJaeger, D.; Noethen, M. High resolution laser writer. *Proc. SPIE*, 1989, *1079*, 27–35.
2. Beiser, L. *Unified Optical Scanning Technology;* IEEE Press, Wiley-Interscience, John Wiley & Sons: New York, 2003.

3. Beiser, L. Design equations for a polygon laser scanner. In *Beam Deflection and Scanning Technologies;* Marshall, G.F.; Beiser, L., Eds.; *Proc. SPIE* 1991, *1454*, 60–66.

4. Marshall, G.F. Geometrical determination of the positional relationship between the incident beam, the scan-axis, and the rotation axis of a prismatic polygonal scanner. In *Optical Scanning 2002;* Sagan, S.F.; Marshall, G.F.; Beiser, L., Eds.; *Proc. SPIE* 2002, *4773*, 38–51.

5. Marshall, G.F. Center-of-scan locus of an oscillating or rotating mirror. In *Recording Systems: High-Resolution Cameras and Recording Devices; Laser Scanning and Recording Systems*; Beiser, L.; Lenz, R.K., Eds.; *Proc. SPIE* 1993, *1987*, 221–232.

6. Marshall, G.F. Stationary ghost images outside the image format of the scanned image plane. In *Optical Scanning 2002;* Sagan, S.F.; Marshall, G.F.; Beiser, L., Eds.; *Proc. SPIE 4773*, 132–140.

7. U.S. patent no. 5,191,463, 1990. Scanning optical system, in which ghost image is eliminated.

8. U.S. patent no. 4,993,792, 1986. Scanning optical system, in which ghost image is eliminated.

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

8

Galvanometric and Resonant Scanners

Jean Montagu

Engineering Consultant
Cambridge, Massachusetts, USA

CONTENTS

8.1 INTRODUCTION

The goal of this section is to offer the reader a comprehension of the parameters that shape the design and subsequently the applications of current oscillating optical scanners. Hopefully this will explain their design and possibly guide system engineers to reach the most desirable compromise between the numerous variables available to them.

It is also my hope that this may stimulate designers to extend the technology or pursue different technologies as they appreciate the constraints and limitations of current oscillating optical scanners and their applications.

This text is the third of a series edited by Gerald F. Marshall[1,2] covering the evolving field of optical scanning. Since oscillating scanners are developed to meet the needs of specific technical and scientific applications, it is constructive to review some of these applications. Applications are the stimulus that underlies past and future developments.

It is evident that the material presented here is evolutionary and a broader treatment can be found in the references. The reader is frequently referred to the previous texts.[1,2] Only Section 8.2.1.4, "Mirrors," and Section 8.2.23, "Induced Moving Coil Scanner," are reproduced here in toto. These are important subjects that are frequently disregarded by system designers, and no meaningful advances have taken place. On occasion, some material is taken from the previous editions in order to present the new material in a consistent manner. Section 8.2.1.3, "Bearings," as well as Section 8.2.16, "Dynamic Performances," contains some material from the previous edition as well as new material.

In addition, this edition is inversely organized when compared with its precursors. The technology underlying the components of scanners is reviewed up front and new applications are at the end of the section. Important evolutions of older applications are presented generally, while referring the reader to earlier texts for basic descriptions of the subject.

The past decade has seen extensive technology evolutions that have brought major changes in the market and manner of use of scanners as well as unexpected performances and designs of oscillating scanners. Improved performances of competing technologies have attracted applications previously the domain of oscillating mechanical scanners. Linear and two-dimensional solid-state arrays now dominate the vision and the night vision market, both military and commercial. Digital micromirror devices (DMDs) and liquid crystal displays (LCDs) have also captured the field of image projection away from oscillating scanners.

On the other hand, the advances of computer control in industry have benefited the laser micromachining industry, which requires a high degree of flexibility and is well adapted to digital control. This has benefited galvanometric scanner manufacturers, who have responded by improving their product.

Improved scanner performances as well as greater choice and more economical associated technologies have broadened the market for scanners and stimulated new applications. This in turn has offered opportunities for new sources of supplies.

8.1.1 Historical Developments

The galvanometer is named after the French biologist and physicist Jacques d'Arsonval, who devised the first practical galvanometer in 1880. Initially it was used as a static measuring instrument. Its dynamic and optical scanning potential were recognized early on when galvanometers were employed to write sound tracks on the talking movies. Miniature galvanometers with bandwidths as high as 20 kHz were used for waveform recording on UV-sensitive photographic paper as late as 1960.

The invention of the laser broadened the applications of galvanometers in the graphic industry during the late 1960s. The first designs were open loop scanners, but very early in the 1970s, the position-servoed, better known as the closed loop scanner, came to reign in meeting the desire for more bandwidth and increased accurate positioning.

The servoed scanner enabled the accuracy of the device, relegated to the position transducer, to be dissociated from the torque motor. The next challenge was to minimize inertia and optimize rigidity. Cross-talk perturbations were mostly solved with the use of moving magnet torque motors and the practice of balancing the load and armature.

Demand for higher speed and greater accuracy forced the design of all the building blocks of scanning systems to be refined. The performance of scanners evolved along the evolution of its constituents: torque motor, transducers, amplifiers, and computers.

The first milestone in the early 1960s was the development of moving iron scanners, as they offered a compact magnetic torque motor. The compact, efficient, and economical design offered scanners beyond the capabilities of moving coils at that time.

The second milestone in the late 1980s came as a consequence of the commercialization of high-energy permanent magnets. Moving magnet torque motors were developed with much greater peak torque. In the same period, a new design of transducers appeared, driven by the availability of much improved electronic elements.

The third milestone that came into being in the last decade of the century is marked by the presence of computer power to mitigate the shortcomings of even the best galvanometers. The clock rate of ordinary PCs has reached the megahertz range and can compensate in real time for position encoder imperfections as well as optimize dynamic behavior of periodic and aperiodic armature motions. The PC has also simplified full system integration.

High-energy permanent magnets were also developed to power resonant scanners, but innovative new designs and the use of PCs form the underpinning of present-day devices.

At this writing, all high-performance optical scanners share a common architecture: a moving magnet torque motor, a position transducer built along a variable capacitor ceramic butterfly for high-precision work and optical sensors for less demanding applications. The performance of the galvanometric scanner is limited by the following parameters, which shall be covered in more detail in the following sections:

1. The thermal impedance of the magnetic structure and specifically the drive coil. This in turn limits the available torque of the magnetic motor and induces unpredictable thermal drift of the position transducer.

2. The thermal stability of the position transducer.

3. The mechanical resonances of the armature and the load as they prevent the system from achieving a step response expected from available torque. An expert servo designer can appreciably optimize system performances if the elements are stable.

The ability to integrate all these disciplines, as well as optimum frame configuration, mirror design and mounting, drift compensation, and software has become a specialty so that a fully integrated subsystem is frequently selected rather than just the scanner. Figure 8.1 illustrates all these elements.

The progress of scanners is application driven; consequently more recent applications will be reviewed in the last section.

8.2 COMPONENT AND DESIGN ISSUES

Scanners are like old soldiers; they fade away but never die. Moving coil and moving iron scanners, as well as some four-pole stepper motors built with laminated stators, are still available and priced competitively as their tooling has been amortized. Their design and performances have not evolved since they were evaluated previously by this author[1,2] and will not be reviewed here. Their desirable features do not compensate for their shortcomings such as iron saturation, nonlinearity, and strong unpredictable radial forces, the major cause of wobble, for the iron-based scanners and flexible armature and poor thermal properties for the moving coil units as well as high cost of entry and manufacturing.

8.2.1 Galvanometric Scanners

All modern high-performance oscillating galvanometer scanners are built with a moving magnet torque motor and all high-performance position transducers employ a two- or four-lobe ceramic variable capacitor.

All oscillating scanners designed in the last decade are built using the NdFeB family of permanent magnets. These alloys can have as much as five times the energy product of the best ALNICO magnet and certain other benefits, but they have a low Curie temperature, possibly as low as 310°C, such as for the 45 MGO material from Ugimag.[3]

The higher the energy product, the lower the Curie temperature.[3] This has two important consequences:

1. A typical magnetic strength temperature coefficient of −0.8%/°C in the range 22°C–85°C.
2. An irreversible flux loss will take place each time the material is heated above 80°C–100°C. The range reflects the particular alloy selected and the magnetic design.

The coil design and its thermal conductivity are critical features of the galvanometric scanner because they are the major cause of transducer thermal drift.

8.2.1.1 Moving Magnet Torque Motor

The torque motor is selected for its ability to integrate with the other elements of the scanner, the mirror, the position sensor, and the electronic driver/controller. It must also support the dynamic performance requirements and those caused by environmental changes and perturbations.
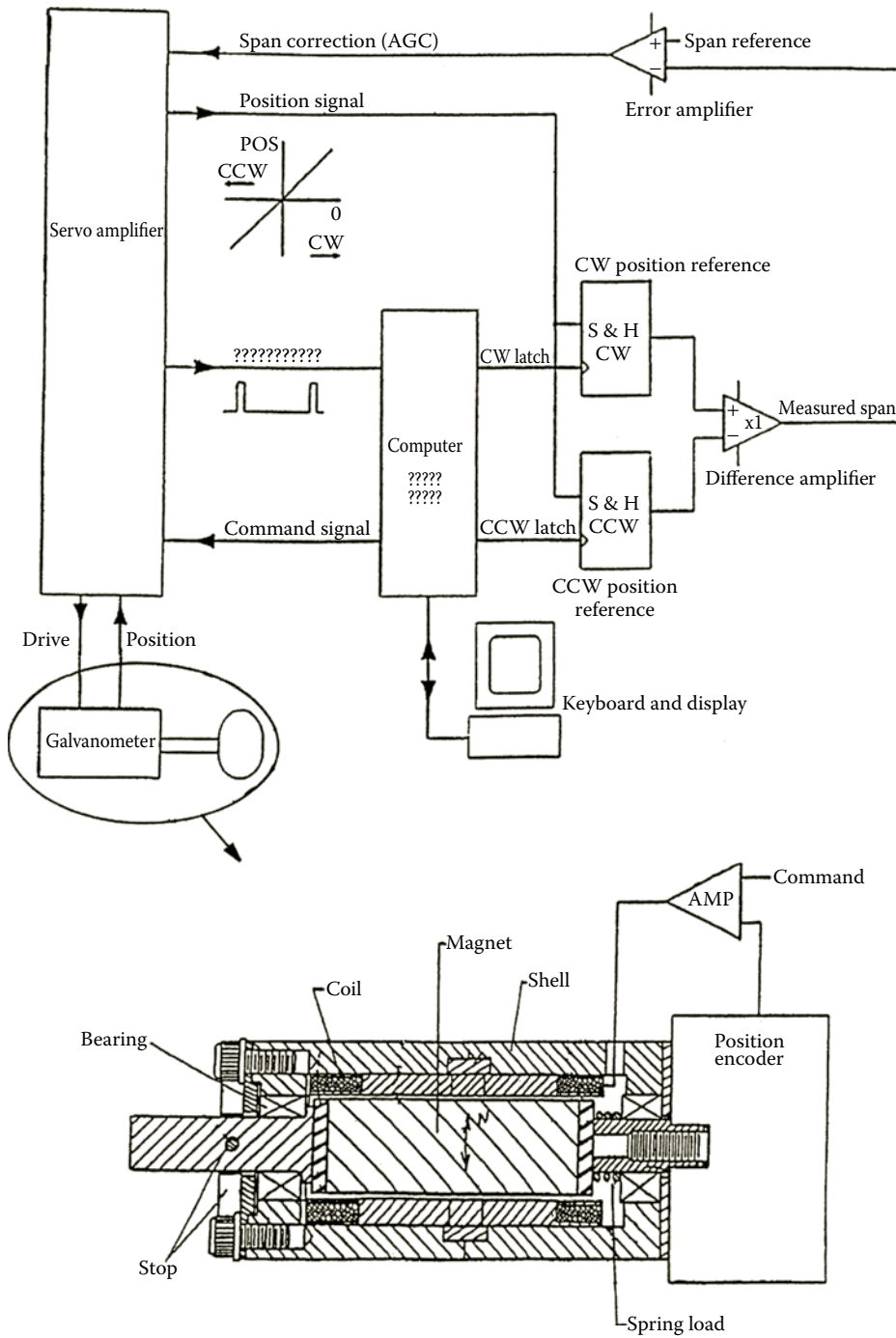
FIGURE 8.1
Galvanometric scanner and system management.

The list of features of an ideal torque motor is long and frequently a compromise is reached where some necessary system properties are obtained through other means. Environmental control and electronic compensation schemes have become standard features of high-performance scanners.

The ideal galvanometric or resonant scanner driver would have the following properties:

1. High torque-to-inertia ratio.
2. Low electrical time constant: inductance/resistance.
3. Linear relationship between torque, current, and angular position.
4. Freedom from cross-axis forces or excitations.
5. No hysteresis or discontinuities.
6. No elastic restraint.
7. Some mechanical damping, constant and uniform is acceptable.
8. Very high rigidity in torsion and bending.
9. A balanced armature.
10. Low power consumption [figure of merit; torque/(inertia*Watt$^{1/2}$)].
11. Immunity to thermal expansion constraints.
12. Good heat dissipation.
13. Demagnetization protection.
14. Simplicity of installation and use.
15. Insensitivity to radio frequency (RF) and other environmental perturbations.
16. Absence of sensitivity to external perturbations.
17. Freedom from self-induced perturbations.
18. Infinite life with stable parameters.
19. Small, light, and cheap.

Additionally, for a resonant scanner, damping properties should be minimum along the axis of rotation and high for all other degrees of freedom.

8.2.1.1.1 Moving Magnet Torque Motor

The high energy of rare earth magnets is practically free of radial forces, which makes them attractive torque motors. They are the choice of all devices listed in Tables 8.1, 8.2, and 8.3. They are large-air-gap devices with comparatively low inductance, and simple to interface with electrically.

The driving stage is a conventional inside–out d'Arsonval movement, as shown in Figure 8.2. The torque can be calculated as the interaction of two fields or the effect of a field on a current. We shall follow the latter method.

Equation 8.1 is derived for coils with a total number of turns N and with conductors at 45° from the plane of symmetry. Equation 8.2 gives the torque generated by a device built with a coil as shown in Figure 8.3 having a uniform conductor density distribution ±45° from the plane of symmetry. This assumes that D is the average diameter of the coil. Also the coil needs to be of tightly wound coils and facing the region of highest magnetic field of the rotor:

$$T = 0.90 K B_r L N I D \cos g$$

$$(8.1)$$

TABLE 8.1

Comparative Performances of Moving Magnet Scanning Galvanometers: Inertia <1 gm cm²

	Model						
	6200	**6210**	**6220**	**RZ-15**	**6860**	**TGV-1**	**6230**
Torque motor							
Rotor inertia, gm cm²	0.012	0.02	0.14	0.34	0.6	0.65	1
Torque constant, gm cm/A	10.8	25	57	40	93	123	114
Resistance, Ω	2.4	4.1	3.4	1.3	1.5	1.4	1.4
Thermal conductivity, °C/W	7.5	4	2		1.5		1
Figure of merit of torque motor							
Torque/inertia (Watt)$^{1/2}$	580	625	221	103	126	160	96.3
Transducer	Opt.	Opt.	Opt.	Cap.	Cap.	Cap.	Cap.
Sensitivity, μA/°Opt.	24	24	22.8	100	29	50	23.4
Gain drift, ppm/°C	75	75	75	50	50	b	25
Null drift, μrad/°C	50	50	50	25	30	b	150
Repeatability, μrad	30	30	30	6	16	4	30
Dynamic performance							
Small angle step response, ms	0.175	0.175	0.25	0.25	0.5	0.18	0.3

ᵃ Angular excursion: All scanners are rated 60° optical pick to pick (ptp), mechanical motion, minimum.
ᵇ Not applicable, scanner has internal feducial references.
ᶜ All angles are in degree optical.
ᵈ All optical detector have linearity >98% and all capacitive detectors have linearity >99.5%.

The outer shell closes the magnetic circuit of the permanent magnet as well as that of the drive coil. It is preferably made of sintered high density 50/50 nickel–iron alloy. Low-carbon cold rolled steel such as C1020 and other steels have similar magnetic properties, but are rarely more economical as the finished part. Its radial thickness is recommended to be about one-quarter the diameter of the rotor when a rare earth magnet is used as the rotor.

The constant K in Equation 8.1 takes into consideration the space allocated for the coil with respect to the dimension of the magnet. It is expressed as

$$K = \frac{1}{1 + 2g \cdot B_r/mH_c \cdot d} \tag{8.2}$$

where for rare earth magnets

$$\frac{B_r}{mH_c} = 1.1 \tag{8.3}$$

B_r is the magnetic intrinsic induction, remanence of the rotor material; H_c is the magnetic demagnetization force, coercive force, of the rotor material; μ is the permeability of air; g is the radial gap between the rotor and the outer shell; and d is the diameter of the magnet. In practice, $1/2 < K < 1$. It is evident that for a given number of coil windings and a given resistance, it is advantageous to minimize the radial gap.

Rare earth magnets have a very high intrinsic coercive force and they are practically impervious to operational demagnetization so that extremely high torques can be safely generated. The torque is only limited by the coil design and construction as it relates to the

TABLE 8.2

Comparative Performances of Moving Magnet Scanning Galvanometers: Inertia > 1 gm cm²

	Model									
	M2	**VM2000**	**6870**	**TGV-2**	**6450**	**M3**	**RZ-30**	**6880**	**TGV-3**	**TGV-4**
Torque motor					Note 5					
Rotor inertia, gm cm²	1.7	1.7	2	2.3	2.4	4	5.9	6.4	7.4	14
Torque constant, gm cm/A	230		180	335	450	500	278	254	550	650
Resistance, Ω	4.5		1.4	2.6	4	4.8	3	1	2.6	2.7
Thermal conductivity, °C/W	2.5		1		5	1.4		0.75	0.7	0.7
Figure of merit of torque motor[e]										
Torque/inertia* (Watts)$^{1/2}$	63.8		76	90	93.7	57	27.2	39.7	46.1	28.3
Transducer										
Sensitivity, μA/° Opt.	11		29	100	43	11	150	44	100	100
Gain drift, ppm/°C	−60	100	50	b	50	60	30	50	b	b
Null drift, μrad/°C	18	30	30	b	30	18	10	20	b	b
Repeatability, μrad	12		16	2	4	2	2	16	2	2
Dynamic performance										
Small angle step response, ms		0.3		0.7	0.3			0.6		

ᵃ   Angular excursion: All scanners are rated 60° optical ptp, mechanical motion, minimum.
ᵇ   Not applicable, scanner has internal feducial references.
ᶜ   All angles are in degree optical.
ᵈ   All optical detectors have linearity >98% and all capacitive detectors have linearity >99.5%.
ᵉ   Moving coil torque motor.
ᶠ   All transducers are capacitive detectors.

cooling capability to prevent either catastrophic failure or excessive thermal drift of the position transducer. Coil design and transducer designs are the critical features of galvanometric scanners as well as armature rigidity. These elements shall be reviewed in the following sections.

8.2.1.1.2 Coil Construction

The use of the space allocated to the coil windings in electromagnetic devices has been the subject of numerous texts and studies since electromagnetic devices have been made. Roters[4] gives a general description of the subject and early patents[5] demonstrate the appreciation that efficient packing density of coils as well as their thermal construction are critical features. Hodges[6] gives a detailed description of the benefits that can be derived from pressing a coil of round conductors into a minimum cross section. He reports that thermal conductivity can be increased by a factor of 3 after compression.

Roters[4] shows that "resistance density" and "thermal conductivity" of a coil are nearly proportional to the "copper density" of the coil. Optimizing the copper density of a coil without prejudicing its reliability or cost has been the pursuit of optical scanner manufacturers as these devices are regularly driven extremely hard.

TABLE 8.3

Performances of Moving Magnet Galvanometers with Flexure Bearings

	Model		
	Harmonicscan	**FM200**	**Slowscan**
Torque motor			
Rotor inertia, gm cm^2	0.3	2.5	8.25
Torque constant, gm cm/A	120	230	278
Resistance, Ω	1.3	4.5	5.5
Transducer			
Sensitivity, μA/° Opt.	70		90
Gain drift, ppm/°C	50	100	50
Null drift, μrad/°C	25	30	25
Repeatability, μrad	5	1	2
Suspension			
Jitter, μrad	4	1	1.7
Wobble, μrad	1	0.5	0
Performances			
Small angle step response, ms	0.2	0.6	1.3

[a] Angular excursion: All scanners are rated 60° optical pick to pick (ptp), mechanical motion.
[b] All capacitive detectors have linearity >99.5%.
[c] All angles are in degree optical.
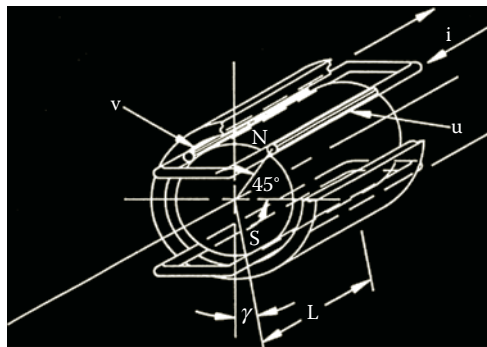
FIGURE 8.2
Inverted d'Arsonval movement.

The thermal impedance of a coil is mostly defined by the electrical insulation of the conductor, the copper density as well as the encapsulation compound. The volume of insulation and encapsulation needs to be minimized. Single insulation of magnet wire occupies about 20% of the volume of the typical conductor used for these devices.

Packing efficiency has been recognized as another critical factor. The highest possible local packing density of large coils with round conductor wound in quincunx is 90.69% and layer winding reaches a maximum of 78.5%. Most common windings have a packing density under 60%, which consequently yields a copper density under 50%.

Two technologies have been developed to improve the copper packing density as well as the thermal impedance of galvanometer coils. One technique compacts the coil and is best described by Hodges[6] and Houtman.[7]
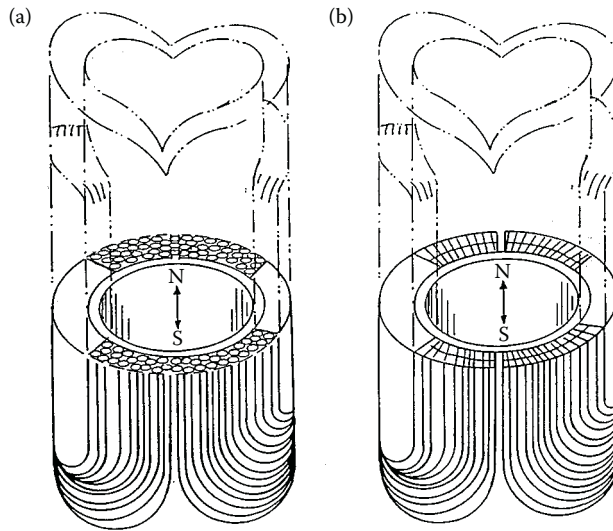
FIGURE 8.3
(a) Coil wound with round wire conductors; (b) coil wound with ribbon wire conductors.

They recommend first preforming a coil with conventional single insulation round wire plus self-adhesive. This preformed coil is then compressed to conform to its final shape and achieve a higher packing density. A suitable choice of insulation and ductile conductor can improve the copper density by 20%, comparable to a quincunx winding. That indicates that when the base coil has a random winding, an increase in copper density from 50% to 70% copper density is possible.[7]

A slightly higher copper density can be achieved with low aspect ratio ribbon conductor. This construction in addition offers the best thermal impedance and a simpler construction. Figure 8.3 compares conventional coil and ribbon coil constructions with about equal resistance. This construction can yield a coil with thermal conductivity four times higher than can be obtained with a conventional random wound coil and 50% higher than can be achieved with a compressed coil construction.

8.2.1.1.3 Heat Dissipation

The heat dissipation constant is a critical parameter as the temperature rise of the scanner has numerous consequences with possible thermal runaway and catastrophic consequences as described earlier in this section. The elements directly affected by a temperature rise are:

1. The coil temperature as its resistance increases with temperature rise as:

$$R_{\mathrm{T}} = R_{25}(1 + 0.0039\Delta T)$$

(8.4)

   where ΔT represents the temperature change from the 25°C resistance value.

2. The thermal conductivity of the coil has major consequences for the temperature of the magnet as it is located at the center of the coil. The average coil temperature rise may be modest, but it may not be representative of the temperature at the center of the coil and the temperature of the magnet.

3. The temperature stability of a hard magnetic material is inversely proportional to its energy product. The ALNICOs are the most stable, followed by the samarium compounds and finally the NdFeB alloys. Most scanners are built with the most energetic NdFeB material with a high negative temperature sensitivity such that, as a first approximation, the magnetic field can be derived, from Reference 3 data, for 22–85 °C from:

$$B_T = B_{22}\left(1 - 0.008\Delta T\right)$$

(8.5)

4. The thermal coupling of the position transducer and the torque motor. Symmetry and mounting designs are critical to minimize transducer drift. One manufacturer effectively separates the torque motor from the transducer with the suspension.

Manufacturers give particular attention to the heat dissipation coefficient of scanners for competitive reasons. It is advisable to obtain from the vendors the mode of measurement and judge if it applies to the application at hand.

The dissipation coefficient is frequently specified as "coil to case" because the best numbers are obtained when the change in coil temperature is derived from its change in resistance, as the device is being powered and held in a large fixed-temperature heat sink. Treating the galvanometer stator as an oven where a thermistor is located within a simulated armature yields the most meaningful values. The heat dissipation coefficient obtained in this manner may be half that of the coil to case values.

8.2.1.2 Position Transducer

The simplest and most economical position transducer is a torsion bar. The torque motor pushes against the torsion bar and positions the mirror. This forms a second-order system where bandwidth and position accuracy must be traded off. These units are best built as moving iron devices typically with very stable ALNICO magnets. They typically exhibit good temperature stability, around 150 ppm/°C.

Most high-performance scanners are closed loop servoed systems. They also must deliver bandwidth and positioning accuracy. High-energy magnet torque motors are very powerful and can deliver the bandwidth but the magnetic material, NdFeB, is temperature sensitive and they depend upon the position transducer to deliver the accuracy.

8.2.1.2.1 Gain and Pointing Stability Considerations

Galvanometric scanners applications can be divided into two groups: image/position acquisition and pointing/designation/laser micromachining. In both cases, systems need to be calibrated and drift needs to be compensated. This becomes evident in the tolerance budget of some advanced positioning systems demanding absolute beam positioning.

For example, as the technology advances, the density of the biology carriers in GeneChips is increasing from the present 4000 pixels per scan line to 10,000 and the number of scans per chip from 4000 to 10,000. Each one of the pixels must be correctly located.

High-precision laser micromachining systems used in the manufacturing of flat panel displays or silicon devices such as DRAMs or the trimming of batches of MEMs (such as air bag accelerometers) or trimming of resistors on circuits on silicon need highly accurate and stable gain and pointing performance. These applications commonly demand

addressability errors to be of the order of 1/20,000 or 50 ppm of the field of view and special application can be two or four times more demanding.

The addressability error has a number of sources:

1. It may not be possible to accurately locate the work in the field of view with respect to the scanners' axis of reference. It is common practice to imbed fiducial marks for optical alignment and calibration.
2. The environment and application may not permit a structural design sufficiently rigid.
3. Scanner/transducer drifts may exceed acceptable tolerances for certain applications. In such cases it may be necessary to close the loop around the drift.

The two applications listed above cannot be satisfied with scanners built with only "open loop drift transducers." This becomes apparent upon the analysis of the temperature regulation required to meet the necessary pointing accuracy. The following case illustrates this point.

Let us consider an application where the optical scan angle is 0.4 rad, ptp, and the scanner exhibits a gain drift of 50 ppm/°C as well as a null (or zero) drift of 50 ppm/°C (20 μrad/°C over 0.4 rad excursion) or a total uncertainty for each measurement of 100 ppm/°C. Because two measurements are required to point the beam, one to define the system's references and one to point at the work, the uncertainty is 200 ppm/°C.

The DRAM application would demand that the scanner be held in an environment controlled to less than 0.25 °C and the BioChip example could tolerate twice that. These are extremely demanding conditions, specially for aperiodic operation. For these reasons, closed loop drift compensation or closed loop drift position transducers are required.

8.2.1.2.2 Transducer Drift

Most high-accuracy capacitive position transducers are derivatives of Rohr's[8] design and built with a ceramic vane mounted on the armature of the scanner, which rotates between two stationary conductive plates as depicted in Figure 8.4. A circuit creates a drive signal and detects changes in capacitance. The combined demands of high system bandwidth (0.5–5 kHz) and resolution (1–10 μrad) as well as stability (5–50 ppm/°C) are extremely difficult to achieve in an analog environment and, so far, no manufacturer offers a digital system able to approach this suite of specifications.

These transducers are complex analog devices with multiple sources of drift, both mechanical and electrical, induced by small changes of thermal or other origins including aging and hysteresis. A typical transducer with a 2-cm diameter ceramic butterfly vane can resolve better than 1 μrad in less than 1 ms. Under these conditions the tip of the ceramic vane has moved less than 0.01 μm. It is quite an achievement.

Unfortunately the stability of the same transducer is two orders of magnitude lower per °C. Scanner manufacturers specify short-term temperature gain drift as well as null drift and occasionally uncorrelated drift of the position transducer. Those quantities are difficult to quantify and it is valuable to know under which condition they were derived, as these may not be representative of the conditions where the scanners may be used.

Normally only gain drift is measured and the null or pointing drift value is derived from gain values measured at two extreme representative positions. The null drift quoted may be more representative of asymmetry in the gain drift.
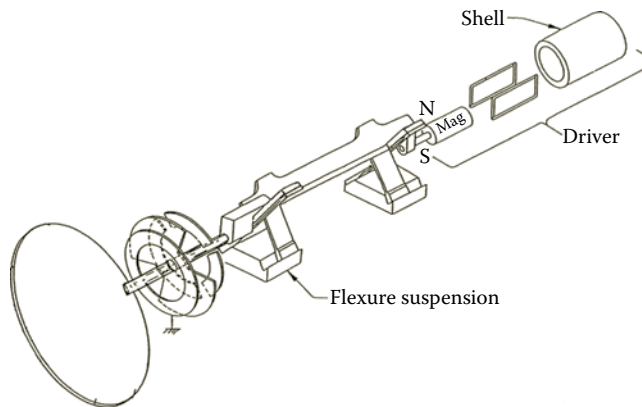
FIGURE 8.4
Scanner with cross-flexure bearings.

Measurements are made with the scanner held and stabilized in a constant temperature oven as the controller is kept at room temperature. This is rarely representative of the operating environment of optical scanners. More common is a burst of energy to rapidly reposition the mirror then followed by rest time when part changes take place.

It is recommended to structure a representative drift test within the environment and process of the scanner application. If this is impractical, drift performances should be derated from the quoted values by at least a factor of 2.

8.2.1.2.3 Closed Loop Drift Transducers

It is frequently necessary that the beam be located with similar or greater precision than the resolution required for the task the instrument is to perform. To define location or correct for drift, reference points, known as fiducial marks, are commonly used. This approach, familiar in astronomy, is also frequently applied to recalibrate gain and null of galvanometric scanners.

It is important to keep in mind that most scanning systems are designed to the limit of the resolution capabilities achievable with the assembled elements: scanners, lasers, mirror size and flatness, focal length, and so on. Commonly, the optical elements performing the task are optimized for dynamic performances as well as optical performances. The addition of other scanner mounted optical sensors or the purpose of absolute position reconnaissance can substantially limit overall performances.

A number of optical techniques have been adopted to recalibrate gain and drift in process. Most of them use split cells—or fiducial marks—located in the work plane, beyond the angular reach of the work area. Weiss and colleagues[9,10] and others report the analyses of these techniques.

A different technique[11] to correct for drift is described by Montagu and colleagues. It takes advantage of the high-resolution capability of the capacitive transducer to incorporate capacitive fiducial features within the transducer proper. The leading edge of a step or a pulse is used for periodic recalibration. In this method, the fiducial features may be positioned on the inside of the operating range of the transducer or beyond. In most applications, this technique offers performances comparable to that obtained with fiducial marks in the work plane and can reduce drift errors by more than an order of magnitude.

8.2.1.2.4 *Optical Transducers*

Optical position transducers offer the attraction of low cost, low inertia, and small volume, as well as low power consumption and an analog signal in an analog environment. Unfortunately the same features are at the root of the difficulties to reach high temperature stability, good signal-to-noise ratio, and good linearity simultaneously.

Optical transducers have recently dramatically improved and the data listed in Table 8.1 lets us compare recent and older designs. Still, the performances of the best optical transducers fall short of those of the best capacitive units. From Tables 8.1 and 8.2 we can see that typical capacitive detectors have stability 50% better than the best optical transducers and their repeatability is about an order of magnitude better. The capacitive transducer also exhibits better linearity: one-quarter the nonlinearity.

Figure 8.5 shows the conceptual construction of three optical transducers from References 12 through 24.

8.2.1.2.5 *Capacitive Transducers*

All capacitive transducers have a common concept where a movable ceramic element rotates between a driving plate and a pair of sensor plates wired in series according to the design of Robert Abbe.[15] The moving element can have two lobes ("butterfly" type) or four lobes ("iron cross" type). Figure 8.4 depicts the basic design mounted at the front end of a flexure-mounted scanner. This symmetrical construction is preferably located between the mirror and the torque motor to simplify the servo loop if the torque motor and the mirror lack a rigid coupling.

The previous sections have addressed the critical issue of stability. Numerous attempts to bring drift within the range of the resolution[16–18] have yielded limited commercial success. The use of fiducial marks in the field of regard has been the only reliable solution for the system designer. Although galvanometers with integrated fiducial marks have been commonly in use for military application, only recently have economical designs become commercially available.

8.2.1.3 **Bearings**

Chapter 5 of this book is devoted to the design of bearing suspension for rotary scanners and the material addressed is applicable here. The design or selection of bearing suspension is critical for both galvanometric and resonant scanners. Galvanometric scanners are built with either ball bearing suspension or flexure bearing suspension. Resonant scanners are built with a variety of suspensions such as cross flexures or torsion bars.

Armatures as well as bearing tolerances of galvanometer scanners are similar to those encountered in the manufacturing of rotating polygons. Oscillating scanners benefit from the periodic motion and a suitable preload should ensure that all elements of the bearing will retrace their path. This should limit wobble to 2 or 3 optical μrad.

Moving magnet torque motors, common to galvanometric scanners, are built with large air gaps—where the drive coil is housed—and consequently impart negligible radial forces. In addition, a properly balanced load should not induce any radial forces. It is therefore practical to preload the ball bearing axially, which is compatible with conventional bearing design. In addition it is possible to use cross flexure bearings, which inherently have low radial rigidity. Torsion bar suspension of resonant scanners may also be employed for the same reasons.
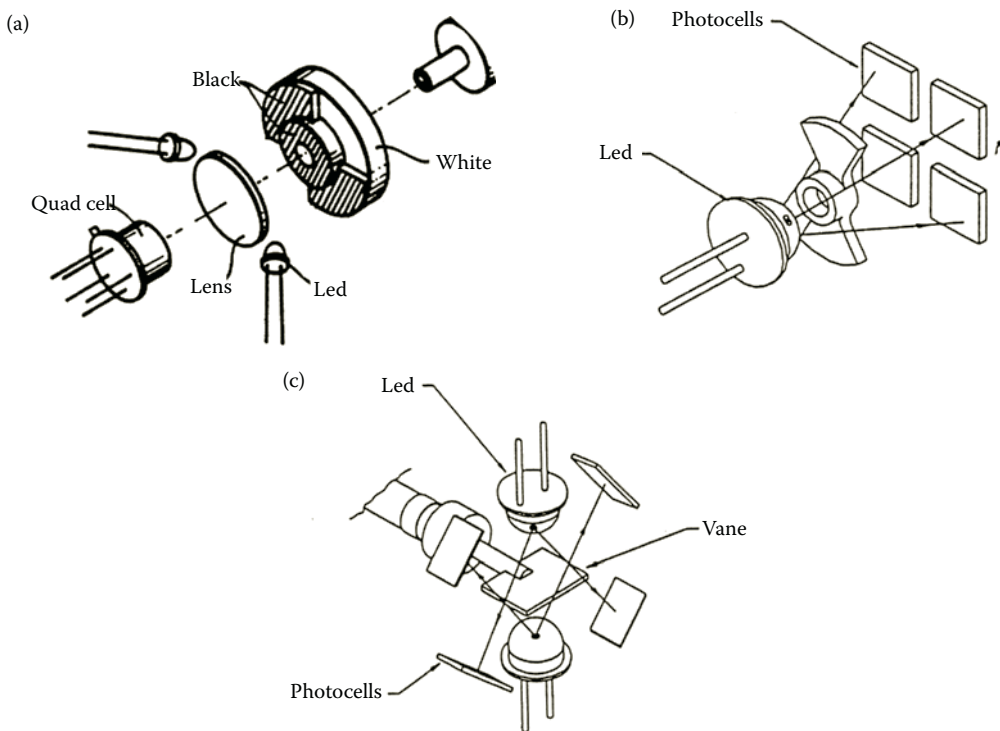
FIGURE 8.5
Construction of three optical scanners: (a) advanced optical position detector (General Scanning U.S. Patent no. 5,235,180); (b) advanced optical position detector (Cambridge Technology U.S. Patent no. 5,844,673); (c) radial optical position detector (Cambridge Technology U.S. Patent no. 5,671,043).

8.2.1.3.1 Ball Bearings

Ball bearings should preferably be selected for their ability to operate at high speed. It is critical to prevent the balls from skidding. To that effect, the choices of lubricant as well as the magnitude of the preload are the major consideration. These are well within conventional bearing standards for scanners operating over a few degrees of motion. Periodic high-frequency continuous small motions, typically under 1 or 2°, are known to cause rapid catastrophic failure, a condition known as "false brinelling" or "fretting corrosion."

Bearing wobble. A typical ball bearing supported spindle, as for a polygon, exhibits 20–50 μrad of wobble normal to scan. This represents spindle-only errors and does not include any sagittal errors of the polygon reflective surface. A typical galvanometric scanner using the same ball bearings on the same spacing will exhibit wobble an order of magnitude lower, and typically under 2 μrad.

The components of a bearing, the inner and outer races as well as the balls, have specified tolerances. The same goes for the bearing seats and the shaft. Errors in excess of 1 μm are associated with each interface and the accumulation of all imperfections defines spindle wobble. Polygon inaccuracies must be added to these.

The armature construction of a well-designed galvanometric scanner has the same tolerances and imperfections as any spindle. The one mirror surface, however, is forced to keep

a constant periodic relationship to all the bearing and other components, so the mirror repeats its sagittal behavior scan after scan and virtually no wobble is present.

The design of galvanometric scanners and mirror systems considers this periodicity as critical to their performances. The sections "Dynamic Imbalance" and "Mechanical Resonances" provide information for analysis of dynamic radial imbalances and confirm the importance of armature mirror balancing.

Bearing preload. The conventional method of achieving radial rigidity of a spindle is to have axial preload for the bearings following bearing manufacturers' recommendations. Moving magnet scanners exhibit extremely low radial forces and are normally preloaded axially. Scanner suspensions have benefited from the improvement of bearings, lubricant, and mounting technology developed for reliable spindles for torque motors that index magnetic heads of CD-ROMs. Ball bearing selection and installation are now well understood and must be tailored to the application and address the following considerations to assure proper life:

1. Preload forces.
2. Radial play and ABEC tolerance number.
3. Lubricant.
4. Acceleration (acceleration greater than 500 g results in sliding ball condition and early failure).
5. Bearing rigidity.
6. Material selection. The choice of ceramic ball is preferred for high speed or high acceleration applications or to minimize fretting corrosion encountered when scanning very small excursion at high frequency.

8.2.1.3.2 *Cross-Flexure Bearings*

Cross-flexure bearings were incorporated in oscillating scanners quite early. Their low radial rigidity to angular rotation ratio has limited their use to angles of only 1 or 2° as they were built with moving iron architecture. The torque motor redesign with high-energy magnets has very much expanded their use. A thorough study of properties and designs of flexure bearings has been conducted by Wittrick[19] and Siddall.[20] It should be noted that radial motion of the axis of rotation is eliminated when the axis is located at one-third of the length of the flexure, as shown in Figure 8.6.

Figure 8.7 is a section of a commercial cross-flexure bearing unit built inside a cylindrical housing. These units come close to the size of a ball bearing. Most scanners are built from photo-etched flexure assembled on a box-like armature designed for minimum inertia and maximum rigidity. Flexures must be stress-free at assembly to avoid catastrophic stress failure.

Figure 8.4 shows a scanner with an armature supported on cross-flexure bearings. Cross-flexure pivot suspensions are totally free of jitter. Flexures are for oscillating scanners what air bearings are for polygons, but at a cost lower by at least one order of magnitude. Both flexures and air bearings have a common weakness however: low radial stiffness.

The advantages of flexure bearings are:

1. Freedom from wobble and jitter
2. Nearly unlimited life in noncorrosive atmospheres
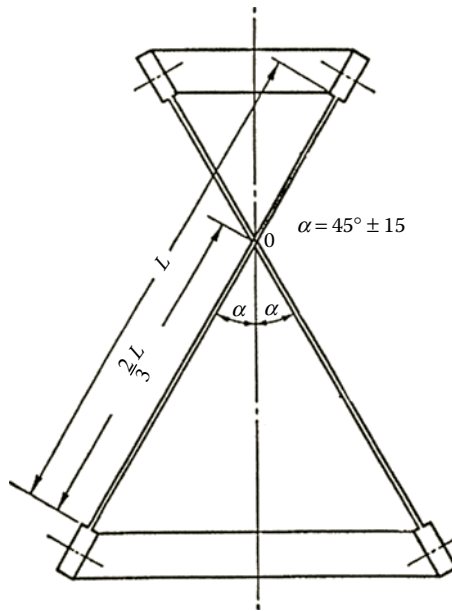
FIGURE 8.6
No center shift cross-flexure bearing.
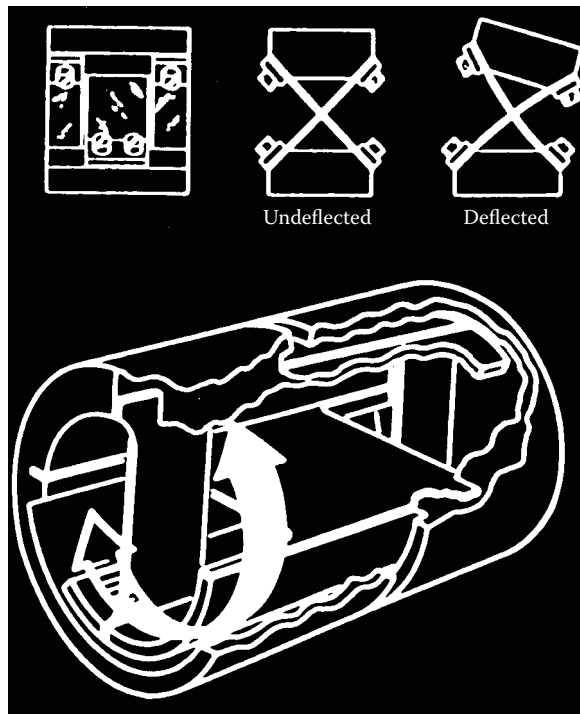
FIGURE 8.7
Free-flex flexural pivot.

3. Operation in a vacuum (non-out-gassing)

4. No contamination of optics (vs. lubricated bearings)

5. Wide temperature range

6. Very low noise

7. Very low damping losses

The disadvantages are:

1. Bulkier than ball bearings

2. More expensive than ball bearings

3. Limited angular excursion

4. Low radial stiffness

5. Rotation axis shift with angle for wide angle units

6. Coupling of torsional and radial stiffness

7. Multiple elastic modes with low damping

8. Difficult installation

9. Intolerance of axial loading

10. Amplitude-dependent rigidity

11. Temperature-dependent rigidity

Fortunately, most of the shortcomings of flexure pivots can be circumvented by trade-offs in scanner design and installation procedures.

Cross-flexure bearings are the preferred suspension for applications with a need for extremely critical repeatability and low wobble. They are also the only solution for small angle excursions, under 4° optical. Small motion does not permit proper ball bearing lubrication, leading to "fretting corrosion" and catastrophic failure in short order. Clearing moves every few seconds in order to assist lubrication may mitigate the damages.

8.2.1.4 Mirrors

Many scanning problems can be attributed to mirror problems. If mirrors could be made infinitely rigid, flat, and reflective, and have negligible inertia, the design and operation of scanners would be extremely simple and there would be less demand for this text. No material able to fulfill these ideal requirements comes to mind; however, available materials do offer practical design solutions.

The facets of a polygon scanner are often viewed as the last link in the chain of components of a scanning system and so burdened with all the system's faults. The mirror of a galvanometric or resonant scanner, however, is rarely perceived in the same fashion and receives comparatively little attention. Actually, they both must meet the same requirements for reflectivity, balance, thermal and dynamic deformation, mounting, and so on.

The condition of the mirror of an oscillating scanner is of concern for that portion of its movement in which optical data is transmitted. Control of dynamic deformations is commonly associated with resonant scanners. Fabrication and installation requirements are commonly associated with galvanometric scanners. Thermal conductivity becomes

an issue with the use of high-power lasers (and galvanometric scanners), which tend to induce temperature gradients.

In oscillating scanning systems, the design and installation of the mirror must be able to preserve the essential system features under all operating and storage conditions. Low wobble, low jitter, precise pointing or scanning accuracy, and long life are typical requirements that have to be addressed.

8.2.1.4.1 Mirror Construction and Mounting

Several guidelines for mirror design and mounting associated with scanning performances have been mentioned previously. These are:

1. The mirror mass must be a minimum.
2. The mirror inertia must be a minimum.

The mirror must be mounted as close as possible to the front bearing of the scanner in order to lower cross-axis resonances. All moments of inertia with respect to the axis of rotation must be balanced in order to minimize wobble induced by angular acceleration and by environmental perturbations. Balancing is most imperative for resonant scanners with torsion bar suspension. Three other performance issues associated with mirror design and installation—alignment, mirror bonding, and mirror clamping/mounting—must be addressed.

Alignment. One consequence of mirror cross-axis misalignment is beam-positioning error, frequently expressed as a "smile" or a "frown." Limited compensation is possible for a line scan, but a precise area scanner, or designator, may require accurate mirror angular positioning. In addition, mounts may need to be electrically as well as thermally isolated. Alignment or balancing must be verified for each mirror along both axes to prevent imbalance and wobble.

An effective way to minimize problems with both alignment and bonding is to precision machine both the mirror's reflective surface and its shaft mounting hole from a single piece of metal to make an integral mirror. Mirror design criteria for high-resonance frequency and mounting stress isolation must be considered.

Mirror bonding. It is extremely difficult to bond a mirror to a mount and have it aligned with an accuracy of 1 mrad. For mirrors smaller than 1 cm the alignment tolerance can be as high as 5 mrad unless optical autocollimation methods and great care in bonding are used.

An improper bond process or mount design can cause mirror deformation when the adhesive cures or temperature changes during shipping or use with higher power lasers. These thermal stresses may also cause the mirror to break or become unbonded.

The elasticity of the bond can cause dynamic pointing errors as well as undesirable resonances that could reduce the system's bandwidth. The rigidity of the bond can cause mirror stresses and deformation when mounting to the shaft of the scanner.

Mirror clamping and mounting. Reiss[21] gives a brief overview of recommended mounting procedures. Clamp-like mirror mounts allow for repositioning and possible removal for replacement. By contrast, both integral and shaft bonded mirror mounting methods are semipermanent conditions.

The most successful removable clamps are a form of collet that provides isolation of the mirror substrate from clamping stress. A disadvantage of the collet approach is potential loosening of the collet clamping forces, with a resulting drift of catastrophic failure.

Collet clamps, if not overtightened, induce only compressive forces that produce no bending movement and thus no distortions in the mirror facesheet. One collet clamp technique is to mechanically isolate the shaft by relieved regions so that the distortions imposed by the clamping screw are not transmitted.

Fastening a mirror mount onto a shaft with setscrews is not recommended, as they can deform the shaft. When setscrews are properly fitted, removing them is nearly impossible. When not fitted properly, they act as a hinge, allowing wobble excitations. Set screws can fatigue and loosen.

Dynamic deformations. The accelerating torques imparted to a scanning mirror can produce significant mirror surface distortions. This is particularly true in scanners driven with a sawtooth waveform and in high-frequency resonant scanners. Brosens'[22] analysis of the deformations induced by accelerating torques yields the approximate formula

$$f = 0.065\left(\frac{s^2 T}{Elh^3}\right) \tag{8.6}$$

where f is the maximum deflection from the original mirror shape; s the width measured across the axis of rotation; E Young's modulus of the mirror material; h the thickness; l the length in the axial direction, and T the applied torque. The torque is related to the angular acceleration a by

$$T = hs^3\ lda/12 \tag{8.7}$$

with f the density of the mirror material. Combining the two equations, we obtain

$$f = 0.0055a\left(\frac{ds^5}{Eh^2}\right) \tag{8.8}$$

This expression points to the desirability of keeping mirrors as narrow as possible. For a glass mirror 1 cm in diameter and 1 mm thick, the resulting deflection at an acceleration of 10^6 rad/s^2 is about 1/25 of the sodium D-line wavelength. Since the inertia of such a mirror is 0.011 g cm^2, the above acceleration corresponds to a torque of only 11,000 dyne cm. The actual mirror deformation may be smaller when a substantial portion of the width of the mirror is cemented to a mount.

Thermal deformations. When a scanning mirror is exposed to radiation, a substantial part of the radiation that the reflecting surface absorbs is transferred in the form of heat to the rear surface. This heat is discharged by conduction to the mount and by radiation and convection to the surrounding atmosphere. The conduction of heat to the rear surface causes differential expansion and, if the incident radiation is particularly intense, significant distortions can occur. Such distortions can be estimated by assuming one-dimensional heat transfer to the rear surface.

The radius of curvature R caused by a uniform temperature gradient is

$$R = \left\{a\left(\frac{du}{dx}\right)\right\}^{-1/2} \tag{8.9}$$

where a is the coefficient of linear thermal expansion, and du/dx is the temperature gradient in the material. From Fourier's law of conduction,

$$\frac{du}{dx} = \frac{q}{kA} \tag{8.10}$$

where q is the heat transfer rate, k is the thermal conductivity, and A is the cross-sectional area. The camber assumed by a plate of width s, when it curves with a radius of curvature R, is given to a first-order approximation by

$$e = \frac{s^2}{2R}$$

(8.11)

Combining this equation with the previous expressions, we obtain

$$e = \frac{aqs^2}{2kA}$$

(8.12)

For a glass mirror of width 1 cm conducting 0.1 W/cm² to the back surface, the resulting camber is 0.5 μm, or about 1 wavelength.

Erosion. When a scanner mirror is moved through air at high speed, the collision of dust particles with its surface can cause gradual erosion of its reflective coating. Experience shows that for any coating the process of erosion does not occur below a critical impact velocity. It is believed that surface erosion occurs when the stress developed at the impact interface between the coating and the dust particle exceeds a value that is characteristic of the coating.

The stress developed by the impact of a rigid body against an elastic mass was analyzed by Timoshenko and Goodier.[23] The stress wave generated by impact is given by the formula

$$S = E\left(\frac{V}{c}\right)$$

(8.13)

where E is Young's modulus of the substrate, V is the relative speed at impact, and c is the velocity of wave propagation (sound velocity) in the substrate.

Experimental evidence shows that AISIO coatings on fused silica degrade through erosion at all points where the speed of motion exceeds 3 m/s at any instant during the scan cycle.

Users of high-speed scanners should take precautions to minimize the presence of suspended dust particles near scanning mirrors. Where such protection cannot be provided, hard coatings should be used in combination with substrates of low Young's modulus.

Material selection. All scanner applications do not have the same performance requirements, so there is no optimum mirror material. The selection of a substrate is application dependent and has to satisfy some or all the performance requirements reviewed earlier.

Table 8.4 lists the properties of materials suitable for mirror substrate as well as mounts. The figure of merit for resonant scanners, E/d^3, has been derived by Brosens and Vudler.[22] This is to be used as a comparative guide for a given geometry. One should keep in mind that design, construction, heat treatment, coating, and installation can each have a dominant influence on the performance of a mirror. The figure of merit for galvanometric scanner mirror design is E/d. This is based on fabrication requirements only (for discussion, see Reference 25).

Cost, ease of fabrication, stability with time and environmental conditions (such as temperature and cyclic stresses), bonding capability, and mirror surface finishing are

TABLE 8.4

Mechanical and Thermal Properties of Substrates

Material	Density (g/cc)	Coef. T-exp. (10^{-6}/°C)	Therm cond. (W/cm °C)	E, Young's modulus (kg/ $cm^2 \times 10^5$)	Fig/merit ($E/d^3 \times 10^5$)	Fig/merit ($E/d \times 10^5$)
BK7	2.53	8.9	0.010	8.22	0.50	3.2
Fused silica	2.20	0.51	0.014	7.10	0.66	3.2
Fused quartz	2.20	0.51	0.014	7.10	0.66	3.2
Pyrex	2.23	3.3	0.011	6.67	0.54	3.0
Silicon	2.32	3.0	0.835	11.2	0.89	5.0
Aluminum	2.7	25	2.37	7.03	0.35	2.6
Iron alloys	7.86	0–20	0.1–0.8	13–21	0.03–0.04	<2.5
Al oxide	3.88	7.0	0.08	36.0	0.61	9.3
Titanium	4.3	8.5	0.20	11.2	0.14	2.6
Beryllium	1.8	12.0	2.10	30.8	5.2	17.0
Magnesium	1.7	26.0	1.59	4.2	0.80	2.5
Diamond	3.5	0.7	10–25	120.0	2.6	34.0
Silicon carbide	2.92	2.6	1.56	31.5	1.4	11.1
SXA	2.96	10.8	1.2	14.5	0.56	4.9
Tungsten carbide	15.3	5.94	0.5	68.5	0.02	4.5
Miralloy	2.10	6.3	1.1	20	2.1	9.5

extremely important to the selection of a substrate material. Fatigue and yield strength are normally irrelevant.

Mirror surface finish. Definitions and specifications of available surface finishes and coatings for glass mirrors can be obtained from numerous sources, including government specifications, and will not be reviewed here.

For metal mirrors, difficulties begin with the form of the metal stock and the machining process. Each case is different and presents its own problems. After a blank has been machined to finished dimensions, it has to be stress relieved and stabilized. Dynamic stressing can also be used. Thermal stabilization can be achieved with three or four cycles of processing from liquid nitrogen to boiling water.

The following processes are used for surface finishing when polishing the substrate is not acceptable.

Plating and polishing. All the surfaces of the mirror can be plated with equal thickness of hard nickel (typically 0.002–0.005 in) to avoid thermal deformation. Any material removal during polishing may need to be allowed for when plating, and additional thermal stabilization may be required. Nickel can be ground and polished to a high-quality surface and then finished by conventional means.

Replication. This is a process in which a reflective surface and one or more coatings are formed by successive evaporation in reverse order onto a master, and then transferred together onto the substrate and bonded. The bonding agent is typically an epoxy layer with a viscosity of 100 centipoise and a thickness of a few micrometers. If the thickness approaches 25 μm, the process introduces alignment errors of 0.1–1 mrad or greater. There is limited usable temperature range due to "bimetallic deformation." Weissman[24] shows that a 1 fringe curl per 25 °C is to be expected for a disc with a 25:1 aspect ratio.

Additional limitations may be introduced by the power to be reflected from the mirror. Brosens and Vudler[22] calculate that the heat transfer coefficient of epoxy could limit

replicated optics to energy pulsed under 0.0017 J/cm² or four orders of magnitude lower than polished copper optics at YAG and CO_2 wavelengths.

Diamond machining. This technology has proven to be exceptionally successful in the manufacture of high-volume low-cost aluminum polygon mirrors with sagittal tolerance <50 µrad. Diamond machining is less attractive for beryllium and coated steel substrates, or when reflectivity requirements necessitate additional processing.

8.2.1.4.2 Mechanical Protection of Mirror Substrates

Table 8.4 lists the material and thermal properties of various mirror substrate materials. It should be noted (last column) that beryllium has the highest figure of merit of all materials but for diamond.

8.2.1.5 Image Distortions

This section reviews the most common scanner-related image distortions. While it is important to identify the source of any error, only the most common imperfections deriving from scanner sources will be discussed here.

8.2.1.5.1 Cosine Fourth Law

Off-axis image distortions have been described by Smith.[26] They occur even when there is no vignetting. The illumination is usually lower than for the point on the axis. Figure 8.8 is a schematic drawing showing the relationship between exit pupil and image plane for point A on axis and point H off axis. The illumination at an image point is proportional to the solid angle that the exit pupil subtends from the point. It is apparent that for small values of f, $f' = f\cos^2 f$, and that OA = OH cos θ. Thus, the solid angle subtended by the pupil from H is reduced by a factor of $\cos^3 \theta$ from that subtended at A. Now the illumination so far has been considered in a plane normal to the direction of propagation; it is apparent that at H the energy is spread over an area that is proportionately larger than at A because the cone strikes the surface at an angle θ from the normal; thus, a fourth cos θ factor must be added, and we find that:

$$(\text{illumination at H}) = \cos^4 q (\text{illumination at A})$$

(8.14)
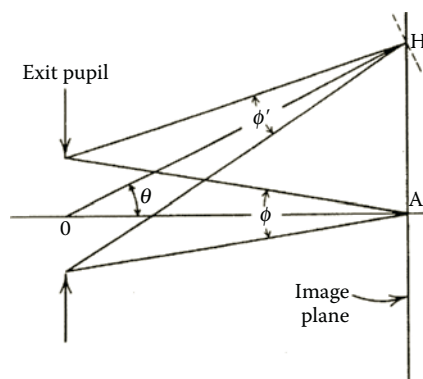
FIGURE 8.8
Cosine fourth image distortion.

8.2.1.5.2 Index of Refraction of Air

The index of refraction *n* for air is strongly dependent on the local pressure *p,* density *d,* and absolute temperature *T.* The pertinent relations are

$$\frac{p}{d} = RT \quad \text{and} \quad \frac{(n-1)}{d} = K$$

(8.15)

where *R* is the gas constant and *K* is the Gladstone–Dale constant, an empirical value. One should keep in mind that the local air density is proportional to its local velocity. In practice this forces the designer to enclose the optical bench and prevent air motion.

Excessive wobble of single-axis or bending of two-axis scanning systems operating at a slow speed (a few scans per second) is frequently due to air motion and can be eliminated by proper baffling and occasional vigorous air stirring, such as with a fan.

Microphonic perturbations are due to the variability of the index of refraction of air.

8.2.1.5.3 Air Dynamics

As we have seen, air is not the ideal medium for light to travel through. Air also adds to scanning difficulties due to the damping effect of its viscosity and the buffeting perturbations caused by its turbulence. These disturbances are pertinent for systems of high speed and high precision. At present they are frequently encountered with high-performance resonant scanners that are selected for their high frequency and large mirror capabilities. Many advanced systems being contemplated operate the scanners in partial vacuum or helium in order to minimize these disturbances.

There is no literature for low inertia scanners equivalent to that of Lawler and Shepherd[27] for polygons. Aerodynamic effects for low inertia scanners are complex due to the extremely low inertia and stored energy of the moving element compared to the effects of the aerodynamic forces generated. The Reynolds number has been used by Brosens for the purpose of evaluation.[28]

The Reynolds number Re is a dimensionless quantity function of the fluid density *d,* viscosity *v,* velocity *V,* and the mirror radius *r.*

$$\text{Re} = \frac{dVr}{n}$$

(8.16)

A practical expression for air at standard atmospheric conditions in the MKS system is

$$\text{Re} = (Vr)6.7 \times 10^{-4}$$

(8.17)

At Reynolds numbers above 2000 the pressure forces proportional to mirror tip velocity add to the viscous losses and are the dominant cause of low *Q* for resonant scanners. This is the region where laminar flow changes to turbulent.

It is also these turbulences that induce jitter in resonant scanners, which can exceed 5 µrad. This, more than any other effect, limits the dimensions and operating frequency of resonant scanners in air.

8.2.1.5.4 Mirror Surface Off Axis

For dynamic reasons, the reflecting surface of a scanner mirror is normally offset from the axis of rotation. This offset *T* causes an additional scan nonlinearity error. In order to

minimize this effect, the beam should be centered below the axis of rotation by an amount K as shown in Figure 8.8. This error E function is

$$E = \frac{(T - K \sin a)}{\cos a}$$

where α is the angle of the mirror to the normal of the work plane.

Figure 8.9a and b show a graph of typical single-axis, flat-field scan error for a mirror with a unity surface offset. The scan angles are beam rotations. The beam scan angle is the rotation added to the reference angles of 37° and 45°, respectively.

8.2.1.5.5 Beam Path Distortions

Beam path distortion (BPD) may come from the scan head or imaging system. It is caused by path length variations for different portions of the beam. Portions of the image may be
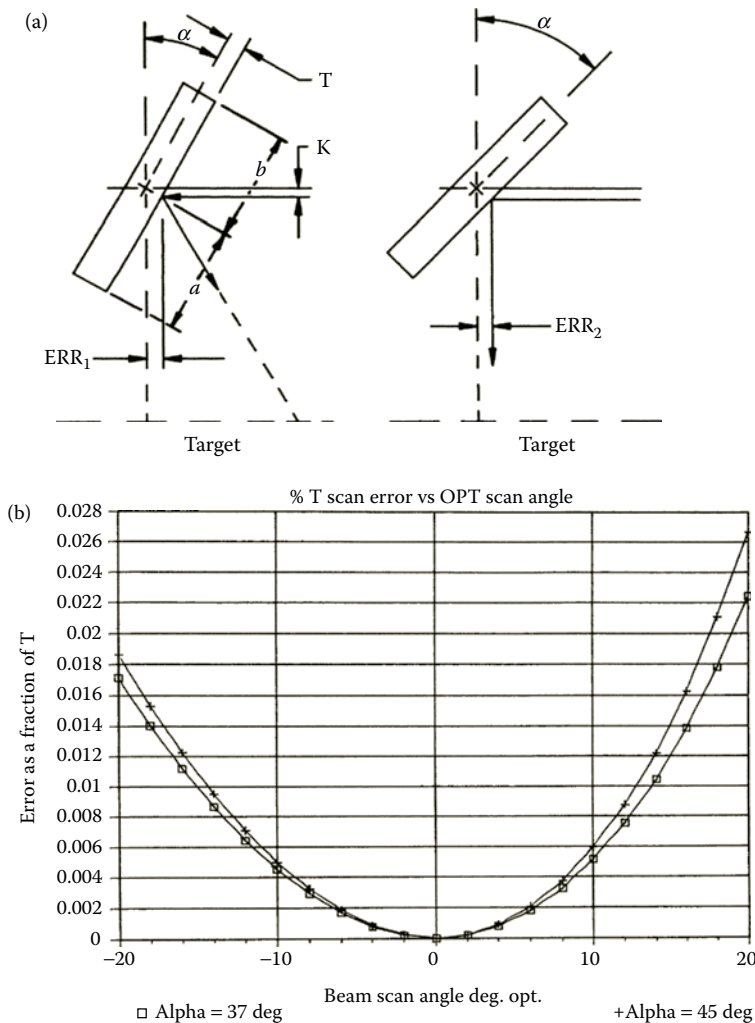
FIGURE 8.9
Single-axis, flat-field scan error: (a) compensation for mirror surface off-axis error; (b) scan error versus scan angle.

blurred or may focus before or after the image plane. Alternately, portions of the image beam may be directed to an incorrect position in the image plane.

In focused laser systems, BPD frequently appears as elongated or distorted spots. These defects appear in different axes in the near and far focus about the optimum focus or as "lobes" of energy projecting from the focused spot.

In vision systems, the image is not always diffraction limited, particularly with the larger apertures. As much as or even 1 wave of distortion may be acceptable.

Mirror nonflatness in imaging systems is a common cause of BPD and is likely to impair the image astigmatically.

Vertical and horizontal axes do not focus at the same plane. This is seen in the two views of the image beam in Figure 8.10. The incorrect imaging position of certain bundles of light typically reduces contrast in the image.

Figure 8.10 shows the front and side views of an image beam reflected from a cylindrically deformed mirror. In the front view, because the beam's convergence is reduced, it focuses at a farther point than the undisturbed focus cone in the side view. The spot diagrams show that, in this case, the spot takes on an oval or enlarged form; it never attains the correct size and shape, shown on the right.

Lens system defects, such as decentering, stress, and other manufacturing-related problems, can also cause image deformities. If image quality does not meet expectations, test the lens system without the scanner. Figure 8.10 shows common distortions.

F–θ lenses commonly exhibit more than 0.5% nonlinearity. Field-flattening lenses also have performance tolerances that need to be specified.

8.2.1.6 Dynamic Performances

Galvanometric scanners are servo-controlled devices and must obey all standard closed loop servo control requirements for stability. Servo system theory is a well-developed discipline and this section shall only review the scanner design features affecting scanner performances and the effect of some common drive profile signals.

As the cost of high-speed computers has come down, they have become an integral element of scanner systems. Programs can be implemented that alter, on the fly, the drive signal to the power amplifier to optimize random addressing of scanners or laser beam in two or three dimensions. This may be used to circumvent amplifier saturation and high frequency excitations. System performances are strongly affected by the armature design as well as the driver amplifier and the algorithm chosen to drive the scanner.
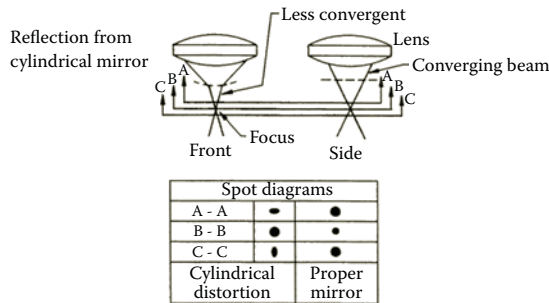
FIGURE 8.10
Astigmatism caused by cylindrical mirror surface.

8.2.1.6.1 Resonances

All mechanical elements have resonances. Galvanometric scanners should preferably be designed such that all elements and subelements be as rigid as possible. Ideally their lowest natural frequency should be higher than the highest frequency of the drive signal or any perturbing frequency that can be transmitted to it. This is rarely possible.

It is customary to identify offending resonances and exclude them from the drive signal. It should be kept in mind that scanners mounted on nonrigid lossy material, for isolation, may lose their registration to the work surface. Mirror installation is a frequent source of imbalance and should be executed with that in mind.

8.2.1.6.2 Dynamic Imbalance

A rotating body can be balanced, but never perfectly. It is necessary to qualify and quantify resulting imbalance forces in order to assess their consequences and judge if they are acceptable for the application.

Bearings are built with some degree of radial play. When they are subjected to periodic eccentric forces that exceed the constraints of their preload, radial or axial, damage follows that can result in catastrophic failure.

Figure 8.11 is a schematic representation of a ball bearing mounted rotor from a galvanometer with an unbalanced load m. The total system inertia is J and a drive torque T imparts an acceleration $d^2\theta/dt^2$. The middle of the rotor is also subjected to a radial force F_5, which is the bearing preload. The derivation uses the symbols of Figure 8.11.

All the torques and forces on the rotor must balance; consequently, the following conditions must be satisfied. This will cause perturbations in the scan axis. Periodic compression and decompression of data will occur. If such a scanner is used to generate gray tone images, they will show lighter and darker waves normal to the raster lines.

$$T = \frac{mr^2 d^2 q}{dt^2} \tag{8.18}$$

The derivation equations shown in the figure:

$$T = J \frac{d^2\theta}{dt^2}$$

$$F_1 = F_2 = mr \frac{d\theta^2}{dr^2} = \frac{mrT}{J}$$

$$F_3 = F_1\, a/b$$

$$F_4 = F_2 + F_3 = F_1 \frac{a+b}{b} = \frac{mrT}{J} \cdot \frac{a+b}{b}$$

Stability criterion:

$$F_5 > 2F_1 \frac{a+b}{b}$$
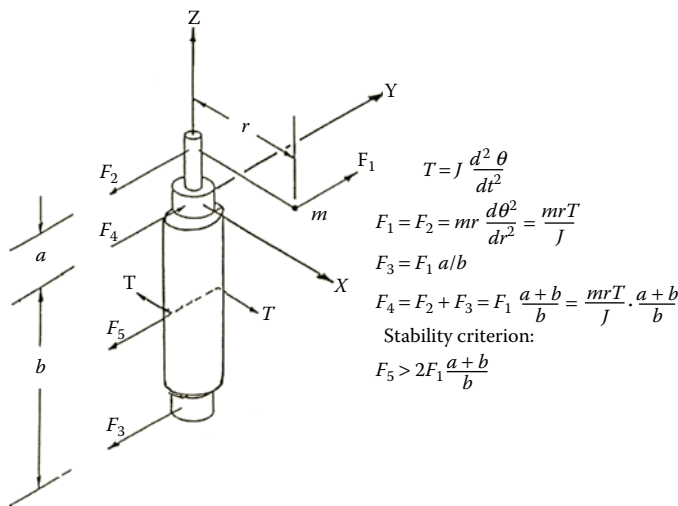
FIGURE 8.11
Dynamic forces on rotor.

$$F_1 = F_2 = \frac{mrd^2q}{dt^2} \tag{8.19}$$

$$F_3 = \frac{F_1 a}{b} \tag{8.20}$$

$$F_4 = F_1 + F_3 = \frac{F_1(a+b)}{b} \tag{8.21}$$

The stability criterion is for the preload to be greater than the period eccentric forces. If it is assumed, according to Figure 8.11, that the eccentric mass represents the effect of the mirror, the front bearing is the most vulnerable. The following relationship must be satisfied:

$$F_4 = \left(\frac{(a+b)}{b*mr}\right)\left(\frac{d^2q}{dt^2}\right) < F_5/2$$

$$F_4 = \left(\frac{mrd^2q}{d^2}\right)\left(\frac{a+b}{b}\right) < F_5/2 \tag{8.22}$$

The most common mirror mounting technique is a mass balanced assembly with the reflecting surface forward of the axis of rotation. Lateral mass balance is equally imperative.

Armature imbalance of resonant scanners causes wobble and excites the instruments' chassis. Unacceptable audio coupling to the chassis can be minimized with massive construction or soft mounting of the scanner. Both are costly and undesirable as compared to a properly balanced armature.

8.2.1.6.4 Mechanical Resonances

A perfectly balanced armature mirror assembly can still produce unacceptable oscillations. These are caused by the excitation of any possible natural frequency of one or more of the elements of the armature or occasionally the stator. These structures commonly have no damping and can be excited by a magnetic imbalance or external shocks and vibrations. Such oscillations are usually sensed by the control circuitry and amplified to cause system instabilities; it is necessary to design the armature so that its first resonance in any mode is substantially beyond the cutoff frequency of the amplifier or is properly damped.

This is the most common limiting factor to the speed of response of a small mirror galvanometric scanner. Two familiar modes of oscillation are reviewed here.

Mirror on a limb. The mirror is overhung at the end of the shaft and behaves like a freely supported beam. With reference to Figure 8.12 it has a deflection angle θ expressed by

$$q = \frac{Mb}{3EI} \tag{8.23}$$
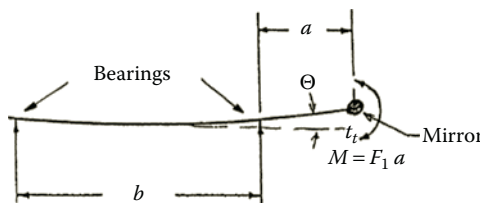
FIGURE 8.12
Bending forces on rotor.

where E is Young's modulus of the shaft's material and I its moment of inertia. For a heavy mirror of mass m the first cross-axis resonant frequency ω is expressed by:

$$w = \left(\frac{3EI}{ma^3}\right)^{1/2} \tag{8.24}$$

For a small mirror, the first cross-axis resonant frequency is the rotor resonance. This has been analyzed by Den Hartog.[29] If the rotor can be represented by an iron cylinder, its resonance can be expressed by

$$w = \frac{500,000\, d}{b^2} \tag{8.25}$$

where ω is in radians per second, d is an approximate value of the diameter of the rotor, and b is its length between bearings, as shown in Figure 8.12 (*d* and *b* are both in inches here). Exciting this resonance will cause the mirror to wobble and/or render the servo unstable.

Torsional resonances. The rotor and mirror are two freely supported inertias connected by a shaft. The resonant frequency of such a system, with reference to Figure 8.13, is

$$w = \left\{\frac{K(J_1 + J_2)}{J_1 J_2}\right\}^{1/2} \tag{8.26}$$

This will cause perturbations in the scan axis. Periodic compression and decompression of data will occur. If such a scanner is used to generate gray tone images, they will show lighter and darker waves normal to the raster lines.

8.2.1.6.5 Armature Construction

The armature of galvanometric scanners is a mass-spring system. The scanner's dynamic performance—step response—is limited by the first uncontrollable resonance. Most attempts for compensation beyond the first uncontrolled mechanical resonance have been so far mostly ineffective. The armature is made as rigid as possible and consequently exhibits very high Q resonances. These resonances are known to shift slightly with changes in temperature of the device or as a function of its mode of operation. This renders analog compensation extremely difficult and digital compensation is limited to slow operating systems. It is imperative to have very rigid constructions to minimize the number of possible resonances and raise all resonances preferably one order of magnitude beyond the desired system bandwidth.

Two armature architectures are commonly encountered and schematically exemplified in Figure 8.14a and 8.14b where the bearings are omitted. Both designs can have the same components but produce very different servo system responses.

FIGURE 8.13
Angular rotor mirror presentation.
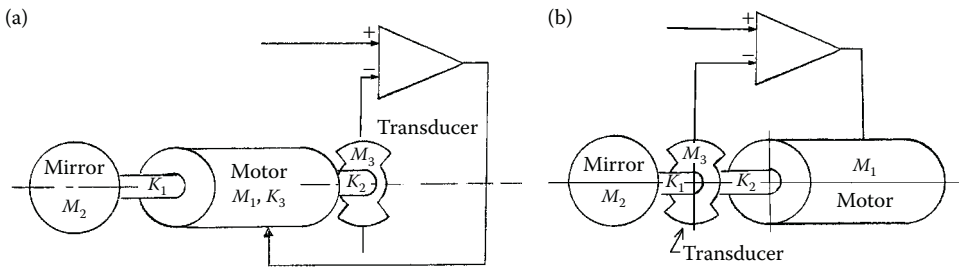
FIGURE 8.14
Armature architectures: (a) galvanometric scanner construction with mirror load and transducer at either end of the armature; (b) galvanometric scanner construction with mirror load and transducer at the same end of the armature.

The construction of Figure 8.14a is more common as it is simpler to build. The mirror and the transducer are located at either end of the torque motor, beyond each bearing. Unfortunately, in order to satisfy the need for a high torque to inertia, the magnet is long and thin and adds one spring to the servo loop. This may also add a low cross resonance that may couple into the transducer signal. This may be inconsequential for a raster scanning application operating at a single frequency where these resonances may be excluded from the drive signal.

The construction of Figure 8.14b offers a simpler servo system and consequently potentially a better response, but it is more complex to assemble because one shaft carries both the transducer armature and the mirror. This construction offers an efficient mounting where the scanner is held near the mirror and also serves as an efficient heat sink. It may be noticed in Tables 8.1 and 8.2 that scanners that have adopted this architecture also claim better thermal stability.

8.2.1.6.6 Drive Signals

It is imperative that the frequency content of the drive signal and the magnetic forces is kept away from exciting secondary resonances of the armature. For example, when designing a vector scanning micromachining system with a 0.2 ms small step response, one would desirably see that all secondary mechanical resonances be beyond 50 kHz.

Using Equation 8.25 it can be shown that the resonance of a magnet 0.5 in in diameter and 2 in long, held on bearings at both ends, has a cross-axis resonance of about 10 kHz, which should be excluded from any drive signal.

Vector scanning. The shortest step response for a given excursion can be derived from the equation of motion of a second-order system under the assumption that current, voltage, resonances, and electrical time constants are negligible. This is an idealized model and should be used with that understanding. Experience tells us that small steps, 1°, can be expected to approximate these theoretical values. In most cases full torque is limited by the power supply and the driver and is delayed by the electrical time constant of the circuits and the torque motor. These limiting conditions are most experienced when large angular jumps are made.

Small angle stepping time can approximate the condition derived from Newton's law and depicted in Figure 8.15a. The minimum stepping time is derived as follows:

$$T = \frac{I d^2 q}{dt^2} \tag{8.27}$$
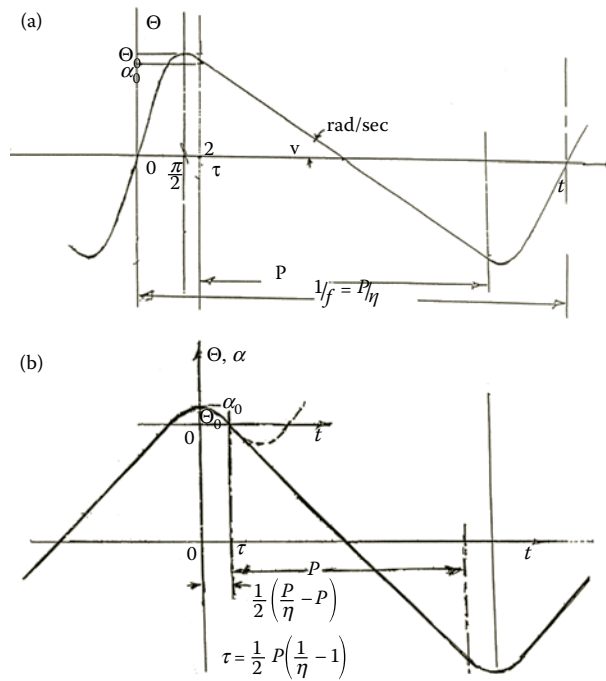
FIGURE 8.15
Derivation of (a) saw tooth motion; (b) triangular motion.

where T is the torque required to impart the angular acceleration d^2q/dt^2 and I is the moment of inertia. To optimize a reciprocating motion system, equal energy and time must be allowed for both acceleration and deceleration of the mirror and armature. The maximum potential mechanical energy W that can be given to a rotating system is expressed as

$$W = \frac{1}{2}(Tb)$$ (8.28)

where β is the total angular displacement (peak-to-peak). This energy can be expressed as equivalent dynamic energy by the kinetic energy equation as follows:

$$W = \frac{1}{2}I\left(\frac{dq}{dt}\right)^{1/2}$$ (8.29)

Solving the above three equations for time yields the expression for the minimum stepping time or step response:

$$t = 2\left(\frac{Ib}{T}\right)^{1/2}$$ (8.30)

In actual cases it is imperative that the drive signal excludes known resonances of the armature. In practice, structured waveforms are created that approximate a section of a sine wave joining the two endpoints of the scan angle. This is known as "cycloidal waveform" and minimizes high-frequency components.

The acceleration capability of moving magnet torque motors is extremely high and in some condition may cause the ball in the bearings to slip rather than rotate. This may lead to rapid catastrophic failure. For this reason, it is advisable to consider a drive signal other than the maximum acceleration condition described above. As expressed above, this condition can be very detrimental with a scanner operating at small amplitude, under 1 or 2°, which does not allow proper ball lubrication and leads to a condition known as false brinelling. A flexure bearing suspension is recommended for these applications.

Raster scanning. The critical points to be attended to for raster scanning are scanner overheating and drive signal-induced vibrations as well as bearing damage as described in the above paragraph. A judicious drive signal design can avoid most of these difficulties.

Two types of raster modes are commonly used: saw tooth and triangular tooth signals.

Saw tooth drive signal. Saw tooth drive signal yields a simpler overall architecture but makes much more demands on the scanner in the turn-around critical points. The fastest possible fly back time can be derived from Equation 8.30 and is often referred to the "constant acceleration" drive signal. Frequently, amplifier and power supply saturation as well as inductance and other phase delays will cause the turn-around time to be longer than calculated. Also, caution must be taken to avoid exciting on-axis or cross-axis resonances that may couple through the transducer and the drive amplifier with positive feedback. Cross-axis vibration can couple into the elements of the transducer to induce an erroneous output. If these resonances are known they should be excluded from the frequency spectrum of the drive signal. The most desirable fly back signal, and frequently the fastest, is shaped as a segment of a sine wave where both the start and ending of the sine wave segment match the slope of the linear part of the signal as described in Figure 8.15a.

The relevant parameters of a saw tooth drive is illustrated in Figure 8.15a. The linear segment of the signal has duration p and slope V. The efficiency η can be expressed as a function of the frequency f of the signal in Hz according to

$$h = pf = \frac{p}{p} + 2t \tag{8.31}$$

The fly back signal is a sine wave in order to minimize the system's bandwidth. The sine wave and the linear segment join at an equal slope at time T somewhat longer than one-fourth period of the sinusoidal signal at time $\pi/2$. The angular position θ of the shaft is therefore expressed as

$$q = q_0 \sin wt \tag{8.32}$$

At time T the angular position is α_0 such that

$$a_0 = q_0 \sin wt \tag{8.33}$$

We match the slope of the sinusoidal motion to the slope of the linear motion such that

$$V = \frac{-2a_0}{p} = q_0 w \cos wt \tag{8.34}$$

The ratio of these equations yields

$$\tan wt = \frac{-qpw}{2a_0} \tag{8.35}$$

At time $t = t, q = a_0$ and the equation simplifies to

$$\tan wt = \frac{-pw}{2} \tag{8.36}$$

As both T and p are defined by the application, ω can be derived by iteration. As a starting value for ω the value derived by setting $q_0 = a_0$ yields $w = p/2t$.

When ω is known, the first equation can be used to derive the value of q_0 as both a_0 and τ are defined parameters.

The pick acceleration a can than be calculated as

$$a = q_0 w^2 \tag{8.37}$$

It is interesting to note that the natural frequency of the fly back is a function of only the scan efficiency $h = p/p + 2t$.

Triangular tooth scanning. Triangular tooth scanning normally yields a system running at more than twice the repetition rate with the same power consumption. The coding/decoding software must take into consideration phase shifts and possible lack of symmetry or linearity of the position transducer. Again the turn-around signal should be structured to minimize power consumption, to limit heating and undesirable frequencies. Again, the preferred drive signal is a segment of sine wave that matches the slope of the linear portion of the signal as described in Figure 8.15b, yielding a more robust system than the "constant acceleration" signal. The analysis of this model given in Equations 8.38–8.41 can readily be adapted for saw tooth drive.

If p is the linear segment of the signal with slope V and the efficiency is the magnitude of the overshoot, a can be derived for a sinusoidal turn-around waveform signal as outlined below according to the symbols of Figure 8.15b.

$$V = \frac{2q_0}{p} \tag{8.38}$$

$$a = a_0 \cos wt \tag{8.39}$$

where

$$w = \frac{2ph}{2p(h-1)} \tag{8.40}$$

and therefore

$$a = \frac{a_0 \cos 2pht}{2p(h-1)} \tag{8.41}$$

We match the slope of the sinusoidal motion to the slope of the linear motion such that

$$\frac{da}{dt} = \frac{2q_0}{p} = a_0 \left\{ \frac{ph}{p(1-h)} \right\} \frac{\sin pht}{p(h-1)} \tag{8.42}$$

At time

$$t = \frac{p(1-h)}{2h}$$

(8.43)

the angle of the sinusoid is $p/2$ and as $\sin p/2 = 1$ and the magnitude of the overshoot is derived from

$$a_0 = \frac{2q_0(1-h)}{ph}$$

(8.44)

The second derivative of *a* is the angular acceleration *a* that shall be imparted to the armature during the turn-around function and it is maximum at $t = 0$. With Equation 8.44 it yields

$$a = \frac{2q_0 ph}{p^2(1-h)}$$

(8.45)

It can be noticed that the overshoot and the galvanometer acceleration, and therefore the torque requirement, are inversely affected by the scanning efficiency.

8.2.1.7 Evaluation Parameters

The accompanying tables list only scanners built as inside–out d'Arsonval movements. These are moving magnet designs with a large air gap in the magnetic circuit. This architecture is preferred for optical scanners as it best meets the list of desirable features listed in Section 8.2.1.1. Earlier optical scanners were built with torque motors incorporating iron in the stator or with moving coil armatures. These devices are reported in Reference 1. To the knowledge of the author no advanced scanner has come from any further development of these two technologies.

As all these units are of similar torque motor design, the figure of merit of the torque motor reflects the coil copper packing density as well as its thermal conductivity. This feature may prompt the unit choice when the scanner is worked hard, such as in raster scanning, or if thermal drift is a critical element.

The choice of position transducers, capacitive and optical, is normally guided by the tolerable error in repeatability dictated by the application as well as by the duty cycle of the application. A low duty cycle with high peak power demand shall not permit the transducer to reach a stable temperature and therefore the actual drift parameters may need to be verified experimentally in the application or the use of fiducial marks should be considered.

The dynamic performances listed are to be used as references only as different standards are used for the various parameters. Dynamic performances also depend on the armature construction, the load, the load attachment, as well as the sophistication of the driver amplifier and often also the drive signal that is used.

The accompanying tables list commercial scanners according to their armature inertia. All the data is derived from published material from the following manufacturers:

1. Cambridge Technology, Cambridge, MA: models 62xx, 64xx, and 68xx
2. GSI Lumonics (General Scanning Inc), Bedford, MA: models Mx and VM2000
3. Nutfield Technology Inc., Windham, NH: models RZ-xx
4. GalvoScan LLC, South Royalton, VT: models TGV-x

Other commercial manufacturers of galvanometric scanners, whose products are not tabulated here are:

1. EOS, Munich, Germany: moving iron scanners
2. Lasesys Corp., Santa Rosa, CA: stepper motor with an optical encoder
3. Laserwork, Orange, CA: entertainment products

8.2.2 Resonant Scanners

The armatures of resonant scanners are low-mass, high-rigidity, and high-Q structures. Large excursion can be achieved with a low-torque simple drive motor. The major advantage of the resonant scanner is its simplicity, small size, long life, and in particular its low cost. Its major disadvantage has been its sinusoidal motion, and sensitivity to external as well as self-induced perturbations. This deficiency corroborates the scarcity of low-frequency resonant scanners. The induced moving coil of Figure 8.16 is an older but simple low-frequency design that is commercially available.

8.2.2.1 New Designs

Taking advantage of new materials, Dean Paulsen[30–32] developed new concepts for the design of resonant scanners:

1. The use of a high-energy permanent magnet, minimizing radial forces
2. The location of anchor points at vibration nodes of the armature
3. The use of highly lossy material, such as "Sorbutane" at the anchor points in order to damp out external as well as self-induced perturbations
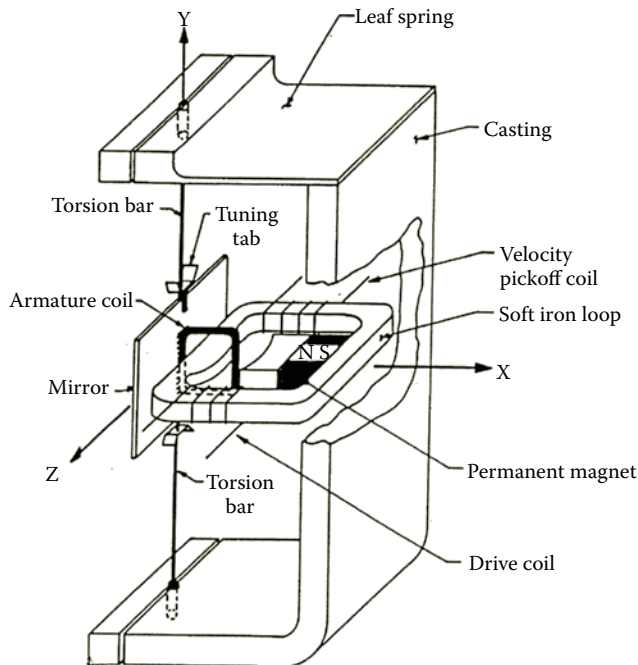
FIGURE 8.16
Induced moving coil resonant scanner.

8.2.2.2 Suspension

Figures 8.17 and 8.18 exemplify these concepts as counterrotating resonant systems. The anchor point or points are located at nodes of the resonant armature that do not experience any movement. This is analogous to holding a musical tuning fork at its base. No energy is lost. High-frequency devices can be built with one support. Low-frequency devices need two supports to be insensitive to orientation or external perturbation. One can also note in Figure 8.17 two orthogonal coils. One is the drive coil and the other the tachometer coil. As they are orthogonal, their fields do not interact, but each coil interacts with the permanent magnet of the armature.

Resonant scanners can also be built with cross-flexure suspensions. Dean Paulsen has also designed the ISX family of resonant scanner at General Scanning.[33]

Tunable resonant scanners have also been built. This author, in References 1 and 34, evaluates two designs. Low-frequency large-aperture resonant scanners are frequently built with cross-flexure suspensions. A common design can be found in Reference 1.

Table 8.5 presents performances of a number of commercial resonant scanners.

8.2.2.3 Induced Moving Coil

Two of the deficiencies of moving coil devices, lack of rigidity and moving electrical connections, can be bypassed by having a single turn coil energized by induction.[35,36] This
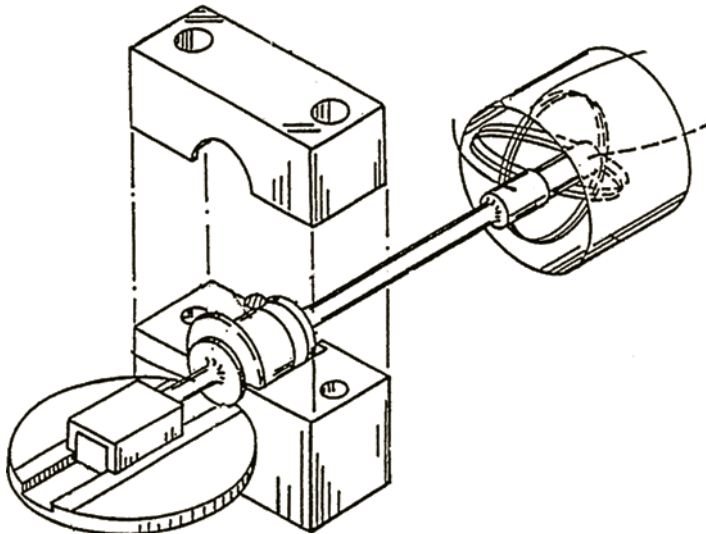
FIGURE 8.17
High-frequency tuned torsion bar resonant scanner.
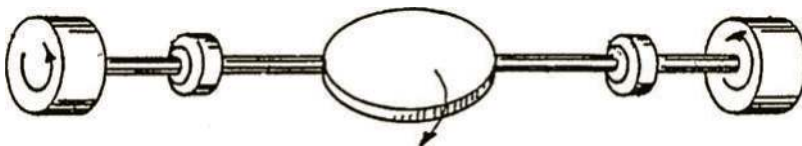
FIGURE 8.18
Low-frequency tuned torsion bar resonant scanner.

TABLE 8.5

Comparative Performances of Commercial Resonant Scanners

	Model							
	GRS	**IMX200**	**IM × 350**	**TRS**	**CRS4**	**CRS8**	**IDS**	**URS**
Mirror/beam diameter, mm	<36	28		<20	12.7	7.8	9	<30
Flatness, wave@633	1/10	1/6	1/6	1/2	1/2	1/4	1/2	
Suspension, type	X-Flexure	X-Flexure	X-Flexure	X-Flexure	Counter Rot.	T Bar	2 T Bars	S-Flexure
Resonant frequency, Hz	<250	200	350	< 10,000	4,000	8,000	< 1,200	<500
Beam rotation max., degrees	72	60	30	60	20	26	60	90
Beam wobble, max., micro rad.	2	2	2	2	100	150	100	NA

Notes:

[1] All units consume less than 1 W.
[2] All units incorporate a tachometer for self-excitation and amplitude control.
[3] Frequency stability 100 ppm/°C, typ.
[4] Line straightness, frequency, and amplitude jitter specifications should be requested.
[5] Large mirror units may impart considerable system vibration.
[6] Manufacturers: GSI Lumonics—IMX 200, IMX 350, CRS 4, CRS8, IDS Lasesys Corp.—GRS, TRS, URS.

technology is applicable to resonant scanners. A resonant scanner implementation is illustrated in Figure 8.16. One of the most novel resonant low-inertia scanners to have been introduced is the balanced torsion bar design with an induction torque driver. The mirror is suspended by two torsion bars in a fully symmetrical arrangement designed to be mass balanced, so that accelerations along the three principal axes of translation will not cause torsional excitation of the mirror.

The drive coil and the armature are magnetically linked in a transformer-like fashion by a soft iron core. The armature is a single-turn, rectangular drive loop with an edge colinear with the torsion current in the drive loop. This current interacts with the return path of the flux of a permanent magnet to create the drive torque:

$$T = \frac{mANIBlr}{LR} \tag{8.46}$$

where μ is the permeability of the iron core, A is the cross section of the iron core, N is the number of turns of the drive coils, I is the driver coil current amplitude, B is the field of the permanent magnet, l is the length of the drive loop in the magnetic field, r is the drive loop acting radius, L is the length of the iron core path, and R is the resistance of the drive loop. In this manner a moving coil driver is obtained without having to provide leads or brush contacts to the moving armature. The motion of the armature induces a voltage and current, which are sensed by a pick-off coil. The actual voltage measured is the sum of this induced voltage, which is proportional to velocity, and of the portion of the drive coil voltage induced by transformer couplings. It is expressed as:

$$E = \left(\frac{mAIN^2}{L} \right) w \cos wt + NBLr \frac{dq}{dt} \tag{8.47}$$

where ω is the resonant frequency. The transformer coupling component is easily subtracted electronically. This velocity sensor permits extremely good, simple, and external amplitude control with drift below 100 ppm/°C of the peak-to-peak excursion. The resonant frequency is stable to ~160 ppm/°C. A derivation of Equations 8.46 and 8.47 can be found in the appendix of Chapter 5 of Marshall.[1]

8.3 SCANNING SYSTEMS

Scanning systems can be treated analytically, but they involve a number of disciplines, most of which are treated superficially in this chapter. In laser machining applications the system performance also frequently requires a good understanding of the interaction of the work and the wavelength, radiation duration, and power of the laser, as well as possible damage to the optical elements of the scanning system. A number of manufacturers offer predesigned scanning packages also known as "scan heads." These are typically two- or three-axis vector scanning systems that include all scanning functions less the laser and the work surface. One of the valuable aspects of these packages is the matched driver-amplifier matched to the inertia of the mirrors. This is frequently the most economical and expeditious manner to construct a vector scanning system such as used for micromachining. The goal of this section is to present the numerous available choices.

The scanning applications may be divided into two major classes: raster and vector scanning. The former frequently involves a galvanometric scanner or a resonant scanner sweeping the fast axis and a stepper motor driving the slow axis—frequently moving the workpiece as this minimizes the size of the lens of a pre-objective scanner configuration. The latter demands that both axes have preferably identical dynamic behavior and is covered in Section 8.3.2. The former is exemplified in Section 8.5.2, "Microscopy", which looks at three raster optical scanning architectures, fixed objective, pre-objective scanning, and flying objective scanning.

8.3.1 Scanning Architectures

Scanning systems can be split into two categories, namely the beam moving category and the moving objective lens category. The two are interdependent but can be described separately. First we will explore the fixed objectives choices.

8.3.1.1 Post-objective Scanning

In this configuration the scanners are located between the objective and the work. This demands an objective with a long focal length to accommodate the scanning mirrors; the focal point will approximately paint a sphere. If the work area is flat, this restricts the scan to a small angle where the depth of field can approximately intersect a plane. If large angular excursions are required, the objective lens is normally translated on a linear stage to accommodate the need. The bandwidth of the translation mechanism needs to be twice that of the fastest scanner, but less accurate position control is required. A galvanometer is frequently the best choice to drive the translation stage[37] and a number of scanner manufacturers offer such an assembly.

When used in microscopy with comparatively fast optics, this arrangement operates at small angles and consequently cosine fourth law aberrations[26] are negligible. Also this

construction benefits for a comparatively small and low-cost objective where chromatic aberrations can be minimized.

This configuration is best applicable to long focal length objectives such as found in laser radars, range finders, and designators.

8.3.1.2 Pre-objective Scanning

This architecture locates the objective lens between the scanners and the work. It is the most common choice for laser micromachining as it permits a comparatively small spot size with high energy density. It is also the primary choice for conventional laser scanning confocal microscopes and Section 8.5.2.2 depicts the original design of Minsky.[38] The *Handbook of Biological Confocal Microscopy*[39] carries numerous examples of this concept.

A number of lens manufacturers offer off-the-shelf lenses for this application configured for YAG or CO_2 lasers. Multiple wavelength telecentric field-flattening objectives of this type, capable of yielding a diffraction limited small spot focus, are frequently the most expensive component of the entire scanning system, including the laser.

All the beam steering systems described in Section 8.3.2 are suitable for use as pre-objective systems.

8.3.1.3 Flying Objective Scanning

This scanning architecture demands that the work be moved in one axis while the beam is moved in the other axis. It is extremely advantageous for microscope raster scanning and is the preferred choice for biochip scanners as described in Section 8.5.2.3. It is also used for random axis scanning in the semiconductor industry for DRAM repairs. It offers a very low-cost system with multiple wavelength telecentric flat-field, diffraction limited small spot performance.

8.3.2 Two-Axis Beam Steering Systems

Oscillatory scanners are best suited for large excursion systems, over a few hundredths of a radian. Single-mirror or small motion two-axis beam steering (TABS) systems have been hotly pursued for SDI applications, and the *SPIE Proceedings,* Vol. 1543 as well as the *IR/EO Handbook*[40] examine a number of them. Conventional gimbal systems, where the torque motor for the second axis is transported on the structure mounted on the first axis are reviewed in Section 8.5.2.

Sections 8.3.2.1 through 8.3.2.6 illustrate some of the most common architectures of two-dimensional scanning systems. They fall into two major classes: vector scanning and raster scanning. Vector scanning applications require that both axes have equal properties and normally they have two galvanometric scanners. Raster systems frequently have a more diversified architecture. Some use a polygon or resonant scanner for the fast or raster scan and a galvanometer, motor, or a linear transport for the other motion. The *LR/EO Handbook*[40] describes a number of such systems.

8.3.2.1 Single-Mirror TABS

The device symbolically described in Figure 8.19 is capable of 1 rad motion in both axis. The drivers and encoders for both axes are stationary. The torque capabilities, range, and angular resolution are dissociated from the inertial load. The optical system behaves like a
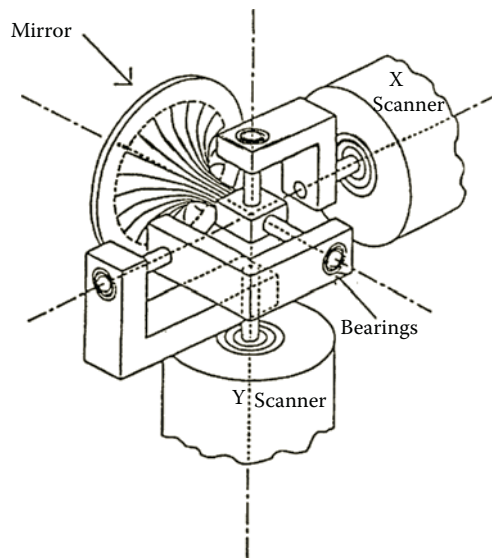
FIGURE 8.19
Single-mirror TABS.

true point source with double pincushion distortions. The central block is on the origin of all three coordinates. The floating "L" bracket is necessary to prevent bearing lock.

8.3.2.2 Relay Lens TABS

This construction guarantees that both mirrors and scanners have equal performances. It is penalized by the added two optical elements with their associated cost and/or distortions. The distortions are again in the familiar pincushion pattern and can be compensated for in the design of the objective lens or the computer program. Transmission optics are frequently used, but when chromatic aberrations are critical, reflective optical elements are preferred, as shown in Figure 8.20. It should be noted that the two axes need not be perpendicular.

8.3.2.3 Classic Two-Mirror Construction

Figure 8.21 is a generic model of this construction and is used to derive image distortions and focus variations. In this configuration the two mirrors have different inertia. In order to minimize their difference, the X galvanometer can be mounted at 15–20° from its perpendicular axis. This allows a closer packing construction and a smaller Y mirror. The incoming beam must be parallel to the axis of the Y scanner.

In this configuration, a is the center of the X mirror, b is the center of the Y mirror, and c is the point at coordinates $(0, Y_i)$; d is the length from b to $(0, 0)$ and e is the length from a to b. The optical scanner angles are θ_x and θ_y and the coordinate (X_i, Y_i) is any point on the target field. It can be seen that when $X_i = Y_i = 0$, then $\theta_x = \theta_y = 0$. The equation that relates Y_i to θ_y is derived from the triangle of points $(0, 0)$, $(0, Y_i)$, and d. Solving for the length $(0, 0)$ to $(0, Y_i)$, which equals Y_i we obtain

$$Y_i = d \tan q_y$$

$$(8.48)$$
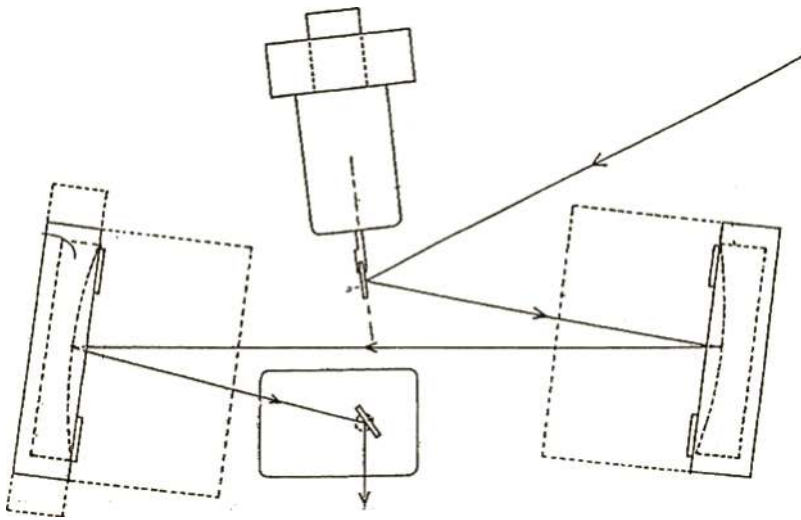
FIGURE 8.20
Relay lens TABS with reflective optics.
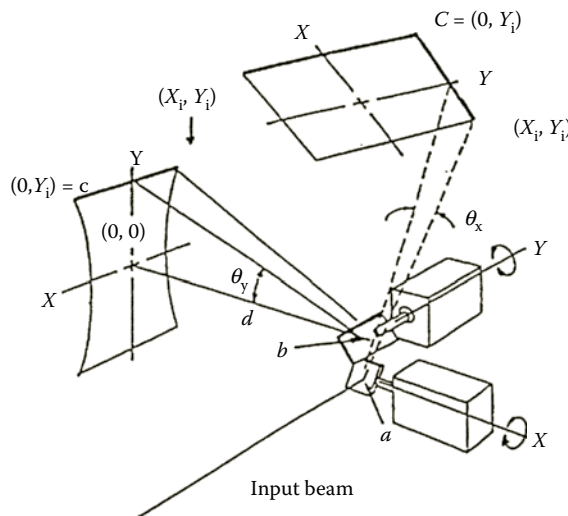
FIGURE 8.21
Two-mirror, two-axis flat-field assembly.

The determination of the X equation is somewhat more complex and is best illustrated by projecting the target image onto the virtual image position of the Y mirror, as shown by the phantom lines and phantom coordinates $(0, Y_i)$, (X_i, Y_i), and a in Figure 8.21. By solving the triangle of points a, $(0, Y_i)$, (X_i, Y_i), for the length $(0, 0)$ to $(0, Y_i)$, which equals X_i, we have

$$X_i = ac \tan q_x \tag{8.49}$$

Since $ac = \left(d^2 + Y_i^2\right)^{1/2} + e$ where $e = ab$, the solution is

$$X_i = \{(d^2 + Y_i^2)^{1/2} + e\} \tan q_x \tag{8.50}$$

If we solve for the length from a to (X_i, Y_i) we obtain the equation for the focus length:

$$f_i = [\{(d^2 + Y_i^2)^{1/2} + e\}^2 + X_i^2]^{1/2} \tag{8.51}$$

The resulting change in focus length for (X_i, Y_i) is

$$\Delta f_i = [\{(d^2 + Y_i^2)^{1/2} + e\}^2 + X_i^2]^{1/2} - (d + e) \tag{8.52}$$

In looking at pincushion errors in simple two-mirror systems, we see that X_i (Equation 8.50) can be combined with Y_i (Equation 8.48) to yield

$$X_i = \left(\frac{d}{\cos q_y} + e\right) \tan q_x \tag{8.53}$$

The pincushion error ε is the ratio of the change in the value of X_i as θ_γ changes from zero to a specified value, to the peak-to-peak amplitude $2 X_i$ at $\theta_y = 0$:

$$e = \frac{X_{iqy} - X_{i0}}{2X_{i0}} = \frac{(1 - \cos q_y)}{2(1 + e/d)\cos q_y} \tag{8.54}$$

8.3.2.4 Paddle Scanner Two-Mirror Configuration

In applications that need large excursion angles, high speed, and high precision, the single-mirror two-dimension scanner techniques described above exhibit handicaps. Paddle scanners are of great interest because they simulate a two-dimensional fulcrum. Figure 8.22 gives a symbolic representation.
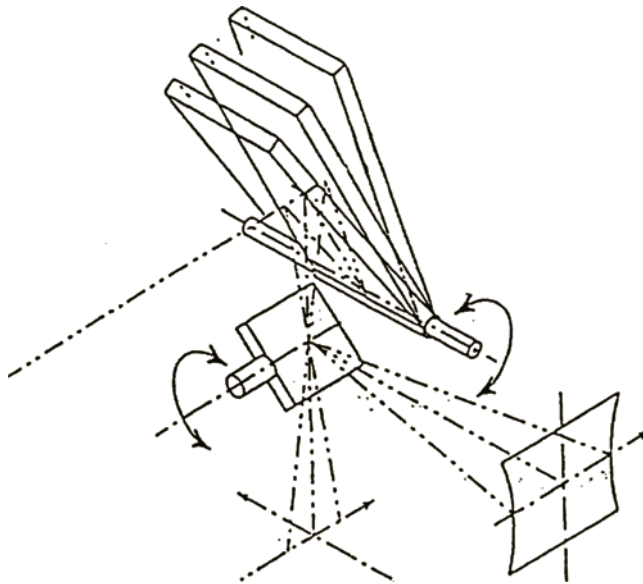
FIGURE 8.22
Paddle scanner configuration.

The inertia of the paddle is greater than the inertia of the other mirror. This configuration adapts well to raster scanning. The second scanner can be either a resonant scanner with sinusoidal, triangular, or saw tooth motion, or a rotating polygon. It should be remembered that at all high angular speeds, the loads should be statically and dynamically balanced in all axes.

The first mirror is mounted on a paddle-like arm and intersects the incoming beam at 45°. The mirror rotation, the size of the beam, and other geometric constraints determine the magnitude of this motion. The magnitude of lateral motion of the reflected beam at its pupil is small and is analyzed in Figure 8.23.

Note that the axis of rotation of the X scanner, the frame scanner, is within the plane of the mirror. In the drawing, the rest position of mirror OM is at 45° of the incoming beam. Radius ON is normal to incoming beam AN. Figure 8.23 also shows mirror OM rotated by an angle α and the reflected beam OM′ is rotated by the angle 2α to line Aa. Lines OB and OD are normal to the reflected beams at the two positions of the mirror. It is evident that point B is on the circle of radius ON. We shall show that point D is also on that circle.

From the above we can write:

$$\text{MEA} = 2a \tag{8.55}$$

$$\text{NOM} = \text{NOA} - a = \frac{p}{4} \tag{8.56}$$
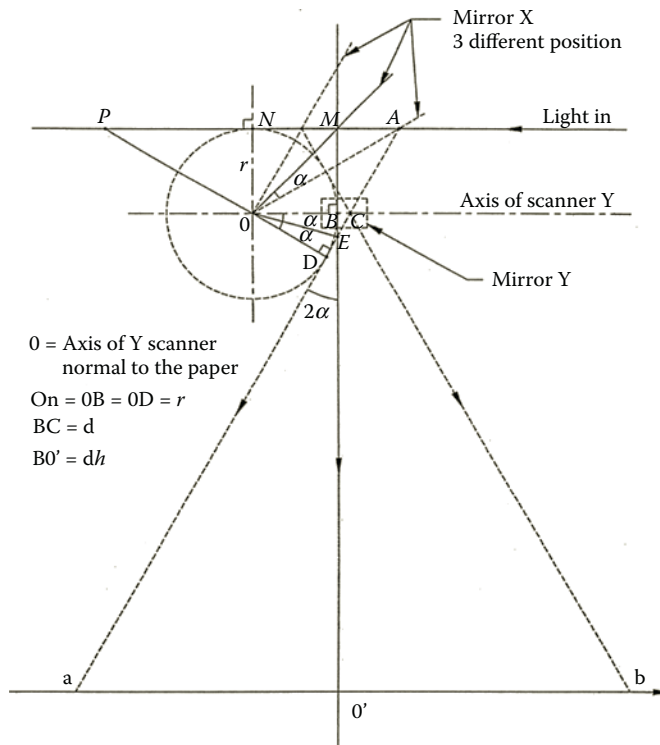
FIGURE 8.23
Paddle scanner pupil motion.

$$NOA + NAO = \frac{p}{2}$$

Consequently angle

$$NAO = \frac{p}{4} - a \tag{8.57}$$

Considering that angles APD and aEO' are equal as both their sides are normal, therefore

$$angles\ APD = aEO' = 2a$$

As triangle PAD is rectangular

$$angle\ NAD = \frac{p}{2} - APD = \frac{p}{2} - 2a = 2NAO \tag{8.58}$$

As angle NAD = NAO + OAD = 2NAO, we can conclude that angles NAO = OAD. The two rectangular triangles ONA and ODA have three equal angles and one common side, therefore they are equal and segments ON = OD and point D is on the circle centered at O and passing through point N and B.

It is therefore possible to derive the amount of translation the beam has on mirror Y as mirror X rotates. It is represented by the segment BC. Considering the triangle OCD, we can write

$$BC = OD\left(\frac{1}{\cos 2a} - 1\right)$$

or, using symbols d and r,

$$d = r\left(\frac{1}{\cos 2a} - 1\right) \tag{8.59}$$

It should be noted that the cosine does not change sign when the angle does and that d is always positive.

For an optical system where the distance between the mirrors is 10 mm and the angle of rotation of the frame mirror on the Y scanner is ±0.15 rad, the beam walk d on the Y mirror is 0.33 mm.

8.3.2.5 Golf Club Two-Mirror Configuration

The diagram of a two-dimensional scanner known as a golf club scanner in Figure 8.24 has similar features to the paddle scanner. The inertia of the page mirror and its mount is typically double that of an equivalent paddle for the same mirror distance. Since some rotation and geometric constraints make this geometry desirable, it is reviewed here.

The defining feature of this arrangement can be seen in Figure 8.24. In a raster scan application, the beam strikes the frame mirror first, the Y-axis, which is held at 45° from the incoming beam. This mirror is mounted on an arm whose axis of rotation intersects and is normal to the reflected beam at the rest position. The axis of rotation of the second mirror (the X mirror with reference to Figure 8.25) is in the plane formed by these two lines. It is
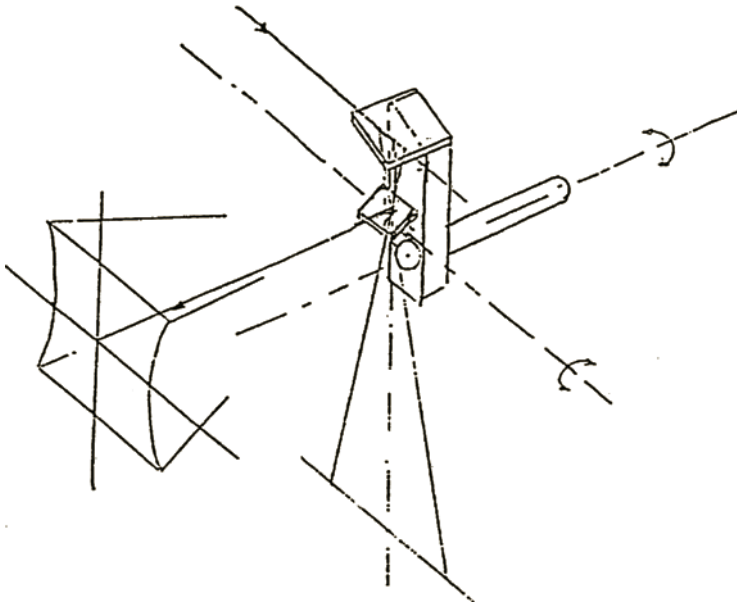
FIGURE 8.24
Golf club TABS.

located approximately halfway between the point of incidence of the beam on the Y mirror at rest and its axis of rotation.

We shall show now that as the Y mirror oscillates, the reflected beam passes through the pupil location X in Figure 8.25. We shall also derive the magnitude of "walk" st of the beam on the X mirror as a fraction of the radius of rotation $P_1Y = r$ of the Y mirror.

As an approximation we have

$$st = \frac{mn}{2 * \tan 2q} \tag{8.60}$$

As

$$mn = P_1 m - P_2 n \tag{8.61}$$

it can be seen that

$$P_1 m = \frac{P_1 P_2}{\tan 2q} \tag{8.62}$$

Also

$$PrP_2 = r * \sin q - ef \tag{8.63}$$

and segment

$$ef = h * \tan(p/4 - q) \tag{8.64}$$

where

$$h = r * (1 - \cos q) \tag{8.65}$$
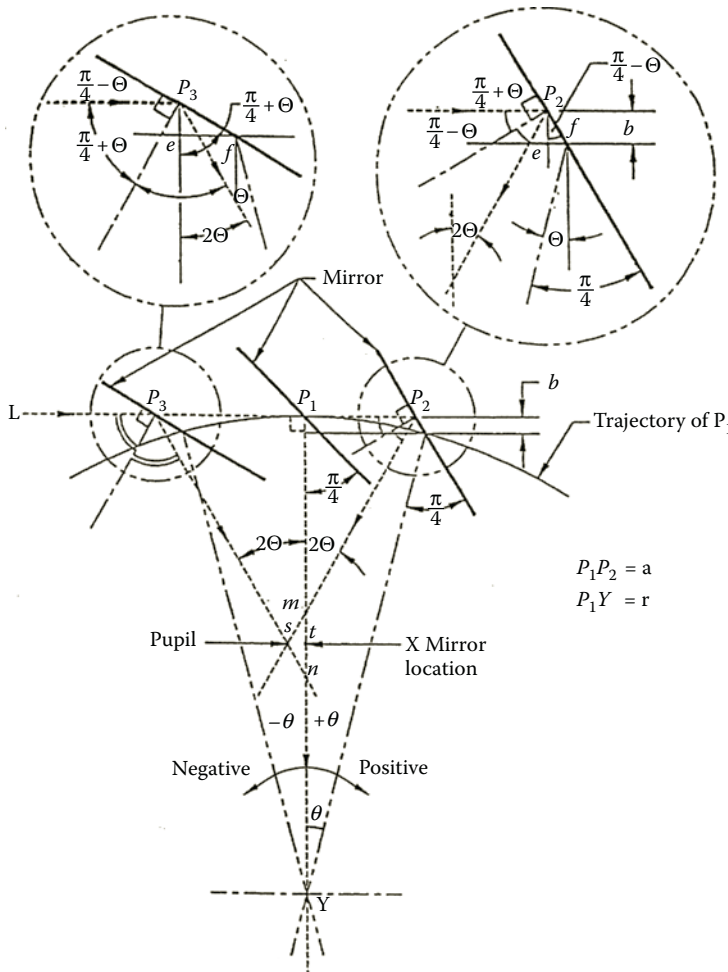
FIGURE 8.25
Golf club scanner.

The above simplifies to yield

$$P_1 m = \left\{ \sin q - [1 - \cos q] \tan\left(\frac{p}{4} - q\right) \right\} \frac{r}{\tan^2 q} \tag{8.66}$$

Values for $P_2 n$ are derived similarly for the angle—θ

$$P_2 m = \left\{ \sin q + [1 - \cos q] \tan\left(\frac{p}{4} - q\right) \right\} \frac{r}{\tan^2 q} \tag{8.67}$$

and going back to Equations 8.60 and 8.61 again, as a first approximation we have

$$st = \left\{ \tan\left(\frac{p}{4} - q\right) + \tan\left(\frac{p}{4}\right) \right\} (1 - \cos q) \frac{r}{2} \tag{8.68}$$

Comparing this geometry to the paddle, for a rotation of ±0.15 rad and a 10-mm mirror spacing, the beam walk $st = 0.24$ mm.

8.3.2.6 TABS with Three Moving Optical Elements

The dominant feature of the design in Figure 8.26 is the emulation of a perfect fulcrum with both axes having approximately the same dynamic capability. The "conditioner" scanner carries a glass wedge and is synchronized to the Y scanner. It translates the incoming beam onto the Y mirror so that the reflected beam always strikes the X mirror in its center. Despite the added cost, this is the preferred design for achieving high speed with large-beam wide-angle vector scanners, especially when focusing optics are needed. The details of this design can be found in Goodman, U.S. patent no. 4,685,775.

8.4 DRIVER AMPLIFIER

High-performance servo amplifiers for galvanometric or resonant scanners are very special products. They are critical to extract the performance of the galvanometer. They are as critical a system component as the galvanometer or the resonant scanner themselves. Commercial drivers are offered by scanner manufacturers and quoted performances are normally guaranteed only when the associated driver is used.

The design of these amplifiers is beyond the scope of this chapter. The list of elements to be addressed is also beyond the scope of this chapter and this author strongly suggests that anyone considering such a task should first refer to Reference 41, where the parameters that need to be addressed to design a high-performance analog servo amplifier are reviewed.
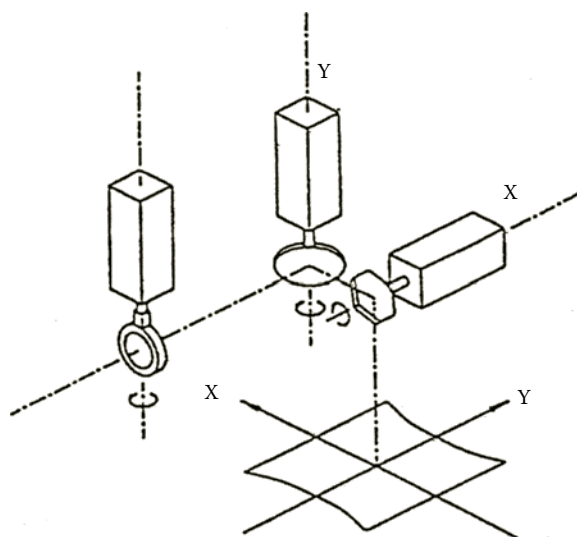
FIGURE 8.26
Three-optical-element TABS.

At this time, all commercial galvanometer driver amplifiers are analog systems. Digital servos are commercially available to drive servo motor-based systems. The numerous attempts to design a high-performance digital equivalent have failed to meet the comparable bandwidth resolution product demanded by optical galvanometer scanners. The potential benefit of a full digital system continues to tempt and elude designers. Hope springs eternal.

Two critical features of these servo amplifiers are:

1. *Low noise.* It is common to demand system resolution of 1 part in 100,000. Consequently each component in the system needs to perform five or ten times better. The design of analog circuitry of that quality is an art in itself.

2. *Frequency response compensation/filtering* for the galvanometer/mirror resonances.

8.5 SCANNING APPLICATIONS

The earlier edition of this text[2] carries numerous examples of applications that have shaped the development and performance of optical scanners, both galvanometric and resonant. Here are two categories that have seen vigorous development in the past decade. These are laser micromachining and confocal microscopy.

8.5.1 Material Processing

The market for laser-based material processing has grown enormously in the last decade and has segmented itself. This is the result of a number of factors:

1. New high-tech applications with no practical alternative solutions
2. Penetration into the commercial market, pencils, and kitchenware
3. Reliable lasers with a choice of wavelengths and sufficient power
4. Scanners with sufficient precision
5. Software packages complete with a powerful low-cost PC

The general system architecture has not changed appreciably. Single lasers, multiple heads, and fiber optic laser light distribution are more common. In general, the market is divided into marking and micromachining.

In the laser-marking field, the CO_2 laser dominates and as the need to ensure good marking visibility suggests large spot size and consequently small scanning mirrors as beam diameters between 10 and 20 mm cover most applications. High scanner speed is desirable to produce high throughput. Since the industrial waveguide laser was developed in the early 1980s—10–50 W originally, and now as much as 200 W—its low cost and reliable design have stimulated this segment of the industry.

The CO_2 laser is also the preferred choice to mark packages on the fly or erode plastic packages to indicate and facilitate opening.

The YAG laser competes in marking applications for plastics, glasses, and paints that incorporate titanium oxide powder. This chemical exhibits very visible and permanent color changes when exposed to 1.06 nm wavelengths. It is frequently used for kitchenware.

The micromachining segment of the industry addresses a variety of applications where different technologies compete. Scanning systems dominate applications where large areas of precisely located features need to be created or where a threshold of power/energy is required. Materials such as metals, polymers, and ceramics are commonly machined in this manner. Typical applications are:

> Nozzles for ink jet printing: 1- or 2-micron feature tolerances with submicron positioning and concentricity are required.

> Screens/sieves/membranes for filtering inks, biomedical fluids, gas separation, and so on. It is common to punch thousands of holes 15 microns in diameter, 20 micron on center, in 10-micron Kapton sheets for such applications.

> Via drilling for printed circuit boards or microchip modules or green ceramic.

> Probe cards for circuit autotesters may have 2000 to 3000 contact wires piercing a 2-in-square support.

In order to reach industrial applications, lasers need to be reliable and easy to maintain. A number of technologies have reached this stage. High-power diode pumped YAG lasers are now available at the following wavelengths: 1060, 532, 473, and 351 nm. The shorter wavelengths, pulsed or CW, compete efficiently with Excimer lasers for most polymer or organic micromachining applications and are also used to stimulate optochemical reactions.

8.5.2 Microscopy

In the last decade scanners have participated in the expansion of microscopy. First came confocal microscopy and more recently large field-of-view microscopy. In these applications, images are assembled on a computer.

The development of confocal microscopy in the early 1950s by Marvin Minsky[38] has opened an array of applications for scanners. In this concept, the image of the work is constructed from single pixels acquired sequentially. This minimizes the demand on the optical design and offers major imaging benefits. Among these are sectioning capability, akin to tomography and the direct digitization of the image. In his original design, Minsky moved the work on a motorized *XY* stage under a conventional microscope. His invention gave images with greater resolution than had been available before, but the images were composed on a computer monitor and a number of decades had to pass before the technique was accepted.

Confocal microscopes are now commonly used either to improve resolution or to capture a large field in a single file without stitching a number of small fields of view. Scanners are most frequently incorporated in both classes of instruments. For simplicity, this review will address only the large field-of-view (centimeters square) model, as they encompass concepts found in all designs used to image arrays of biological material.

Arrays of biological material consist of large number of features such as dots or squares bonded to a glass substrate and labeled with fluorescent molecules. The role of imaging systems is to view the entire array in a single image and present it for analysis to a specialized software program. High-density arrays may carry as many as one million separate 10 microns square. Low-density arrays have 400 spots per square centimeter; here each spot may be 150 micron in diameter, placed on a 200–300-micron grid. Large arrays are typically 22 × 66 mm square. Three basic technologies are used to expand the reach of

conventional confocal fluorescent microscopes to meet the needs of biochip imaging and in all cases galvanometer scanners are the preferred actuator.

Scanning architecture	Manufacturers
XYZ stage: Minsky design	Packard, Brown Lab (website)
Pre-objective scanning	Agilent, MD/Amersham
Flying objective scanning	Affymetrix, Virtek/BioRad, Axon

The target typically measures 1–2 cm square and the resolution needs to be between 2 and 10 microns. Fluorescence is an extremely inefficient phenomenon, so the image quality is extremely dependent upon the energy collection capability of the instrument. All of these instruments are epifluorescent pseudoconfocal laser scanning microscopes.

8.5.2.1 Pre-objective Scanning

Conventional laser scanning epifluorescent microscopes are thoroughly described in the *Handbook of Confocal Microscopy*[39] and have been commercially available for the past 25 years. Zeiss, Nikon, BioRad, and Olympus are common names in that field. They are all built from conventional microscope objectives. Because fluorescence is a very low-energy phenomenon, a high numerical aperture is preferable. This yields a very small field of view, well under 1 mm square, so that multiple images are required to cover the large field of view associated with microarrays. By limiting the spectral range to 1 or 2 wavelengths and accepting a lower NA, it is possible to design, on this model, an instrument with a larger field of view (up to 10 mm). The Avalanche from Molecular Dynamics/Amersham, symbolically depicted in Figure 8.27, is a scanner of this construction. It has a nine-element objective designed for a 10-mm field of view with a comparatively low NA, approximately 0.25. A similar instrument made by Agilent (formerly HP) has a 15-mm wide field of view. In order to make best use of the lens, the biochip or the microscope slide is translated under the objective and the beam scanned on the other coordinate.
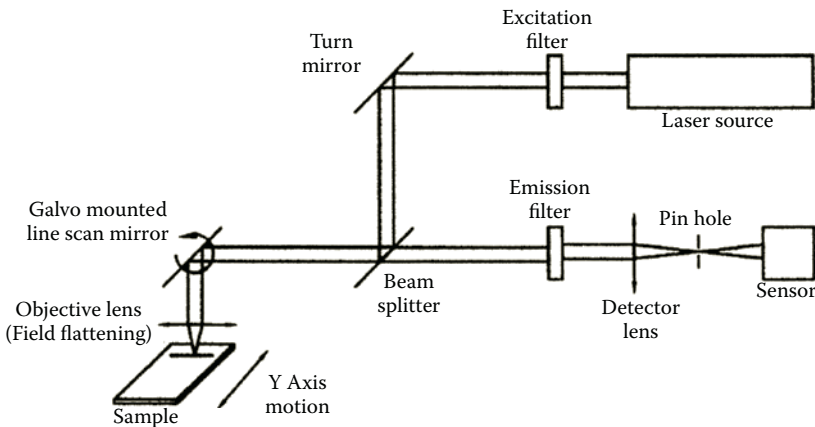
FIGURE 8.27
Pre-objective line scan architecture.

8.5.2.2 The Marvin Minsky Confocal Microscope

The Minsky design shown in Figure 8.28 can be built with a commercial microscope objective. The working clearance of high-power objectives is quite small and the slide needs to be held reliably to avoid catastrophic interference in a dynamic environment. The slide is mounted on a motorized XY stage. An optical scanner is preferably used to drive the high-speed scan stage as it outperforms a DC motor. The slide anchor mechanism must also be light enough to permit a reasonable scan rate without excessive vibration.

8.5.2.3 Flying Objective Scanning Microscope

The flying objective architecture transposes the difficult requirements posed in designing a large flat field-of-view lens by replacing it with a relatively simple objective lens, which is moved across the sample area. In other words, instead of moving the slide and its holding device in the fast axis, the beam of light is moved instead. This approach requires that good mechanical alignment be maintained throughout the entire range of motion, and carries the additional benefit from the use of a much larger numerical aperture objective lens. It also offers a high degree of measurement uniformity across the field of view as all pixels are acquired in an identical manner so that field flatness is not a problem.

Two basic designs are used: one design oscillates an objective lens mounted on a linear rail, and the other design oscillates an objective lens mounted at the end of a rotating arm.

8.5.2.4 Rectilinear Flying Objective Microscope

Hueton[42] describes this architecture. As shown in Figure 8.29, the objective is translated on a linear stage and it is optically "tromboned" to the stationary optical elements. The objective can be driven by a linear motor, a voice coil, a stepper motor, but again, optical scanners offer the fastest drive mechanism. The scanning system offered by the Virtek Instrument oscillates a lens using a scanner to drive a stage in a similar construction.
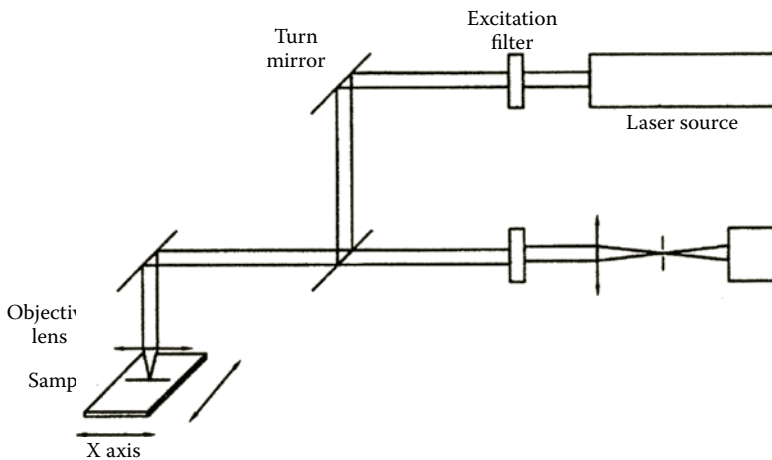
FIGURE 8.28
Moving stage architecture.
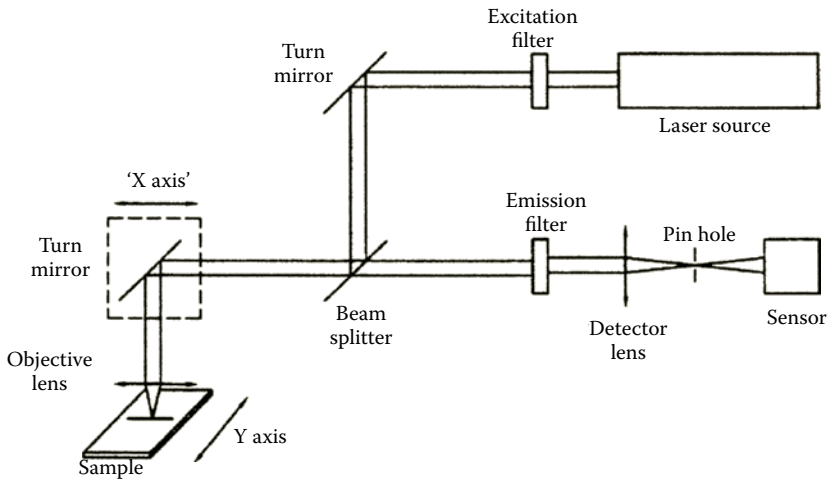
FIGURE 8.29
Rectilinear flying objective architecture.
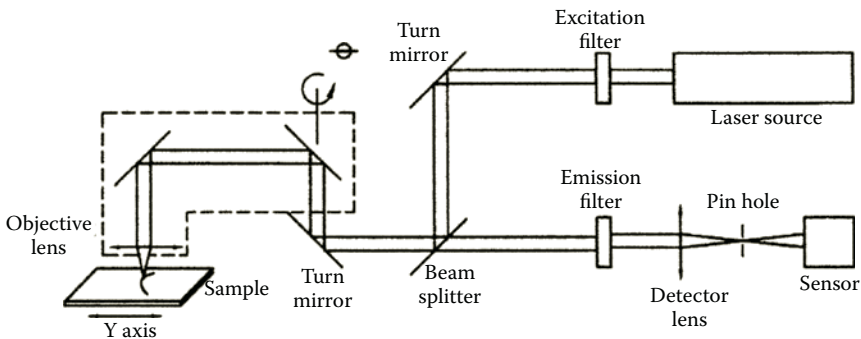
FIGURE 8.30
Rotary flying objective architecture.

8.5.2.5 Rotary Flying Objective Microscope

An alternative method to create a rapidly scanning beam of light is the rotary architecture described by Overbeck and shown in Figure 8.30.[46] It offers both speed and a constant optical path length. A periscope couples the scanning objective to the stationary elements of the microscope and thus keeps the light path length constant as it oscillates to scan the light beam. As the arc covering the width of the slide is in polar coordinates, the resultant image is instantly converted to Cartesian coordinates by the instrument's computer, so the image is then directly correlated with the sample spot pattern. A micro, aspheric, one-element lens is carried on a counterbalanced arm that extends perpendicular to the shaft of an optical scanner. The arm also carries mirrors that receive laser light on the axis of rotation via a periscope, send it out the length of the arm, and down through the lens. The optical axis is always normal to the biochip.

8.6 CONCLUSIONS

This chapter has described the current state of the art of optical scanning galvanometers as well as recent development of resonant scanners. These devices have developed over the past 40 years as part of the general electro-optic revolution.

All components of the electro-optic industry have witnessed a drastic price reduction in conjunction with a dramatic increase in demand. Galvanometric and resonant scanners may be a rare exception to that pattern of evolution. Prices have grown with demand.

The elements used to construct these oscillating devices are the same as those found in the head movers of every computer hard drive or in CD players or high-quality motors that can be purchased for a few hundred dollars.

High-volume applications have stimulated the development of other technologies: rotating polygones, MEMS, LIC scanners, and so on.

Presently the total worldwide annual sales of oscillating scanners lies between $20 million and $30 million with an additional $15 million to $20 million of annual sales of subsystems integrating scanners with their driver amplifiers and some optical elements.

GLOSSARY

This short list of definitions is offered so that it may help the reader. It is a subset of terminology recorded by Alan Ludwiszewski,[43] and is commonly accepted in this industry.

Accuracy: The maximum expected difference between the actual and commanded position. This includes any nonlinearities, hysteresis, noise, encountered drifts, resolution, and other factors.

Bandwidth: The maximum frequency for which a system can track a sinusoidal input with an output attenuated to no less than 0.7 (−3 dB point) of the command. For open loop frequency responses with a phase margin of 90°, the open loop crossover frequency will equal the closed loop bandwidth. For other phase margins the

relationship is not as straightforward. A complete treatment of these relationships can be found in any control theory text.

Drift, Mechanical Null: The drift of the steady-state position of an unpowered scanner when the armature is restrained with a torsion bar. This drift can occur with time and with temperature. Drift is usually specified in terms of the change in optical angle per amount of correlated influence, such as time or temperature.

Drift, Position Detector: The change in relationship between the output of the position detector and the position of the output shaft. It consists of the sum of the gain drift, null drift, and others.

Drift, Position Detector Gain: The change in scale factor of the position detector. Since the absolute magnitude of this change is dependent upon angle, it is specified in terms of the ratio of the change in output over the output per unit time or temperature (that is, ppm/°C or %/1000 h). This takes into account that the effect is seen most at extreme angles where the output is greatest.

Drift, Position Detector Null: The drift of the electrical zero of the position detector with time and temperature. Drift that occurs with temperature change is specified in units of angle per degree temperature (that is, μrad/°C). Drift that occurs with time is specified in units of angle per units of time (that is, μrad/1000 h).

Drift, Uncorrelated: Drift that cannot be attributed to a change in a particular external condition, such as time or temperature. Often caused by mechanical ratchetting or noncatastrophic damage to the system due to overstress.

Jitter: Nonrepeatable position error fluctuations caused by velocity perturbation in a scanner. Generally described in units of optical scan angle and often expressed as the standard deviation of the maximum jitter error observed in each scan line of a large number of consecutive scans. Some applications may require specification of the frequency as well as the magnitude of acceptable jitter.

Nonlinearity, Best Fit Straight Line: This method of quantifying nonlinearity involves finding a first-order linear function that is the closest approximation to the measured data. The nonlinearity is then calculated as the maximum observed deviation from this line. This will result in the smallest measurement of nonlinearity.

Nonlinearity, Pinned Center: Pinned center nonlinearity uses a straight line that intersects a given datum point, such as the mechanical or electrical null of a scanner, and has a slope that best approximates that of the measured data. Sensor nonlinearity is then calculated from this reference.

Null, Electrical: The zero output point of the position transducer.

Null, Mechanical: The steady-state position of an unpowered scanner. This position is determined by the torsion spring, if any, and the magnetic spring of the scanner. In many scanners without a torsion spring, the magnetic spring is not of great enough magnitude to overcome frictional forces and make this an absolute position.

Repeatability: The inaccuracies in final position encountered while implementing a series of identical command inputs.

Repeatability, Bidirectional: The inaccuracies in final position encountered while attempting to return to a position from different directions.

Resolution: Resolution is the ability to discern individual spots in the target field. This is not to be confused with accuracy, which includes gain and offset drift, noise, resolution, and other factors. Dependent upon system design, the limit to resolution may be due to optical considerations, digital resolution, or position detector signal-to-noise ratio and drift.

Resolution, Optical: The optical resolution of a scanning system can be described as the number of separately resolvable spots that can be produced. For diffraction limited optics, this is dependent upon the aperture width in the scan direction, the aperture shape factor, the wavelength of the source, and the total scan angle. These factors are related through the scan equation, which can be found in many optical texts.

Resolution, Scanner: Scanner resolution is limited by the noise and drift of the position detector. The RMS signal-to-noise ratio will determine the statistical resolvability of a given level of command in a given frequency range. Filtering can improve low-frequency resolution, but drift factor will also come into effect.

Resonance, Cross Axis: Structural resonances that cause motion perpendicular to the scan axis are referred to as cross-axis resonances. These resonances may be accentuated by poor mirror design and will cause periodic wobble, possible system instabilities, and may be a limit to achievable system bandwidth.

Ideal distribution of pixels along a scan line

Jitter

Distribution of pixels with jitter error

Jitter: Error from average pixel position, measured in the scan direction

Wobble

Distribution of pixels with wobble error

Wobble: Error from average pixel position, lmeasured in a direction perpendicular to the scan direction
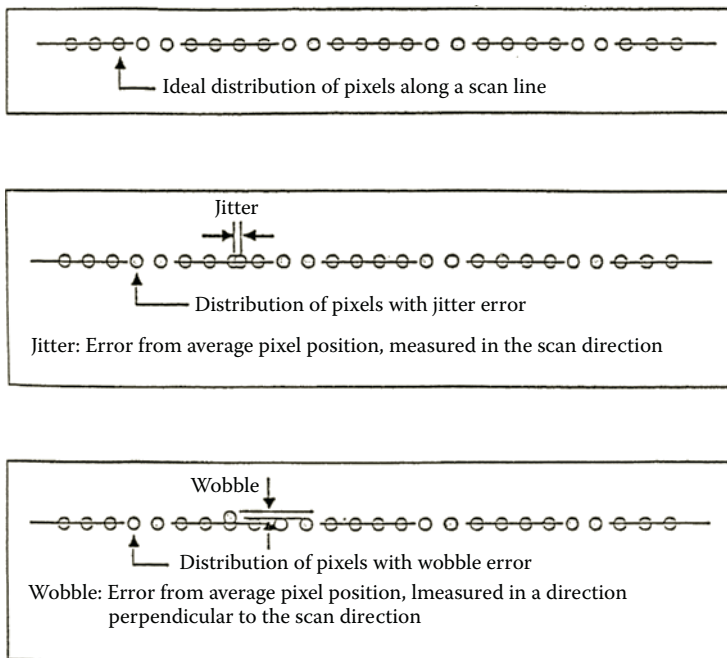
Resonance, Torsional: An on-axis resonance that appears in scanners due to the distributed masses on a compliant rotor shaft the flexible coil of a d'Arsonval system. These resonances can appear as periodic jitter and may cause controllability difficulties due to the resonant peak created in the scanners' transfer function. Mirror design and mounting will have a significant effect on torsional resonance.

Response Time: The response time of a scanning system is defined as the tracking error divided by slew rate. Owing to the characteristics of the controller, stage saturation, aerodynamics, and other nonlinearities, response time will not necessarily be a constant. Although not a constant, at least for vector tuned controllers it is approximately so, if slew rates are neither driving the tracking error to near zero nor approaching the maximum slew rate. Response time is an indication of the relative speed of a scanner and the ultimate performance obtainable with a given load and tuning.

REFERENCES

1. Marshall, G.F. *Optical Scanning;* Marcel Dekker, Inc.: New York, 1991.
2. Marshall, G.F. *Laser Beam Scanning;* Marcel Dekker, Inc.: New York, 1985.
3. *Permanent Magnet Data Sheet;* UGIMAG, Ugimag 45M2 material.
4. Roters, H.C. *Electromagnetic Devices;* John Wiley & Sons, Inc.: New York, 1955.
5. Keller. U.S. Patent nos. 985,420 and 1,041,293.
6. Hodges. U.S. Patent no. 3,348,183.
7. Houtman, J.A. U.S. Patent no. 3,528,171.
8. Rohr, B. U.S. Patent no. 4,864,295.
9. John Weisz, *Proc. SPIE* 1991, 1454, 265–271.
10. Dillon, R.; Trepanier, P. U.S. Patent no. 6,000,030.
11. Montagu, J.; Honkanen, P.; Weiner, N. U.S. Patent no. 6,218,803.
12. Montagu, J.I. U.S. Patent no. 5,235,180.
13. Ivers, R. U.S. Patent no. 5,844,673.
14. Ivers, R. U.S. Patent no. 5,671,043.
15. Abbe, R. U.S. Patent no. 3,990,005.
16. Rohr, B. U.S. Patent no. 41,864,295 and 4,142,144.
17. Stokes, B. U.S. Patent no. 5,099,368.
18. Dowd, R. U.S. Patent no. 5,537,109.
19. Wittrick, W.H. Properties of cross flexurempivots and the influence of the point at which the stripcross. *Aero*. Quat. 2, 1951, 272–292.
20. Siddall, G.J. "The design and performance of flexure pivots for instruments." M. Sc. Thesis, University of Aberdeen, Sept. 1970.
21. Reiss, R.S. Optomechanical system engineering: optomechanical instrument design, OE *reports SPIE* 0817, 1897, 154–170.
22. Brosens, P.J.; Vudler, V. *Opt. Eng.* 1989, *28,* 61–65.
23. Timoshenko, S.; Goodier, J.N. *Theory of Elasticity;* McGraw-Hill: New York, 1951.
24. Weissman, H. Replicated mirrors. *Opt. Eng. 5,* 1976, 435–441.
25. Yoder, P. *Opto-Mechanical System Design;* Marcel Dekker: New York, 1986; 71–77.
26. Smith, W.J. *Modern Optical Engineering;* McGraw-Hill: New York, 1966; 132–133.
27. Lawler, A.; Shepherd, J. *Laser Beam Scanning;* Marshall, G.F., Ed.; Marcel Dekker, Inc.: New York, 1985; 125–147.
28. Marshall, G.F. *Optical Scanning;* Marcel Dekker, Inc.: New York, 1991; 560.
29. Den Hartog, J.P. *Mechanical Vibrations;* McGraw-Hill: New York, 1956; 396.
30. Paulsen, D.R. U.S. Patent no. 4,878,721.
31. Paulsen, D.R. U.S. Patent no. 4,919,500.
32. Paulsen, D.R. U.S. Patent no. 4,990,808.
33. *J. Laser & Optronics* 1998, *17* (2).
34. Montagu, J.I. Tunable resonant scanner. *Proc. SPIE817,* 1987.
35. Montagu, J.I. Induced moving coil resonant scanner. J.I. *Electro Optics* 51–56, May 1983.
36. Montagu, J.I. U.S. Patent no. 4,502,752, 1985.
37. Montagu, J.I.; Pelsuel, K. U.S. Patent no. 4,525,030.
38. Minsky, M. Microscopy Apparatus. U.S. Patent no. 3,013,467.
39. Pawley, J.; Ed. *Handbook of Biological Confocal Microscopy,* 2nd Ed.; Plenum: New York, 1995.
40. Roggatto, W.D., Ed. *IR/EO Systems Handbook;* SPIE Press: Bellingham WA, 1993.
41. Albert Bukys. *Proc. SPIE* 1991, 1454, 185–195.
42. Huerton, I.; Van Gelder, E. High Speed Fluorescence Scanner. U.S. Patent no. 5,459,325,1995.
43. Alan P. Ludwiszewski. *Proc. SPIE* 1991, 1454, 174–185.
44. Hueton, I. High Speed Florescent Scanner. U.S. Patent no. 5,459,325.
45. Montagu, J. Positioner for Optical Elements. U.S. Patent no. 4,525,030.
46. Overbeck, J. U.S. Patent no. 6,335,824.

9

Flexural Pivots for Oscillatory Scanners

David C. Brown

Cambridge Technology Inc.
Lexington, Massachusetts, USA

CONTENTS

9.1 INTRODUCTION

The memory is the knapsack of the mind. When setting out on a journey, particularly one that promises to be long and challenging, it is better to fill it with maps and recipes, fish-hooks and twine, matches and a few candles rather than tinned goods. These things are more useful, and are sure to last far longer. For those of you who cannot resist the quick fix, we hope that there are a few chocolate bars scattered about as well.

Flexures are quite ancient, and their use as pivots is also ancient. Long before the use of the most primitive bearings, leather strap flexures were used as trunk lid hinges and the like. Early war engines, for example the ballista of the Romans, the limbs of technically advanced hand bows, such as those attributed to the Turks, and the crossbows of the 14th century all employ flexures as their enabling technology.

Many pendulum clocks suspend their pendula by means of a flexure. The mechanical metronome, a specialized inverted form of clock, relies heavily on the flexure suspension in its design. It is at least arguable that the tuning fork is a pair of coupled flexures, and the music box comb is a set of flexures tuned to self-resonate at desirable frequencies.

All of these examples, and of course there are many others, exploit the flexure pivot because of its simplicity, reliability, lack of internal clearance, long service life, ease of construction, and often, its high mechanical "Q." The flexure pivots used in scientific instruments, including optical and laser scanning equipment, exploit these very same attributes.

Given the great attention devoted to flexure pivots during the 1960s and 1970s as a result of the need for rugged, reliable, light-weight, unlubricated pivots and bearings for space exploration applications, and the fact that the best and brightest were naturally attracted to what was undoubtedly the biggest science of its time, so much progress was made in flexure pivots and suspensions that one might be tempted, like Charles H. Duell, Commissioner of U.S. Office of Patents, who urged President William McKinley to abolish his office in 1899, to conclude that everything useful had already been invented.

However, as the second section of this chapter suggests, flexure technology may yet again provide one of the enabling elements in mankind's next leap of technological advancement, universal connectivity by means of light. Of course, like the flexure itself, this is an old idea, widely practised on a small scale for many centuries. The Romans communicated between the watch towers along the borders of their far-flung empire in this way, and the indigenous people of North and South America did so as well.

The organization and content of this chapter are intended to reflect the contrast between the rather settled state of macroscale flexure pivots, and the as yet entirely unexplored

space of the microelectromechanical flexure scanners (MEMS) flexure pivot. In the case of the former, the primary author has chosen to dwell largely on the details of manufacturability, which are the fruit gained from many years spent extracting the stones from the fertile soil of the garden of flexures—the practicum, as it were, of the craft. This is the sort of information that is not generally gained through textbooks, and, in the end, differentiates a possible mechanism from a practicable, not to say elegant one. In the case of the latter, the primary author has endeavored to present the theoretical underpinnings of integrated benders, as well as the fabrication and characterization of high-speed, large-angle optical scanners constructed with them.

9.1.1 Introduction to Macroscale Flexure Pivots

First, let us begin by defining what we shall mean by the term "pivot."* We must begin by modifying the term "pin," because flexure pivots sometimes have virtual pins or axles. A good working definition of pivot in this context might be "A device which defines a virtual axis of rotation, over a limited angle, while fixing the other five degrees of freedom. A device which is able to resist moments except in one axis of rotation, and may be able to resist relatively small moments in that as well." In general, these pivots are of interest in instrument quality or scientific applications, as opposed to things like lorry fifth wheel applications, where they would work, but would have no particular advantages over the more common commodity sorts of bearings, such as sleeve, ball, roller, and the like.

What then are the attributes that equip flexure pivots for instrument and scientific applications? In no particular order, they are:

1. Low mass
2. Zero clearance
3. Intrinsic restoring force
4. Infinite life
5. No lubrication
6. Low hysteresis
7. No viscous losses
8. No friction (at least none in the usual sense)
9. No pariculate generation
10. No outgassing
11. No intrinsic temperature limit
12. Extremely high load capability
13. Extremely high force linearity over small angles
14. Flexibility in design configuration, in the sense that the pivot members may be pierced to permit transparency, or even to permit the passage of solid mechanical parts or objects (in contrast with conventional bearings, which are pretty much impenetrable three-dimensional solids)
15. Low design and tooling cost
16. Very low part cost

---

* P. 978, *Chambers 20th Century Dictionary,* W&R Chambers Ltd, 43–45 Allendale street, Edinburgh EH7 4A2.

17. Predictability

18. The ability to withstand high shock and vibration loading

19. Short lead time

20. Others for very specific applications

The choice of material for a flexure is driven by conflicting requirements. For example, it is often the case that the precision of location of the pivot point is secondary to some other requirement, such as long life, low operating force, or high strength, or all the above. It may be that linearity of operating force over the desired angle of motion is also unimportant. There are often operating environment considerations. It may be that easy replacement of a damaged pivot is required. The choice of leather for trunk lid flexures results from meeting requirements such as these.

On the other hand, the designer of a bow limb has the opposite requirement that the operating force be large. Because, however, he must deal with stresses in one direction only, unlike the trunk lid flexure designer, he will exploit that difference to his advantage, and construct a flexure that is a composite. The front or tension side of his limb will be constructed from horn, which is quiet stiff in tension, while the belly, or compression side, will be constructed of sinew, which is stiff in compression. He will rightly conclude that the respective stresses in this structure are controlled by their separation from the neutral plane, and will provide a "filler" of wood properly shaped to maximize the stresses in the working "skins" of his bow without precipitating failure.

This pair of examples is designed to illustrate the broad range of potential applications of flexure pivots, and in particular the degree to which clever design is able to cope with many-order-of-magnitude differences in an application-specific parameter. Leather is able to store only a few tenths of a Joule per kilogram, whereas the best Turkish bows store over 750 Joules per kilogram, eclipsing steel and rivalling the best of the modern composites. The attribute of energy storage, related to the density, strength, and Young's modulus of the materials, may be of critical importance, required to be very high or very low, or of no importance at all in a particular application.

The absence of mechanical noise is a benefit of flexure pivots as well. Flexures are inherently low-noise pivots. Because they have no loose parts, it is unnecessary to preload them to remove clearances. As they are monolithic structures, they display very little "noise" in their force versus displacement curves. Other pivot types, comprised as they are of loose parts, make some mechanical noise during operation, and as they wear and the clearances increase, the noise level rises. Any particulates released during the wear process, of course, can cause sudden changes in the position and parasitic torque of the pivot, as well as mechanical noise, if they jam between the moving parts. Bearing "rumble" is widely recognized and even quantified as an inherent characteristic of ball bearings. In many cases, the presence of this bearing noise places an upper limit on the allowable noise bandwidth of any servo system associated with the pivot.

The possibility of a pivot design with inherently low mechanical loss is an area where flexure pivots shine. All other types of pivots known are lossy. Losses resulting from lubricant viscous friction and ball and race deformation energy not only place upper limits on the speeds at which these devices can operate, but also consume energy, limiting the mechanical efficiency of all known moving-element bearings to a low value.

Flexures, on the other hand, have no lubrication requirement, and are limited only by the internal energy of deformation. Mechanical "Q" approaching 3000 is regularly achieved, even in very high speed designs. This attribute of flexures makes possible

resonant scanners operating above 10 kHz at very low power levels. Of course, the absence of lubrication is itself a benefit in applications in optical instrumentation, spectroscopy, space research, medicine, and semiconductor processing, where even minute contamination levels are an issue.

Flexures work extremely well for small-angle applications, where stiction, traction, surface finish limitations, mechanical tolerances, lubrication distribution requirements, and so forth put demands on other pivot types which they cannot meet successfully. Flexures are, of course, free of any looseness or play, so there is no "backlash" whatsoever associated with their use; moreover, because they rely on molecular stretching, their intrinsic hysteresis over small angles near the neutral position is always less than the unbalanced forces resulting from unavoidable asymmetries in their mountings (which produce some hysteresis in realizable designs; hysteresis levels of less than 0.1% are difficult to achieve, but possible). On the other hand, unlike other pivot types, flexures are not capable of continuous rotation. While several hundred degrees of rotation seem plausible, the authors know of no application of flexure pivots designed to operate over 90°, and most designs are for much smaller angles.

Other precision pivot types, such as ball bearings and jewel ("watch") bearings are limited in their geometrical precision by the limits of accuracy that their respective manufacturing processes impose on them. For example, the best class of bearing balls, class 3, has an allowed out-of-roundness of 3×10^{-6} in. If averaged in a bearing with nine balls, the best error in "wobble" of the whole assembly will be about 9×10^{-2} less, or about 1 µin. If two such bearings 1 in apart support an axle, then the axle will have a wobble of about 2 µrad, not including errors in the concentricity of the bearing rings, axle to ring mounting, and so on. This kind of error in pivots, that is, error that is associated with geometrical features on the parts, tends to be coherent with the motion of the pivot, and so is periodic. As a result, these errors give rise, in raster scanning systems in particular, to undesirable moiré patterns. Moirés are vastly more noticeable by human observers than are nonperiodic or random errors, and can sometimes lead to rejection of a scanner that actually meets its design requirements for wobble. This effect is particularly noticeable in systems designed for printing applications, where the eye is sensitive to moirés generated by periodic angular errors in the 1/10 µrad range. Flexure pivots have none of these sources of error. Production printing engines using flexure scanners in a raster mode built at the authors' factory regularly achieve a level of periodic error below 1/10 µrad, and were developed specifically to solve the moiré problem.

There is a tendency for conventional pivot bearings to become more fragile as the level of precision required of them is increased. This is not surprising, because the accuracy of the pivot is established by careful dimensional and geometrical control of several loose parts. The single 1 µin dent in a bearing ring, which is undetectable in a 10 µrad application, becomes the limiting factor in a 1 µrad application. Flexures, on the other hand, are extremely robust, impervious to dust and other mechanical and most chemical contamination, and are very insensitive to shock and vibration. In most cases, a flexure pivot will survive indefinitely in a factory-floor or outdoor application in which other pivot types have very limited service life, if they operate at all.

Lastly, flexure pivots have inherently infinite life if operated at a peak stress level below the fatigue limit for the material from which they are made. Some care is required in eliminating any sort of stress riser in the design, and also in determining the effective fatigue limit for the material. Cambridge Technology (Lexington Massachusetts) regularly warrants its flexure products for 5 years of operation in any environment, at any duty cycle, with no statistically significant failure rate. For an 8-kHz scanner operated continuously, this amounts to more than 10^{12} cycles. (Of course, ball bearings won't operate at 8 kHz in

an oscillatory mode. However, by contrast, in applications below 1 kHz where they will operate satisfactorily, their life is generally limited to between 1 and $5 \times 10^{*}9$ cycles.)

Of course, flexure pivots are not perfect. In general, the transverse stiffness of the flexure pivot is inferior to that achievable with ball bearings, resulting in unintended cross-axis motions in the presence of significant environmental stimuli, or in gyroscopic reaction to very large accelerations of the axle. It is also more difficult to achieve very large scanning angles with flexure pivots, for the reason that the curl of the flexures at extreme angles reduces even further their transverse stiffness. The allowable stress limit at the extreme angle legislates in favor of quite thin flexures, further reducing their stiffness. While flexure scanners with scan angles up to 80° optical have been produced, their tolerance of environmental vibration and axle acceleration is low because of these effects, particularly when equipped with the very thin flexures required to obtain extremely long lifetimes.

Flexure pivots generally are designed to constrain the minimum number of degrees of freedom required by the application. Often, for example, a translation of the axle parallel to its axis is allowed, although even when the translation is not obnoxious optically, it has the potential to set up undesirable vibrations, and, in the limit, catastrophic mechanical positive feedback. With crossed-flexure pivots, it is possible to construct a geometry with zero axle shift, and practical to expect such a design to produce translation of the axis of less than a few microns over a small angle of rotation.

9.2 FLEXURE DESIGN

The possible combinations and permutations of flexures in a design approach for an arbitrary application is so vast that no attempt will be made here to cover any but the simplest form of flexure, the single leg. In the case of symmetrical designs, the easiest approach is to multiply the width of a single leg by the total number of flexures to model the flexure as a unit. In the case of asymmetric designs, other approaches to fit the circumstances will require invention. The formulas below in Section 9.2.1 assume fully reversed stress; that is, that the flexure is bent symmetrically through the same angle on both sides of the neutral position. As in the case of the Turkish bow, unreversed, or partially reversed stress provides an opportunity for clever design and perhaps nonmonolithic flexure materials. In fact, the definition of "fatigue limit" as it is used here, "That stress which represents the maximum stress which will allow the flexure to operate indefinitely without failure under a particular set of circumstances," is a hotly contested topic, and those who believe they have a suitable answer guard their secret with the most intense passion. The case of the partially reversed stress is an even more volatile subject. Obviously one of the corner stones of flexure design is an understanding of the safe upper limit of flexure stress. In the absence of reliable published data, and for the purpose of designing for a standard laboratory environment, it is the considered opinion of the authors that, for ferrous material, a value of 35% of the ultimate strength is safe under dry operating conditions.

9.2.1 Useful Formulas

The primary parameters of interest to a flexure pivot designer include the rotational spring rate of the pivot, the maximum scan angle achievable under some maximum stress level, the maximum stress applied to a flexure bent through some scan angle, the allowable

thickness for a flexure whose fatigue limit, length, Young's modulus, and angle of operation are known, and the first resonant frequency of the axle–pivot assembly.

It will be obvious that these formulas are a first-order approximation. It has been found, however, that the errors produced are small enough so that minor variations in the density of the materials used, the actual effective length and thickness of the flexures, and so on, dominate the result, and any application that requires very precise control of a particular parameter, such as the frequency of a mechanical resonator, will require some "tweaking" mechanism for final tuning on an individual basis.

These formulas also assume that the flexure mountings are infinitely rigid, support the flexure completely and uniformly, introduce no stress risers, and allow no relative motion. Except for the rigidity, these conditions are usually met satisfactorily by careful design. The first-time designer would be well advised to do a careful finite element analysis (FEA) of the mounting rigidity, or allow for an iteration or two in the design schedule. The formulas are given as follows for constant-section flat springs:

For Crossed-Flexure Pivots:

Rotational spring rate

$$K = EWT*3/12L$$

Peak mechanical angle in radians

$$A = \frac{2LS}{ET}$$

Maximum stress

$$S = \frac{ETA}{2L}$$

Thickness

$$T = \frac{2LS}{EA}$$

For Cantilever Flexure Pivots

$K = P/d = 3EI/L*3 = WT*3E/4L*3$

$S = 6PL/WT*2$

$A = d/L$

$P = Kd$

$d = AL$

$T = (6PL/WS)*1/2$

$U = 1/2d*2$

Resonant frequency

$$F = 1/2p(K/J)*1/2$$

where *E* is Young's modulus of the flexure material, *W* is the width of the flexure, *T* is the thickness of the flexure, *L* is the effective active length of the flexure, *A* is the peak scan angle of the flexure in radians, *S* is the peak stress in the flexure, *P* is the load at the tip of the cantilever, *d* is the deflection at the tip of the cantilever, and *F* is the first rotational resonant frequency of the flexure/axle system or the fundamental vibrational frequency of the cantilever/mirror system, where *J* is the combined moment of inertia of the respective system. Of course, use of consistent units is required.

There is an extensive bibliography on the subject of detailed design of flexure pivots, of which several are listed in the reference section.

9.2.2 Flexure Materials

It will be noticed immediately from inspection of the relationships among the physical constants pertinent to flexure materials that the stress is directly proportional to Young's modulus. The allowable stress is proportional to the fatigue limit of the flexure material, which most workers agree is itself proportional to the ultimate strength of the material. (Since it is now generally agreed that the so-called proportional limit, or yield strength, is not a useful concept, we shall avoid using it here.) The natural conclusion from this train of reasoning is that one could rank flexure materials by a figure of merit that is the result of dividing the material's fatigue limit by its Young's modulus. Because the exact value of the fatigue limit is unknown, or at least unpublished for the materials of interest, we shall use the ultimate strength instead, since the two are believed to be directly proportional.

Leaving for those who follow us those exotic materials whose currently available form or lack of observable malleability or other defect make their use questionable, we can construct Table 9.1.

There are some applications in which weight, or minimizing it, are paramount. In this case, specific strength, rather than ultimate strength, can be used in the construction of a figure of merit list, which might look like Table 9.2.

As one would expect, aluminum and titanium have moved up the list, but little else has changed. One could, of course, construct other lists for other specific purposes, such as minimization of inertia. Since the inertia of the flexures themselves is rarely significant construction of this list is left to the interested reader. The units of inertia are mass × radius squared, so dividing specific strength by dynamic stiffness would produce such a list (dynamic stiffness is Young's modulus divided by density squared to account for the radius squared term).

However, there are two caveats. First, most engineering enterprises are constrained by commercial goals, and so the use of exotic materials is quite properly frowned upon unless there is robust technical or other justification for their use. To the knowledge of the authors, precious metal alloys such as BeAu and spring gold are used almost exclusively in the manufacture of very high quality fountain pen nibs. In this application, they are justified because the unique combination of fatigue resistance, corrosion resistance, wear resistance, appearance, and "feel" enhance the marketability of the end product. Secondly, the fact is that very little is actually known, or at least published, about the long-term fatigue resistance of nonferrous alloys. This is partly the result of the fact that ferrous alloys have been in use for a very long time, and their overall combination of characteristics have recommended them for so many demanding applications that they are both well studied and in ready supply. One could argue that copper alloys have been in use longer, and while that is true, these alloys were generally eclipsed by their ferrous brethren as soon as they

TABLE 9.1

Parameters for Various Materials, Including Ultimate Strength

Material	Young's Modulus (psi)	Ultimate Strength (psi)	Ratio
Carbon/graphite	$<2 \times 10^6$	375×10^3	0.19
Diamond	150×10^6	7.69×10^6	0.051
Glass reinf. epoxy	5×10^6	240×10^3	0.05[a]
Silicon	27.5×10^6	1.02×10^6	0.037
Be/Au	15×10^6	210×10^3	0.014
Spring gold	15×10^6	180×10^3	0.012
Be/Cu	18×10^6	180×10^3	0.01[b]
Ti^{-6}Al^{-4}V	19×10^6	205×10^3	0.01
7075 Al	10×10^6	98×10^3	0.009
Udeholm 718	30×10^6	265×10^3	0.009
17-7PH	30×10^6	235×10^3	0.008
Inconel	31×10^6	250×10^3	0.008
302SS	28×10^6	200×10^3	0.007[c]

[a] This material's lack of electrical conductivity may limit its usefulness.

[b] Be/Cu has published figures of fatigue limit of 40 ksi.[1]

[c] Interestingly, this material not only work hardens, but its Young's modulus rises with work as well. One should, therefore, try to get material rolled as thoroughly as possible, so that the rise in modulus during use will be minimized. The range of moduli published are $24–28 \times 10^6$ psi.[1]

TABLE 9.2

Parameters for Various Materials, Including Specific Strength

Material	Young's Modulus (psi)	Specific Strength (psi)	Ratio
Carbon/graphite	$<2 \times 10^6$	3509×10^3	1.75
Glass reinf. epoxy	5×10^6	3200×10^3	0.64[a]
Silicon	27.5×10^6	12300×10^3	0.448
Diamond	150×10^6	61000×10^6	0.407
7075 Al	10×10^6	961×10^3	0.10
Ti^{-6}Al^{-4}V	18.5×10^6	1120×10^3	0.06
Be/Cu	18×10^6	557×10^3	0.03[b]
Udeholm 718	30×10^6	936×10^3	0.03
17–7PH	30×10^6	830×10^3	0.03
Be/Au	15×10^6	301×10^3	0.02
Spring gold	15×10^6	301×10^3	0.02
Inconel	31×10^6	749×10^3	0.02
302SS	28×10^6	697×10^3	0.02[c]

[a] This material's lack of electrical conductivity may limit its usefulness.

[b] Be/Cu has published figures of fatigue limit of 40 ksi.[1]

[c] Interestingly, this material not only work hardens, but its Young's modulus rises with work as well. One should, therefore, try to get material rolled as thoroughly as possible, so that the rise in modulus during use will be minimized. The range of moduli published are $24–28 \times 10^6$ psi.[1]

became available. The stainless steels have replaced bronzes where corrosion resistance is important, and BeCu and phosphor bronze for springs except for applications where magnetic susceptibility or electrical or thermal conductivity are issues.

In fact, it seems from the published literature on the subject that nonferrous metals have disappointingly low fatigue limits.[1] Experimental work at Cambridge Technology with BeCu flexure materials reinforces this view.

However, the primary reason for their use is probably simply that ferrous materials make very good flexures, and except for very unusual circumstances, the apparent technical superiority of the "other" material, if actually realizable, would require the investment of such large development or other resources that return on the investment might well be negative. As a case in point, Cambridge Technology has abandoned the "stress-proof" steels, number 8 on the list, in favor of 302SS, which is last in both Tables 9.1 and 9.2. While the manufacturer of air compressors has available to him thicknesses of material appropriate to his flapper-valve needs, no one stocks the thin foils needed for flexures. This material cannot be rerolled, and so a new thickness requirement calls for a mill run minimum at a cost of $25,000 or more and 2 years wait, and which yields a supply of material which is vastly in excess of the needs of any flexure product lifetime yet known, and must be stored carefully, inventoried, and so forth for many, many years against possible future demand, or else expensed and discarded.

Of course, not all pivots are required to last indefinitely. It is likely that an exploratory scientific mission to Mars, for example, might require only a few hundred thousand actuations. In circumstances such as these, since low weight and high reliability are much more important than life, it is probable that resources can be saved by designing a limited-life flexure. It is easy to think of other applications in which this may be true, and the author designed and flew successfully flexure systems required to operate only a few tens of cycles during atmospheric testing of atomic weapons.

Unfortunately, it appears that the borderline between indefinite life and finite life is clifflike. It is fairly easy to predict accurately life versus stress for small to moderate cycle lives, because testing to validate the model is practical. As an example, the Bendix Corporation (Utica, NY) published graphs[2] showing life versus cycles for their standard flex pivots under various conditions of loading. These graphs show load lines at 35,000 cycles, 220,000 cycles, and indefinite life. These curves cover a factor of 2 in angle, and 0%–100% loading. Interpolation is possible, but as one approaches indefinite life, small errors in actual loading and angle can cause one to fall off the cliff, with disastrous results. Testing a device to demonstrate "indefinite" life takes a very long time, and various forms of accelerated testing should be approached with caution. That said, many manufacturers of flexures have developed accelerated testing methodology. They, like Cambridge Technology, protect these methods as trade secrets. A rule of thumb, in keeping with the most widely accepted theory of fatigue failure, is that a flexure design that passes 3×10^7 cycles without failure is destined to operate indefinitely in a dry environment at this or lower level of stress.[3]

As the next section discusses, there are some hidden requirements for the qualities of the raw material form that should be understood before extravagant expense is lavished on the development of a material whose soundness of internal structure and finishability are questionable. That said, whenever an exotic, hard-to-work, expensive, material has significant technical justification, then it is worth exploring. The example of the folks who succeeded in bringing diamond phonograph styli to the mass market should be firmly held in mind.

9.2.3 Stress Risers

If the failures of flexures were ranked by cause, it would be found that, except for instances that resulted from exceeding the design stress, all failures resulted from either stress risers or corrosion or a combination. The next section deals with the germane corrosion issues. Here we will deal with the problem of stress risers, which generally arise from poor finish on the surface of the flexure foil, nicks in the edges of the flexure as manufactured, mounting defects, or inclusions in the flexure material itself, in about that order of importance. The sections on flexure manufacturing and flexure mounting deal in some detail with the avoidance of stress risers resulting respectively from manufacturing and from mounting.

A stress riser is anything that raises the stress in the flexure, either locally or globally, above the value that the designer intended that it have at that point. Because most flexures of interest to this discussion are made from rolled material, it is important to note that such materials are not really isotropic in their mechanical characteristics, the "grain" having been elongated more in one direction (the direction perpendicular to the axis of the rolls) than the other; in other words, the grain is longer in the length direction than it is in the width direction. As a result, the fatigue resistance of the material is anisotropic as well, being higher across the grain than it is parallel to the grain. It is therefore necessary to specify the direction of the grain when flexures are produced from the sheet of foil, and if the as-rolled coil is cut up into smaller sheets, the direction of rolling must be marked on each sheet in order to preserve the grain direction information. Since fatigue resistance is paramount in flexures, it is usual to specify that the flexure be cut out of the sheet in such a way as to cause the bending in operation to be across the grain.

The stress risers most prevalent in oriented flexures are the result of scratches in the surfaces of the rollers used to produce the foil. While it may be nonintuitive, the flexure stress is increased in the regions where the thickness is greater. The best quality rolled finish obtainable is none too good for flexure material, particularly when the flexure is thin. For example, a flexure 0.001 in thick, made from material with a standard mill finish of 32 μin roughness average (RA), is allowed to have scratches on both sides that are up to 0.0001 in deep. If two of them lie on top of each other, then the local thickness of the flexure is 0.0008 in. If the scratches extend to the edge of the stressed part of the flexure, the 20% reduction in area is very likely to act like a notch and induce a crack. On the other hand, if the *rolls* which made the foil have scratches, the foil will have inverted scratches, that is raised areas which have the effect of increasing the thickness locally. Using the same exemplar dimensions, because the stress is proportional to the cube of the thickness, the local stress will exceed the designer's intent by 95%, and is likely to cause failure. Of course, an inclusion of microscopic size will produce the same effect.

Cambridge Technology specifies 4RA or better finish for its rolled flexure strip. For critical applications, it is well to have the rolled sheet 100% X-rayed to find any inclusions or cracks or voids or other defects inside the material.

The last important item is not associated with the quality of the raw material, but is one of the most easily overlooked elements of a stress riser free design. Nearly every flexure has one or more radical changes in width, usually associated with attachment. It is absolutely critical that these section changes are made with a stress-distributing fillet or radius in the corner. Otherwise, failure cracks will inevitably originate at these locations.

Figure 9.1 is a drawing of one of the flexure assemblies in current production at Cambridge Technology. This part is actually two flexures joined together by a bar (the horizontal part along datum "B"). Making a pair of flexures this way ensures the utmost in symmetry, since they are made at the same time from nearby regions of the same piece of metal. Also,

Notes:
1. Material: stainless steel, type 301 or 302, full hard condition.
   surface quality 4RA (rahns 2n, photoetch quality).
2. Finish: Passivate, per QQ–P–35.
3. Photoetch attachment points along datum B only
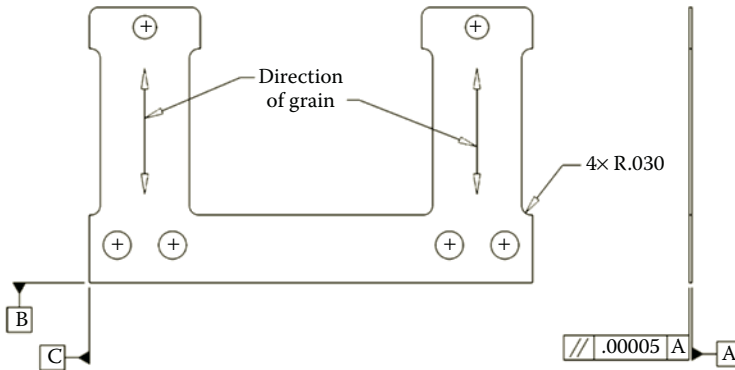
FIGURE 9.1
Typical flexure.

the "self-fixturing" feature of the bar eliminates the necessity for adjusting their parallelism and individual effective lengths during assembly.

Attention should be paid to the specification legend in the upper left-hand corner, as well as to the grain-defining arrows and the corner radii at every section change.

9.2.4 Corrosion

It is assumed here that the flexure designer will take into account the environment in which the equipment will live, and take suitable precautions against atmospheric or environment corrosion. We will discuss here only the forms of corrosion that result from the highly stressed conditions of the flexures during use, the effects of galvanic couples, and the hydrogen embrittlement associated with electroplating.

Corrosion in flexures is intimately connected with stress risers, because most of the corrosion effects encountered begin in cracks and crevices either caused by stress cracking or enhancing to stress cracking. However, there is one area of corrosion difficulty encountered that is not directly associated with stress, but is instead a form of electrolytic corrosion resulting from the contact between metals far enough apart on the electromotive series. This effect is not significant in very dry conditions, but the presence of an electrolyte can cause rapid erosion of the anodic member of the couple. Even atmospheric water held in cracks by capillary action can be a powerful electrolyte under conditions of stress. It is up to the designer to decide which member of the couple he would prefer to be dissolved; in general, it is preferable to protect both to the extent possible. Theoretic galvanic couples are presented in Table 9.3.

In general, MIL-STD-186 allows adjacent groups, or, in some cases, metals two groups away from each other to be coupled. This does not mean that galvanic action does not take place. It just means that for most purposes the rate of corrosion is so slow as to be unimportant.

At Cambridge Technology, standard product joins group 5 flexures with group 14 mounting materials using group 13 fasteners with satisfactory results. Because the aluminum is

TABLE 9.3

Theoretical Galvanic Couples

Group	Material	EMF with Respect to Calomel in Sea Water
1	Gold, platinum	+0.15 V most cathodic
2	Rhodium, graphite	+0.05
3	Silver	0.00
4	Nickel, monel, titanium	−0.15
5	Copper, Ni-chrome, austenitic SS	−0.20
6	Yellow brasses and bronzes	−0.05
7	High brasses and bronzes	−0.30
8	18% chromium SS	−0.35
9	12% chromium SS, chromium	−0.45
10	Tin, tin–lead solders	−0.50
11	Lead	−0.55
12	2000 series aluminum	−0.60
13	Iron and alloy steels	−0.70
14	Wrought aluminum other than 2000	−0.75
15	Nonsilicon cast aluminum	−0.80
16	Galvanized steel	−1.05
17	Zinc	−1.10
18	Magnesium	−1.60 most anodic

quite anodic to the flexures, it is slowly dissolved, but the more sensitive flexures are unaffected. For demanding applications, all the ferrous parts are tin plated before assembly, and the aluminum is anodized.

In this context, it is useful to discuss a side effect of electroplating of high-strength steel, the so-called hydrogen embrittlement. During electroplating, the electrolyte, if aqueous, contains hydrogen ions, which are hydrogen atoms stripped of their electrons. These very small ions are able to squeeze into the grain boundary lattice structure of the steel, where they can produce pressures approaching 13,000 atmospheres. Under conditions of stress, the combined forces of the hydrogen and the external stress rupture the metal. It is one of the cruel jokes of nature that the very high strength steels such as those used for flexures are severely afflicted by this process, and that the heat treatment required to drive out the hydrogen severely limits the strength that can be obtained in precipitation-hardened stainless steels such as 17–4PH. This is one of the reasons for choosing an age-hardening grade of steel such as 302 for flexures that might require electoprocessing.

The theory of stress-corrosion cracking favored by the authors is the electrochemical theory.[4] According to electrochemical theory, galvanic cells are set up between metal grains and anodic paths are established by heterogeneous phases. For example, the precipitation of $CuAl_2$ from an Al–4Cu alloy along grain boundaries produces copper-depleted paths in the edges of the grains. When the alloy, stressed in tension, is exposed to a corrosive environment, the ensuing localized electrochemical dissolution of the metal, combined with plastic deformation, opens up a crack in the metal. Supporting this theory is the existence of a measurable potential in the metal at grain boundaries, which is negative with respect to the potential of the grains. Once the crack exists, capillary forces aid the delivery of electrolyte to the tip of the crack, and it propagates. Figure 9.2 is an illustration of this concept.
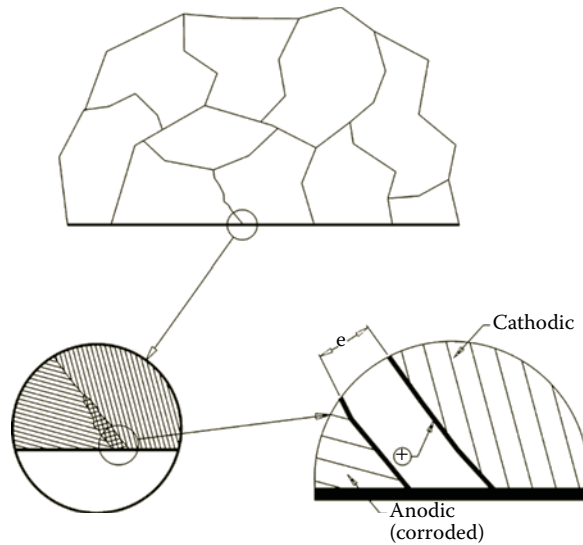
FIGURE 9.2
Stress-corrosion cracking.

9.3 FLEXURE MANUFACTURING

9.3.1 Manufacturing the Material

Flexures have been manufactured by every conceivable process, but for the purposes of small precision pivots, carefully rolled sheet or foil produces the best densification, uniformity of metallurgy, and surface finish. In addition, many work-hardenable materials come off the rolls with the proper temper for use. It cannot be overemphasized that surface finish is critical to the performance of thin flexures. No process now known is able to improve over the uniformity of thickness, flatness, freedom from scratches, and the consistency of temper of ferrous materials precision rolled on machines specially designed and constructed for the purpose. However tempting it may be to obtain experimental stock by grinding and polishing or etching down existing material, whether chemically or in combination with ion impact, these processes will change the performance values of the material. Unless one is prepared to make all the required flexures by an identical process stream, likely extremely uneconomical, the experimental flexures should be made from a material process exactly as the production flexures will be processed. In fact, considering that 100 pounds of flexure material such as type 302 stainless can be obtained as a minimum run from a reputable reroller for under $3000, and 100 lb of foil will make many tens of thousands of flexures, there is hardly any reason not to make the experimental parts from the production material, and there are several reasons why this is the best approach.

First, the process of arriving at the desired thickness, for the sake of argument 0.00500 in thick ± 0.000020 in, at the required as-rolled temper, usually requires that the reroller start with carefully annealed material whose thickness is the nearest standard thickness above ten times the finished thickness, in this case, probably 1/16 in. Even so, the finished temper of the material will vary by a few percent run-to-run, the thickness will vary by 10 to

20 millionths run-to-run, and the cross-sheet crown will also vary by a few millionths run-to-run. As a result, even if the edges of the sheet are trimmed off the material used for flexures, which is good practice, for the best uniformity flexure-to-flexure, it is best to buy all the material needed for the product lifetime in one mill run. If this is not possible, or if product demand exceeds all expectations, it is best to requalify each succeeding flexure material run, and to make adjustments in the flexure width to compensate for any variations in modulus of elasticity or spring constant discovered. The stiffness of the flexure is linear with width, and a new photo tool is only a few hundred dollars, and so adjusting the width is a reliable and economical way to titrate the flexure stiffness.

It is also worth the expense, for critical applications at least, to 100% X-ray the material to find any inclusions, cracks, or voids.

9.3.2 Cutting Out the Flexures

Flexures may be cut out of sheet or foil by punching, conventional machining of stacked blanks, laser, water jet, or E-beam cutting, or chemical etching of photolithographically produced patterns in photo resist. The last process is the standard process at Cambridge Technology, and is that preferred by most workers under most circumstances, because it is universally applicable, widely available, accurate to microinch tolerances, repeatable, fast, localized to the area to be cut, the photoresist protects the sensitive surfaces, and the process is inexpensive.

The other methods cause, or can cause, performance variation as the result of modified metallurgy in the zone near the cut. These potential modifications include heat effects, work hardening, and changes in the composition or phase of the alloy. Changes in the thickness of the flexure locally, and the introduction of stress risers, invariably accompany the relatively large-scale, relatively uncontrolled motions of mechanical machining and punching, including fine blanking, and are also associated with jets of abrasive or abrasive in water, laser cutting, and E-beam cutting.

It is also difficult to provide adequate surface protection during processing by these methods.

Even photoetching produces artifacts that have the potential for undesirable effects. For example, the edge pattern left by double-sided etching is a double cusp that is irregular, sharp, and full of stress risers. Figure 9.3 shows such an edge.

Also, the position of the attachment points between the flexures to be and the parent sheet should be specified in such a place, usually on an unstressed tie bar or mounting tab, that the stresses associated with their disconnect are not obnoxious. The best way to remove the cusps and their associated stress risers is by tumbling in appropriate media, chosen to batter rather than to scratch, so that the cusps are beaten down, but the surfaces are not scratched. The edges of the flexures should be 100% inspected under suitable magnification, and any flexure with a nick or scratch visible at 20× should be rejected.

9.3.3 Corrosion Protection

Most of the materials usually chosen for flexure material have intrinsic resistance to laboratory or office environments which is adequate protection. However, in shop floor, marine, outdoor, or specialized applications, further protection is often desirable. Such protections include sealing and purging the entire mechanism. When this approach is not practical, then the flexures themselves (and, if necessary, their mated parts) can be protected. Of course, with the flexures in particular, the protection should not interfere any more

Notes:

1. Photoetch attachment points along datum B only.
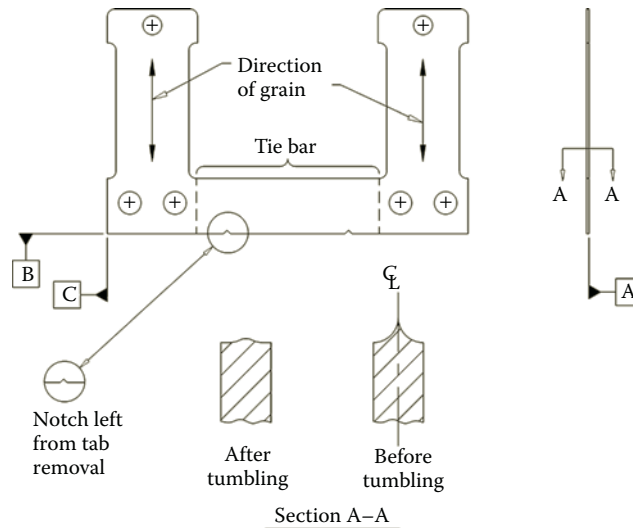
FIGURE 9.3
Edge stress risers.

than necessary with the underlying functional performance of the part. At Cambridge Technology, hot-oil reflowed tin plating of the ferrous parts, including clamps and screws, is the protection method used. Tin has excellent protection in very thin layers, is anodic to the type 302 flexure material, has a very small Young's modulus, is a lubricating film, helps to distribute the clamping forces uniformly, and has low mass. It is, however, cathodic to 5, 6, and 7000 series aluminum, as well as type 355 and 356 cast aluminum. It is, of course, extremely cathodic to magnesium and its alloys. The answer here is anodizing of any aluminum or magnesium mated parts, brought to completion, and sealed with a suitable material such as sodium silicate. Although MIL-STD-186 allows tin in contact with magnesium, the best practice is anodization of the magnesium, or, in the case of fairly porous castings, vacuum impregnation of the casting with epoxy or equivalent followed by anodization.

9.4 FLEXURE MOUNTING

The careful mounting of flexures is essential if the full potential of careful flexure design is to be reached. There are three aspects of mounting that deserve thought. The first is the transition from the flexure itself to the mounting provision for that flexure, whatever it may be. For example, flexures may have integral mounting tabs intended to be supported with reenforcing loose clamps, clamps attached by adhesive or solder to the flexure tab, flexures without tabs intended to be clamped or soldered or cemented into a slot, flexures with or without tabs intended to be welded in place, and so on. Whatever the method used, it is essential that the distribution of stress from the highly stressed operating region to the

mount is gradual, controlled, and without stress risers. In general, neither the mounting area nor the body of the mount can be made to carry the concentrated stress loads that the flexure itself carries. Typically, the flexure is flared out in the region of the mounting so that the stressed region is increased in cross-sectional area, and the intensity of the stress is reduced thereby to a level low enough to prevent fracture of the mounting region before the main flexure. While any of the above methods may be used, the standard practice at Cambridge Technology is to produce a flared tab design similar to that in Figures 9.1 and 9.2. The corner radii are essential to achieving a transition without a stress raiser in the corner.

Of course, the presence of a radius in the corner raises the question of just where the effective region ends, and where the mounting region begins. It has been found by experiment that if the effective length of the flexure is taken to be the distance between the points of tangency between the constant-width section of the flexure and the radii at the ends, plus one radius, the error in the result will be dominated by tolerances on the flexure thickness.

Secondly, alignment of the flexure(s) to the stationary and moving members must be thought out. It is often the case that a multiplicity of flexure pairs or sets positioned along the axis of rotation is required in order to resist the cross-axis moments of the system. Thus, as many as four or six individual flexures may require individual alignment in a typical system. In this case, it is convenient to manufacture the flexures attached together by means of a tie bar, which may either be left attached or removed after assembly. This reduces the number of alignments required by at least two. Such a tie bar is illustrated in Figure 9.3.

Thirdly, the designer must decide whether or not to try to make the flexure mountings on the fixed or on the moving or both elements indefinitely stiff. If he succeeds, then the design formulas given earlier will produce predicable results. If not, then some iteration may be required before satisfactory performance is obtained.

However, before leaping to a conclusion here, the designer should take into consideration the fact that the economical materials of choice for the rotor and/or stator are often not high fatigue strength materials, and may not even be highly qualified materials and so may have relatively loose, or unknown specifications of physical parameters such as Young's modulus and strength. In addition, it may well be that some other parameter is of more importance to their overall function than fatigue strength. One might posit thermal conductivity for the stator, and low inertia for the rotor. Whatever the reason, it is often desirable to allow the flexure-mated parts to be made of less intrinsically stiff material than is the material of the flexure itself. One then has the option of increasing the cross-sectional area of the mounting region appropriately. However, if the reason for using, say, an aluminum–magnesium alloy for the rotor had to do with inertial minimization, then bulking up the extreme-from-the-axis parts (where the flexure ends inevitably attach) to increase their stiffness is not likely to find favor. It may, in fact, be a better overall solution to allow the mountings to bend a little under load, and recalculate the flexures, if necessary, to compensate. It should be recognized that some deflection of the flexure mountings is required anyway, because if they are as stiff as or stiffer than the flexures, the clamping loads will be at least partially transferred to the flexures, potentially overloading them at the worst possible location, the mounting transition region. In this context, it is well to pay as much attention to the flatness and surface finish of the flexure-contact pads as to that of the flexures themselves. For this reason, slots are discouraged because good surface finish on the sides of slots is difficult to achieve, and inspecting them is next to impossible. A carefully machined and lapped

pad, with a loose, lapped clamping plate under the screw heads to distribute their local loading is much better, and less costly in the long run. A layer of tin or indium foil or adhesive between the flexure and its mating surface on each side is advised as a method of filling microscopic voids and the valleys between the peaks on the surfaces, as well as providing corrosion protection at the interfacial joints. Care in mounting flexures cannot be overemphasized, particularly when attempts have been made to make the mating structures extremely stiff.

9.5 CROSSED-AXIS FLEXURE PIVOTS

9.5.1 General Introduction

Of course, there are many uses for flexures, and many flexure types. Flexure hinges, previously mentioned, are perhaps the most ancient. Straight-line motion mechanisms, arguably the most difficult bearing types to design, have reaped great benefit from flexure technology. The author spent 20 years designing and building bearings for scanning Michaelson interferometers, including diaphragm-flexure, "porch-swing" parallel-flexure, Bendix pivot porch-swing, and others. Because these straight-line motions are not generally considered pivots, they will not be discussed in detail here. The reader interested primarily in straight-line motions is recommended to References 4 and 5. Flexures of the torsional type are discussed in detail elsewhere in this book.

There is, however, a very interesting flexure pivot type whose design is due to the late Niels Young, and is illustrated in Figure 9.4. This pivot began life as a straight-line mechanism for a scanning Michaelson interferometer. The diaphragms, instead of being corrugated or plane, were pierced by a number of quasi-radial curved slots. These slots increased the axial compliance of the diaphragm considerably without materially decreasing the
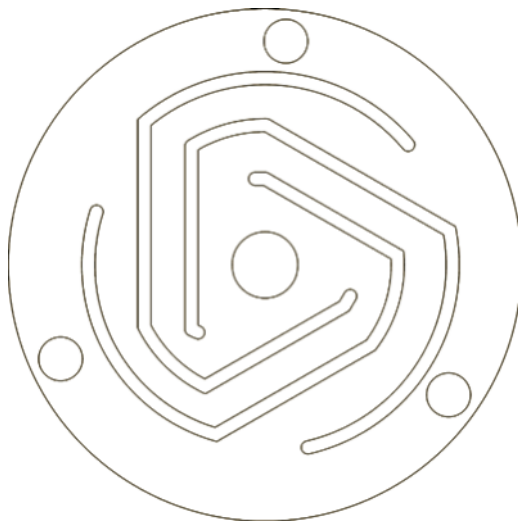
FIGURE 9.4
Support diaphragm for straight-line mechanism.

radial stiffness, or introducing a departure from the straight-line motion. It was discovered, however, that if two of these diaphragms were mounted to their separating pillar with the slots aligned, a small rotation of the pillar accompanied translation. This was, of course, undesirable in the application, so the diaphragms were mounted reversed with respect to each other, and the rotation was restrained thereby.

Somewhat later the author having need of a very tiny precisely adjustable rotation in order to align an optical component, recalled the "defect" of the Young flexure, seized upon it as a virtue, and succeeded in putting it to good use.

Since the 1960s it has been generally recognized that the crossed-axis flexural pivot has the most widely applicable characteristics of any single-flexure pivot type. This wide adaptability was so compelling that the Bendix Corporation designed and manufactured a family of self-contained flexure pivots in various materials and sizes, and was so successful in its implementation that this type of pivot has come to be known worldwide as the Bendix pivot.

9.5.2 The Bendix Pivot

The Bendix "free-flex" pivot was introduced to the world in November 1962 in a seminal paper in Automatic Control magazine, entitled "Considerations in the Application of Flexural Pivots."[6]

Available still from Lucas Aerospace, these high-quality, standardized, well-quantified pivots should be considered whenever there is space to include them, and the product envisioned will be produced in modest quantities. Much time and effort, particularly in qualification testing, can be saved by using these devices. Made in both single-ended (cantilever) and double-ended types, Figure 9.5 shows the general principle of construction.[7]

This general form of layout, the 90° symmetrical cross, has become the standard of construction, and the departure point for many specialized designs. However, this layout has the defect that translation of the axis of rotation takes place during angular motion of the pivot. This motion is neither linear nor small, and while many optical applications

Undeflected   Deflected
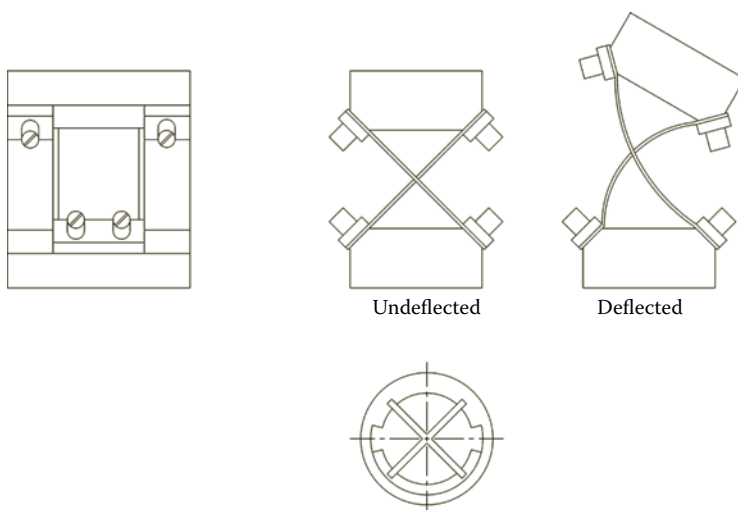
FIGURE 9.5
The Bendix pivot.

are quite tolerant of translation perpendicular to the plane of a mirror, the speed of the system may be limited by vibration produced by the translation of significant mass. As mentioned in the introduction to this chapter, it is possible to design a zero translation crossed-axis flexure pivot, and in principle possible to construct one, although the precision of assembly required is very high. The amount of translation, and the shape of the angle versus translation curve is dominated by the ratio of the leg lengths of the flexure. In the Bendix design, the leg lengths are equal, and the point of crossing, which lies on the axis of rotation, is initially central to the mounting tube. As the pivot is rotated, the axis of rotation departs from the center along a curve whose cusp is central and symmetric, but whose shape depends on the ratio of lengths of the flexure arms at intersection. Since the axis of rotation is usually a line fixed in space with respect to the rest of the mechanism, the axle is constrained by its load, and the flexures must cross on this axis, the only convenient way to change the ratio of the flexure arms is to extend the arm from the axis toward the stator (leaving the arm length between the axis and the rotor attachment point fixed) until the desired arm ratio is obtained. This inevitably has the effect of increasing the diameter of the stator flexure attachment point with respect to the axis of rotation. Since the ratio of arm lengths required for theoretical zero translation is 12.5% to 87.5%, this can become a big deal, so most designs find a compromise that is workable.

9.5.3 Cambridge Technology Crossed-Flexure Design Example

Figures 9.6 through 9.9 show the successive stages of assembly of a flexure-pivot optical scanner mechanism, which represents the state of the art at Cambridge Technology.

This scanner, designed in 1995 and still in production, is used in a high-quality large-format printing engine to produce multicolor magazine illustrations. It produces 30° optical scans of an elliptical 30-mm aperture mirror at 160 Hz, with line straightness of a few
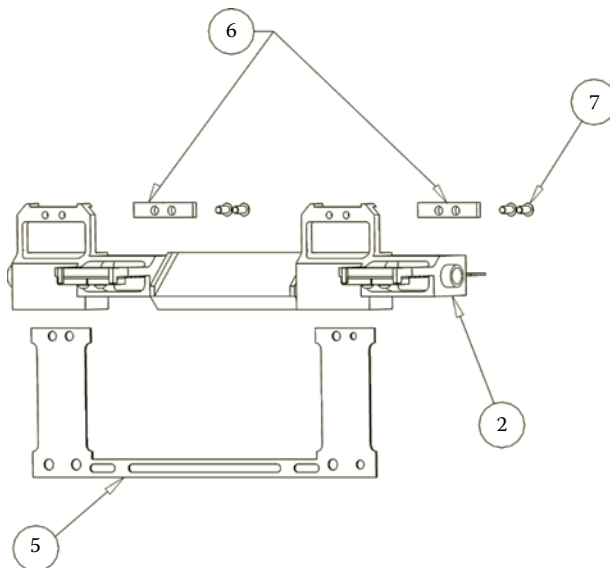
FIGURE 9.6
Rotor-flexure assembly.
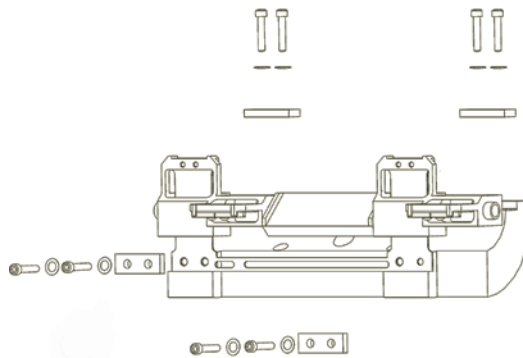
FIGURE 9.7
Rotor-stator assembly.

microradians. Every feature of design and construction highlighted in this text as preferred is illustrated in the design. Of the several thousand of these scanners delivered, of which many have more than 20,000 h of operation at 160 Hz (4×10^9 cycles), none has ever been returned for any reason.

Figure 9.6 shows the assembly of the flexures to the rotor. Note the radii in the mounting transition region of the flexures, the tie bar 5, and the lapped loose clamps 6 under the screw heads. These flexures are photo-etched, tumbled, 100% edge inspected under 20× magnification and tin plated. Figure 9.7 shows the assembly of the rotor and its flexures into the stator. There are registration notches on the mounting pads on both the stator and the rotor, which position the ends of the flexures and their clamps, assuring equal effective length of the flexures and parallelism between the axis of the rotor and the axis of the stator. Notice that the flexure mounting pads are islands, which can be machined, lapped, and inspected easily. Once assembled to the stator sector, the entire pivot mechanism is complete, and can be inspected and tested easily without unnecessary obstruction. Once qualified, it is mounted into the final housing, such as the one shown in Figures 9.8 and 9.9.

An interesting variation of this design is illustrated in Reference 8. This scanner was required to operate in a mechanism that was quite sensitive to microscopic levels of vibration and other mechanical noise. Since crossed-axis pivots operated at high speeds produce a periodic translation of the axle, they have the potential to transmit vibration to the rest of the mechanism through their mounting means. In this case, the stator–rotor assembly, instead of being bolted directly to its housing, is supported on a set of flexures that permit an additional small rotation between the stator and the mounting. These flexures allow the pivot assembly to oscillate torsionally and transversely with respect to the housing. The second flexures therefore isolate the pivot assembly from the housing, and allow the housing to be bolted to the optical system without imparting an undesirable level of vibration in it. The second set of flexures allow the stator to rotate in response to the reaction torque from the rotor. The rotor and the stator counterrotate, and the relative amplitude of the angular oscillations of the rotor and the stator is approximately in inverse proportion to their respective inertias. A typical ratio of inertias for the rotor and the stator is 1:150, so that the angular deflection of the second set of flexures is approximately 1/150 as great as that of the rotor flexures. The second set of flexures also allows the stator to translate in response to the oscillatory translations of the rotor's center of
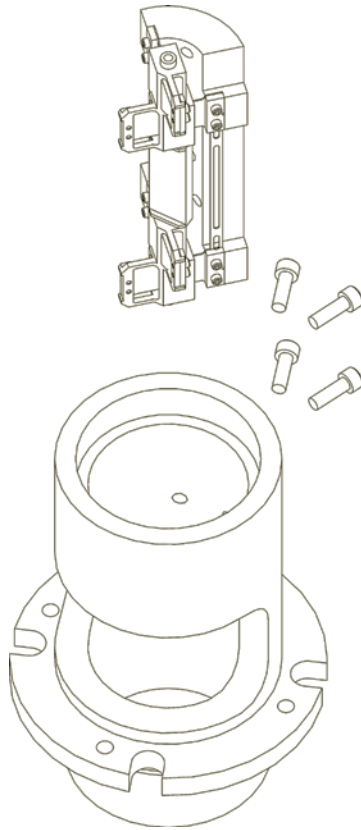
FIGURE 9.8
Stator-housing assembly, exploded view.

gravity, which result from the translation of the rotor as it rotates. The forces required to dynamically accelerate the rotor and the stator largely balance each other. The amplitude of the residual translation force, compared with the translation force of the same pivot assembly bolted directly to the housing, is given approximately by the ratio of the pivot assembly mass to the housing mass, typically 1:15 or greater. One could, of course, apply ever more stages of isolation by this means of flexure pivots within flexure pivots if one had the need to do so.

9.6 LOW-COST CANTILEVER SCANNER

The discussion of the previous section was devoted to the details of construction of a typical very high performance scanner where the unique attributes of the flexure pivot are devoted to solving issues unmanageable by any other known bearing type. On the opposite end of the spectrum of suitable applications, flexure pivots lend themselves to the construction of very low-cost scanners which yet are rugged, long-lived, and precise.
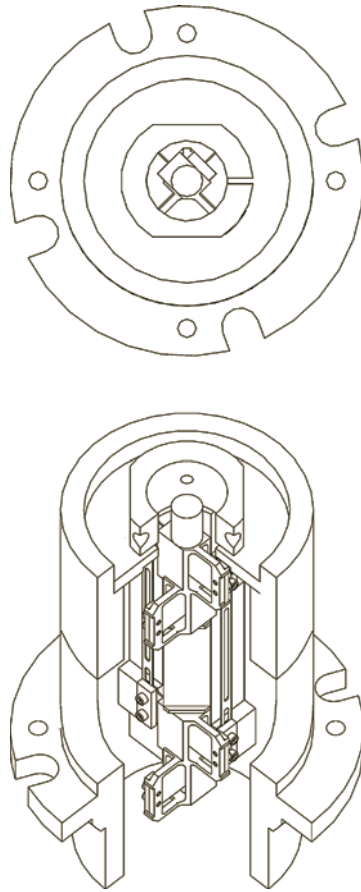
FIGURE 9.9
Final scanner assembly.

9.6.1 General Features

Figure 9.10 shows a cantilever flexure scanner designed for a digital radiography application. It operates at 50 Hz +/− 1 Hz over a peak-to-peak optical scan angle of 40°. The mirror aperture is designed to accommodate a 12-mm diameter beam at a nominal angle of incidence of 45°. Figure 9.10 shows a plan and an elevation view of the scanner. Figure 9.11 shows a picture of the flexure itself. Figure 9.12 shows an exploded view, where 1 is the flexure, 2 is the mirror, 3 is the flexure clamp, 4 is the flexure support, 5 is a counter mass, and 6 is the drive magnet. Referring now to Figure 9.10, the elevation view has an arrow which indicates the motion of the mirror. Cantilever scanners pivot the mirror a long way from the center, and so the mirror "walks" on the beam during the scan; the cross-axis dimension of the mirror is long. The elastic curve of the flexure depends on the way the drive is applied. In this particular design, the drive is by means of a small magnet attached under the center of the mirror at the tip of the effective length of the flexure. The magnet dips into a pair of coils in the stator block, wound at 90° to each other and mutually at 45° to the axis of the magnet. One of the coils provides the drive flux, and the other serves as a velocity coil for feed back. (This coil construction is intended to minimize the transformer coupling between the drive and feed back coils.) This kind of drive coupling
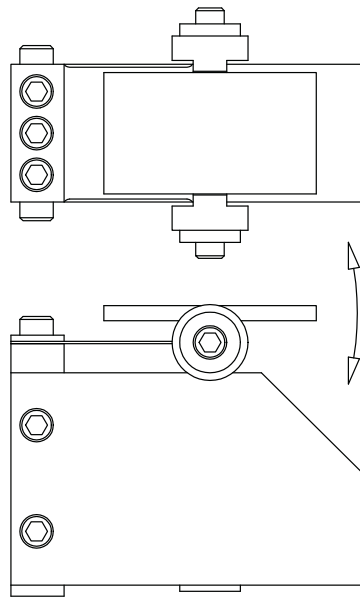
FIGURE 9.10
Low-cost cantilever scanner.
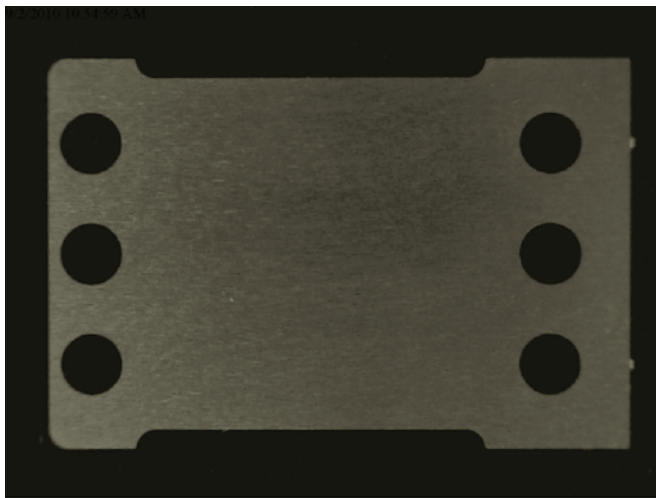
FIGURE 9.11
Flexure.

is a combination of point loading at the tip and loading as a moment around the tip. As a result, the slope of the tip of the flexure, and therefore the slope of the mirror surface does not obey either the $-PL^{*}2/2EI$ or the $-ML/EI$ textbook expressions, but instead a combination, which is sensitive to the exact relationship between the magnet and the back iron which makes prediction of the slope uncertain. Of course, other more predictable drive
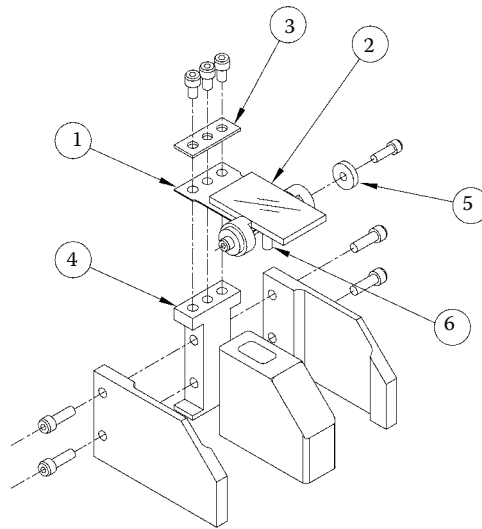
FIGURE 9.12
Cantilever scanner, exploded view. 1. Flexure, 2. Mirror, 3. Clamp, 4. Mounting post, 5. Tuning weight, 6. Drive magnet.

coupling methods are possible if prediction of the slope is important to the design. In any case, the fact that the flexure assumes a curve has the effect of reducing the magnet deflection required for a particular scan angle compared to that of a simple hinge, and this means that the peak stress, which occurs at the clamped end of the flexure, is reduced. It also means that, while the motion of the mass-spring is sinusoidal, the motion of the deflected beam is partially linearized. In fact one could depart from the simple constant-section rectangular flexure in favor of a triangular flexure, or a rectangular flexure tapered in thickness in the form of a cubical parabola, both of which have as their elastic curve a circular arc, and more completely linearize the velocity of the reflected beam. Both of these variations have complexities which increase the cost, and so are inappropriate for a design which is intended, as this one is, to be low cost.

9.6.2 Design example

It is useful to begin with the inertia of the moving parts, since in general the size of the mirror and the scan angle are defined *a priori.* The essential moving parts are the mirror and the drive coupling, in this case a magnet. The inertia of the flexure may usually be ignored; however, including one-third of the flexure inertia produces a rigorous result. There are usually ancillaries to the required rigid connection between the flexure, the mirror, and the drive which add inertia, in this case an aluminum spool, a flexure clamp (not shown, but essentially identical to part 3), and three small screws. In addition, this design includes provision for tuning weights. This is not essential, because conventional dimensional control of the flexure and moving parts, once the desired frequency is established empirically, is adequate to assure high yield in production even with fairly tight frequency specification (+/–2% in this case). However, the provision of tuning weights means that the empirical frequency "tweak" is straightforward, and furthermore, the weights permit the

basic scanner to be configured for a variety of other mirror sizes and resonant frequencies simply by changing the weights at final assembly.

9.6.3 Motor Size Required

Assuming a conservative mechanical "Q" of 1000, the motor will need to deliver about 1/1000 of the stored energy per cycle. The stored energy is given by $U = fd/2$. As pointed out above, we don't know exactly what the value of the displacement will be, but a conservative estimate is given by ignoring the fact that the flexure will bend in an arc, and assume a deflection given by the multiplying the length of the active part of the flexure times the peak scan angle in radians. Let us assume a flexure length of 0.5 in, a spring rate of 6 lb/in, and a peak mechanical angle of 10° or 10/57.3 rad. The displacement is then 0.5 × 10/57.3 = 0.087 in, the force is 0.522 lb, and U = 8.3 × 10^{-2} in lb. The minimum motor size required then will be such as to supply 8.3 × 10^{-2}/1000 = 8.33 × 10^{-5} in lb to the drive. To put this all in perspective, the power of the minimum-size motor in this case is 8.33 × 10^{-5} × 50 = 4.2 × 10^{-3} in lb/s or 1.3 × 10^{-7} Watts. Inevitably there are $I*2R$ losses and windage losses, and the scan angle will take some time to build up with a minimum-size motor, so it is usual to make the motor somewhat oversize. That said, it is evident that such low-power scanners make possible long-life battery powered equipment.

9.7 VIBRATING-WIRE SCANNER

Figure 9.19 shows the general lay out of a vibrating wire scanner. Using the expression $F = \sqrt{T/LW}$, well known to string musicians, and dating from around the year 1300 we can design a scanner to use two strings vibrating 180° out of phase to tip a mirror.

F = Desired frequency of vibration
T = wire tension
L = length of wires
W = mass of moving parts

The spacing between the wires is arbitrary. The mirror is fastened to the midpoint of each wire as shown, and the drive is attached to the underside of the mirror between the wires. In general, the weight of the wires may be ignored, since the mirror and drive are the major source of mass, but again, adding one-third of the weight of the wires is rigorous. Both ends of the pair of wires are clamped, and some means of adjusting the tension of the wires must be provided. Rearranging the expression, $T = LF*2W$ in appropriate units. The drive may be arranged as with the cantilever scanner, and all the "Q" and power considerations described in that section apply. This scanner, with suspension support at both ends, is more resistant to vibration than the cantilever scanner, and is suitable for a wide range of operating frequencies, just like the Piano.

9.8 MICROELECTROMECHANICAL FLEXURE SCANNERS

Microelectromechanical scanners (MEMS) integrate flexures and electromechanical actuators to yield small scanners of high performance and low cost. These are interesting in a number of applications but none more than optical switching in telephony.

The scanning mirror may be formed direct onto the MEMS carrier, or it may be a separate component bonded to the mechanical assembly. Its aperture is typically between 0.1 × 0.1 mm and 3 × 3 mm. Currently, optical scan angles up to 20° are practical. Resonant frequencies are typically in the range 10–40 kHz and are dependent on system design, mirror size, and maximum scan angle.

There is a huge body of expertise in the manufacture of precise, small structures from silicon or on silicon substrates. It is fortuitous that silicon possesses attractive properties for small flexures, so we can exploit the existing semiconductor manufacturing infrastructure to make MEMS. This same manufacturing technology is well suited to the fabrication of the piezoelectric actuators.

9.8.1 MEMS Design

Piezoelectric bimorphs consist of two strips of piezoelectric elements joined over their long surfaces, provided with electrodes in such a manner that when an electric field is applied, one strip elongates and the other contracts. This results in a bending motion of both strips. The motion of the tip can be considerable, which is what makes these devices such useful tools as converters of electrical energy to mechanical energy and vice versa. They are used in different applications such as ultrasonic motors,[9] laser beam detectors,[10,11] fans for cooling electronics,[11] numeric displays,[12] filters,[13] accelerometers,[14–16] optical choppers,[17] and more recently as the legs of microrobots.[18] They are suitable as converters of electrical signals into sound (speakers) and similarly as pickup elements for the detection of sound.[19] They are also used as the control element to reduce vibration in space-borne structures such as solar panels and in the walls of offices for the reduction of sound transmission.[20] While there is a huge variety of concepts that could be realized, we confine this discussion to a relatively simple layout as represented in Figure 9.13.

The two piezoelectric actuators are of a form that has been used for the last 70 years or more.[9] This "bender," as we shall refer to it, has various names in the literature, including "piezoelectric bimorph," a term not strictly accurate in this case. It comprises two bars of
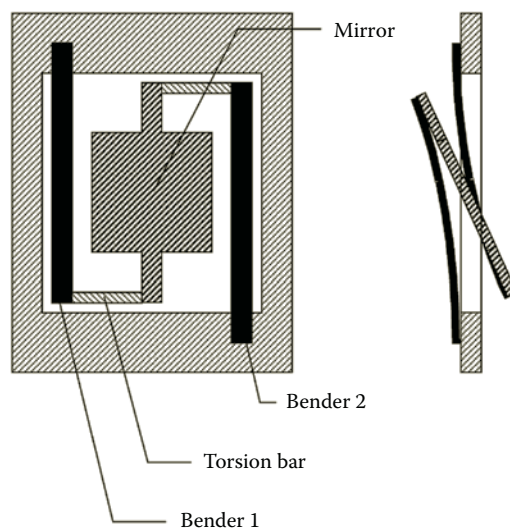
FIGURE 9.13
Simple piezoelectric "bender" scanner.

piezoelectric material, typically PZT, having opposite polarities and sandwiched between two electrodes as indicated in Figure 9.14. A voltage applied to the electrodes results in a field that causes the length of one bar to increase and that of the other to reduce. This produces curvature of the structure similar to the bending of a bimetal structure as a result of a temperature change. In our case, the benders are, strictly speaking, monomorphs; that is, there is only one active layer of piezoelectric material.

The free ends of the two benders are connected by torsion bars to the mirror substrate. The benders are driven by equal and opposite polarity so their motion is equal and opposite. The mirror substrate rotates about a central axis. This geometry has been studied in depth by Smits.

When using PZT material, an electric field is applied across two plates separated by poled ferroelectric material. PZT ceramics are isotropic and not piezoelectric prior to poling. Upon poling, they become anisotropic and display directionally dependent piezoelectric and mechanical properties. Poling is accomplished by heating the material to its Curie temperature and applying a DC field to the crystal, which then aligns its previously randomly oriented dipoles parallel to the field. Upon cooling, the dipoles maintain this preferred arrangement. As a result, there is a crystal distortion that causes growth in the dimension parallel with the field and also in the dimension perpendicular to the field. The axial strain resulting is typically small (0.2%) and is accompanied by hysteresis.

The displacement of the PZT material in the actuators when an electric field is supplied is the source of the bending moment of the "J" arms connected to the mirror platforms. When the electric field is applied, the material has displacement in two directions: parallel to the field and at a right angle to the field. The parallel and perpendicular displacements are opposite in sign so that when the film expands in parallel with the field it contracts its dimension perpendicular to the field and vice versa. The polarity of the parallel displacement is determined by the direction of the electric field with respect to the domains of the material. (The directions of the domains were established during the polarization process.) If the polarization and electric field are opposed, that is, the domains are in the opposite direction to the applied electric field, the material will expand parallel to the field and shrink perpendicular to the field. Opposed domains are stable in the sense that they are already arranged to cancel the applied electric field. If the domains and electric field are parallel, then the material will shrink as the electric field gets stronger until 50% of the domains have switched their direction to be opposed. This is effectively repolarizing the
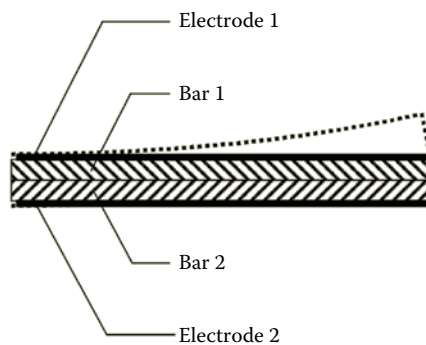
FIGURE 9.14
Piezoelectric "bender" structure.

material. When more than 50% of the domains have switched, the material will begin to expand again.

The amount of displacement caused in the case of the opposed situation can be estimated by the following formula:

$$D_3 = +d_{33} \times (V_3)$$

where D_3 is the displacement in the "3" direction, (the "3" direction is perpendicular to the capacitor plates), V_3 is the applied voltage between the plates of the capacitor, and d_{33} is the coefficient for displacement parallel to the field, typically 7×10^{-12} m/V².

9.8.2 MEMS Manufacture

Exploitation of the existing silicon fabrication technology leads naturally to the manufacturing process represented by Figure 9.15. Each MEMS module is typically 4×4 mm. The modules are produced in arrays on a silicon wafer to yield 50 to several hundred modules per wafer.

The correct operation of the MEMS depends on symmetry of the performance of the two benders in the assembly. This symmetry and repeatability of performance from module to module are achieved only through precise dimensional, process, and material control.

The fabrication process is as follows:

1. A silicon wafer with a buried oxide layer is oxidized (Figure 9.15a).
2. Front- and backside alignments are applied in photoresist (not shown).
3. The wafer is etched in BOE (buffered oxide etch), removing the oxide layer, leaving identical marks on front- and backside (Figure 9.15c).
4. The wafer is etched in NaOH leaving a protruding pyramid at the backside and an inverted pyramid on the front side. (Figure 9.15c).
5. The wafer is reoxidized (Figure 9.15d).
6. The backside is now etched to depth of 350 microns, which leaves a varying thickness of the membrane on the front side, according to the location of the point where the measurement is made. A Pt under electrode is deposited and PZT is applied as a sol–gel. A Pt top electrode is deposited (Figure 9.15f).
7. On the front side the outline of the double monomorph optical scanner structure is etched out to the depth of the insulating buried oxide layer, while the backside is protected (Figure 9.15g).
8. The wafer is reoxidized (not shown).
9. 25 microns is removed from the backside, to etch free the double monomorph optical scanner structure. The etching stops at the buried oxide layer (Figure 9.15h).

Figure 9.16 shows a schematic of a double monomorph optical scanner. Figure 9.17 shows an SEM photograph of a scanner at 50× magnification.

9.8.3 Operation of the Scanner

In Figure 9.18 a cross section is given, in which the eye is located in the plane of the wafer. The angle β is the angle the mirror makes with the plane of the wafer.
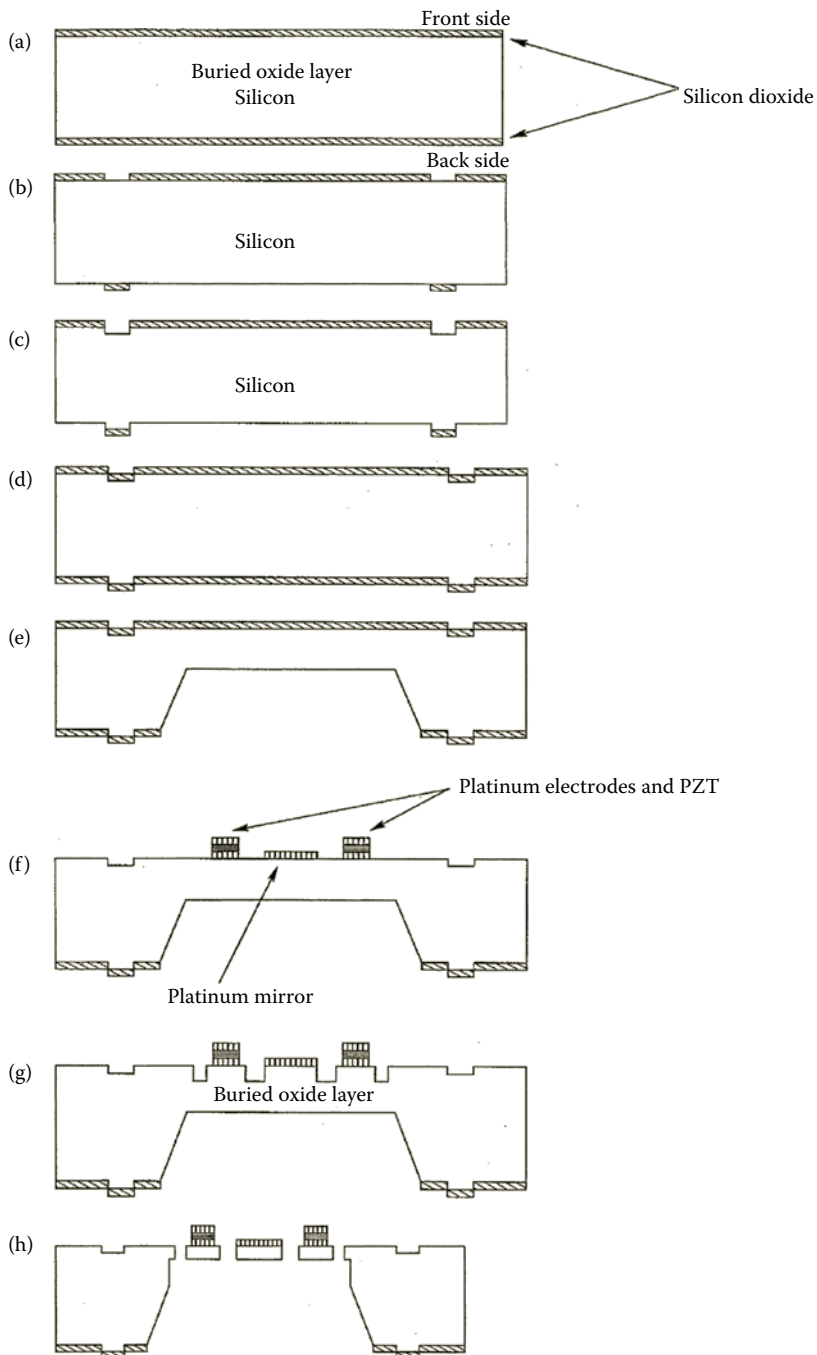
FIGURE 9.15
An illustration of the fabrication process. (a) Oxidized silicon blank, (b) First etch, (c) Second etch, (d) Second oxidation, (e) Third etch, (f) Deposition step, (g) Fourth etch, (h) Final etch.
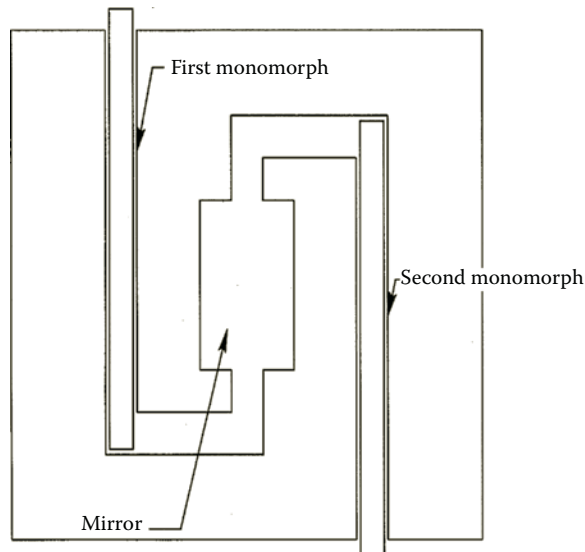
FIGURE 9.16
Schematic of double monomorph optical scanner.
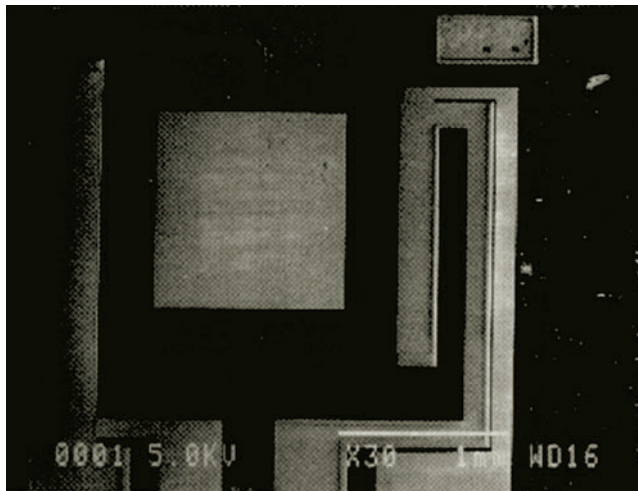
FIGURE 9.17
5× SEM photograph of MEMS scanner.

9.8.4 Material Properties

The properties required of the driving materials are a reasonably high coefficient of strain versus applied field, low or predictable hysteresis, and high repeatability. The properties required of the support materials are ease of fabrication, very high stability, and a high fatigue limit. These are met satisfactorily in the silicon mechanical material chosen: it was in fact, these demonstrated characteristics of silicon that led to its choice as a basis material. On the other hand, while the PZT material has a satisfactory coefficient of strain, these constraints remain. Although these materials are excellent for submicron positioning, they
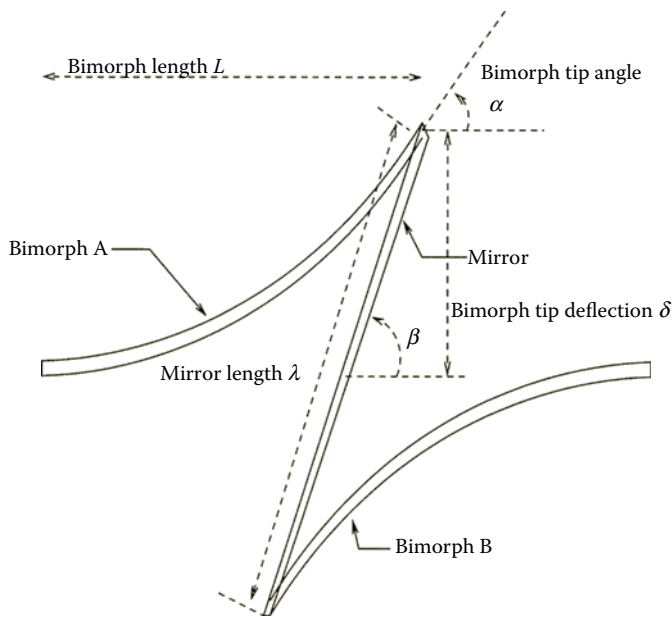
FIGURE 9.18
Cross-sectional schematic of double monomorph optical scanner.

have inherent hysteresis and creep, and so they lack the level of repeatability required in a practical open-loop positioning device. Long-term creep can approach 15%, and hysteresis can approach 12%.

9.8.5 Static Performance

9.8.5.1 Hysteresis

Hysteresis appears as apparent "backlash" upon reversal of the direction of motion, but unlike the backlash of conventional mechanical systems and stiction, which can largely be predicted and compensated, hysteresis depends on the recent history of motions and is difficult to model, predict, and compensate. The creep is, at this time, completely unpredictable. As a result, open-loop piezoelectric devices are limited to applications in which repeatability of 10% is satisfactory. Of course, when combined with a position feedback system, this defect can largely be overcome, but the cost of this level of control is disproportionate to the economics of the production of MEMS.

9.8.5.2 Linearity

Linearity of motion over mirror angles of 10° mechanical is in the range of 2%, smoothly changing, monotonic, and predictable. This characteristic is considered satisfactory for most laboratory uses. However, PZT has a rather large strain sensitivity to temperature, so that linearity of motion can be compromised by changes in temperature.

9.8.5.3 Uniformity

Uniformity of performance scanner to scanner within a wafer is entirely a question of process control, and is not expected to pose long-term problems. Standard levels of production

process control are also expected to address wafer-to-wafer performance uniformity issues satisfactorily. At the moment, we are achieving uniformity of characteristics of about 10% across the wafer and wafer to wafer in a pilot production phase.

9.8.5.4 Yield

Yield is an issue that falls into the process control purview as well. At the moment, we are achieving yields above 80%, with most of the dropouts the result of mirror defects.

9.8.6 Dynamic Performance

9.8.6.1 Dynamics

These scanners have a first torsional resonance typically well above 20 kHz with integral mirrors. As a result, the dynamics permit full-scale (10°) steps in well under 100 μs. With high-quality mirror flakes 150 microns thick cemented over the integral mirror, the resonant frequency falls to about 8 kHz, which is satisfactory for most microscanning purposes. The flatness of the present examples of integral mirrors is poor, as is the surface finish, but attached flakes achieve any required surface quality. Testing has verified that 150 micron thickness is adequate to preserve quarter wave or better flatness over the 1.8 × 2.6 aperture of the standard scanner because of the underlying silicon support ring.

9.8.6.2 Life

Life testing has revealed no observable damage after 10^{10} 60° optical scans when driven by an 8-kHz oscillator. Flexure stress is expected to be below the endurance limit for silicon at all angles of operation up to 30° mechanical.

9.8.6.3 Degradation Processes

No degradation process is known, other than excessive voltage breakdown inside the PZT crystal, and heating above the Curie point. High-speed operation (above 20 kHz) has the potential to encounter impact degradation of the tips of the bender arms and the edges of the mirror if operated in open air. The standard packaging presently envisioned for individual scanners is a hermetically sealed TO5 package.

9.8.7 Application Rules

9.8.7.1 When and When Not to Use MEMS

The current state of the art in PZT open-loop scanners has limited repeatability and "staring" positional stability because of creep and hysteresis. As a result, MEMS scanners should not be specified for applications in which positional precision better than 10% of full scale is required unless some form of active position feedback is anticipated. On the other hand, the device behaves as though it were a pair of capacitors of a few nanofarads capacity. Therefore, self-heating even during aggressive scan profiles is not an issue, and cooling is not required except in environments above 100 °C. Because the thermal transport path from the mirror through the structure is long, the effect of power coupled from the scanned beam may be significant. In this case, it is desirable to provide a direct mirror cooling path through a suitable gas, such as helium, directly to the scanner case.

Vibration and inertial forces applied to the scanner are likely to cause deformation of the flexure structure, and, since the system is very poorly damped, are likely to produce long-lasting settling periods (Figure 19.9). Otherwise, these scanners have no particular susceptibility to environmental stimuli such as temperature, barometric pressure, humidity, magnetic and electric fields, and so forth.

9.8.8 Anticipated Developments

It is anticipated that electrostrictive actuators may replace piezoelectric actuators in MEMS in the near future. A typical electrostrictive material, such as lead–magnesium–niobate (PMN), should provide an order of magnitude improvement over PZT in positional stability.

For PMN materials, the change in length is proportional to the field voltage squared, so the coefficient is a factor of two higher than PZT. Unlike piezoelectric materials, PMN crystals are not poled. Positive or negative voltage changes result in an elongation in the direction of the applied field, regardless of its polarity. Because PMN is not poled, it is inherently more stable than PZT, resulting in a reduction of long-term creep from 15% to 3%. Also, PMN materials have better hysteresis than does PZT. While PZT exhibits 12% hysteresis, PMN displays only 2%.

Two properties of PMN contribute to its thermal stability. First is the strain sensitivity to temperature. The PMN is much more robust than PZT in this regard, especially over large temperature ranges. Secondly, the coefficient of expansion of PMN is twice that of PZT.

9.8.9 Conclusions

It is clear that MEMS have not yet grown from infancy. Unlike their mature brethren, the macroscale flexure scanners, the only conclusion one could legitimately draw at this time is that the future is bright for small scanning and pointing devices that draw little power, pack closely together, demonstrate extreme reliability, and offer a price–performance index beyond the capability of any macroscale device. It seems certain that these attributes of MEMS will be exploited. Just how MEMS will look in their maturity years hence is anyone's guess.
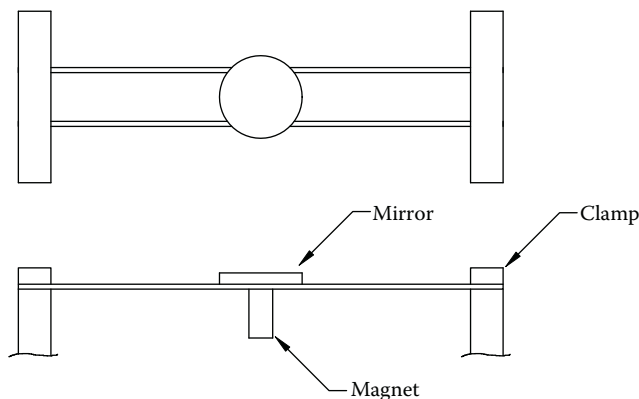
FIGURE 9.19
Vibrating wire scanner.

9.9 CONCLUSION

It seems that flexure pivots have come of age. They have been used in a great diversity of products where their attributes of sensitivity, accuracy, and repeatability are enabling. They are also low cost, lubricant-free, low mass, high "Q," and capable of storing considerable energy. They can be cascaded to provide vibration isolation. There are standalone commercial versions to suit many purposes. The published data on detail design rules are considerable, and no competent engineer should have unusual difficulty in producing a workable pivot on the first try. It is the hope of the authors that the information in this chapter will help neophytes to avoid most of the impediments to success that lurk in the mysteries of material selection, processing, and the mounting of flexures and that many more useful applications of flexure pivots will be found and pursued with success.

ACKNOWLEDGMENTS

REFERENCES

1. Weinstein, W.D. Flexure pivot bearings. *Machine Design* 1965, *37,* 136–145.
2. Bendix flexural pivot. *Bendix Electric and Fluid Power Division,* Application Notes, Catalog, Bendix Corp., Utica, NY.
3. Sines and Waisman, Eds. *Metal Fatigue;* McGraw Hill 1959, 89–111.
4. Boyer, H.E., Ed. Failure analysis and prevention. In *Metals Handbook,* 8th Ed., Vol. 10; American Society for Metals: Metals Park, OH; 208–249.
5. Paros, J.M.; Weisbord, L. How to design flexure hinges. *Machine Design* 1965, *37,* 151–156.
6. Neugebauer, G.H. Designing springs for parallel motion. *Machine Design* 1980, *52,* 119–120.
7. Troeger, H. Considerations in the application of flexural pivots. *Automatic Control* 1962, *17*(4), 41–46.
8. Paulsen, D.R. Flexural Pivot. U.S. Patent 4,802,720, February 7, 1989.
9. Brosens, P.J. Resonant Optical Scanner. U.S. Patent 5,521,740, May 28, 1996.
10. Sawyer, C.B. The use of Rochelle salt crystals for electrical reproducers and microphones. *Proc. Inst. Radio Eng.* 1931, *19*(11), 2020–2029.
11. Smits, J.G.; Dalke, S.I.; Cooney, T.K. The constituent equatons for piezoelectric bimorphs. *Sensors and Actuators* 1991, *28,* 41–61.

12. Kugel, V.D.; Xu, B.; Zhang, Q.M.; Cross, L.E. Bimorph based piezoelectric air acoustic transducer: A model. *Sensors and Actuators.*
13. Caliano, G.; Lamberti, N.; Iula, A.; Pappalardo, M. A piezoelectric bimorphstatic pressure sensor. *Sensors and Actuators A* 1995, *46*(1–3), 176–178.
14. Coughlin, M.F.; Stamenokic, D.; Smits, G. Determining spring stiffness by the resonance frequency of cantilevered piezoelectric bimorphs. *IEEE Trans. Ultrasonics, Ferroelectrics and Frequency Control* 1977, *44*, 730–733.
15. Kielczynski, P.; Pajenski, W.; Salewski, M. Piezoelectric sensors for the investigation of microstructures. *Sensors and Actuators A* 1998, *65*(1), 13–18.
16. Juan, I.; Roh, Y. Design and fabrication of piezoceramic bimorphvibration sensors. *Sensors and Actuators A* 1998, *69*(3), 259–266.
17. Van Mullem, C.J.; Blom, F.R.; Fluitman, J.H.J.; Elwenspock, M. Piezoelectrically driven silicon beam force sensor. *Sensors and Actuators A* 1991, *26*(1–3), 379–383.
18. Naber, A. The tuning fork as a sensor for dynamic force control in scanning near-field optical microscopy. *J. Microscopy-Oxford* 1999, *194*(2–3), 307–331.
19. Yamada, H.; Itoh, H.; Watanabe, S.; Kobayashi, K.; Matsushige, K. Scanning near-field optical microscopy using piezoelectric cantilevers. *Surface and Interface Analysis* 1999, *27*(5–6), 503–506.
20. Kielczynski, P.; Pajewsli, W.; Sealcwski, M. Piezoelectric sensor applied in ultrasonic contact microscopy for the investigation of material surfaces. *IEEE Trans Ultrasonics, Ferroelectrics and Frequency Control* 1999, *46*(1), 233–238.
21. Edwards, H.; Taylor, L.; Duncan, W.; Melemed, A.J. Fast, high-resolution atomic force microscopy using a quartz tuning fork as actuator and sensor. *J. Appl. Phys.* 1997, *82*(3), 980–984.

10

Holographic Barcode Scanners: Applications, Performance, and Design

LeRoy D. Dickson

Wasatch Photonics, Inc.
Logan, Utah, USA

Timothy A. Good

Metrologic Instruments, Inc.
Blackwood, New Jersey, USA

CONTENTS

10.1 INTRODUCTION

Significant changes have occurred in the field of laser scanning since the first edition of this book was published over 10 years ago. Specifically, visible laser diodes (VLDs) have become the laser light source of choice in the scanning industry, allowing scanners to become much smaller, in the form of hand-held and wearable scanners. Holographic scanning, however, does not yet have a very significant presence in these applications.

In the first edition, much attention was given to supermarket scanners and most of the examples were given in reference to such designs. Over the past decade, however, a great deal of growth has occurred in the industrial scanning market, and the adaptability of holography has helped it grow into this market more significantly. Accordingly, more examples are given in this edition with respect to the industrial scanning market.

The fundamentals of scanning, however, have not changed. Neither has the presence of several decades of barcode and scanner design specifications and laser standards been diminished. For example, printing specifications of barcodes are based on reflectance properties at the wavelength of helium–neon lasers, predominantly used years ago but no longer in scanners today. As such we will begin with a basic discussion of barcodes.

A barcode is a sequence of dark bars on a light background or the equivalent of this with respect to the light reflecting properties of the surface. The coding is contained in the relative widths or spacings of the dark bars and light spaces. Perhaps the most familiar barcode is the universal product code (UPC), which appears on nearly all of the grocery items in supermarkets today. Figure 10.1 is an example of a UPC.

A barcode scanner is an optical device that reads the code by scanning a focused beam of light, generally a laser beam, across the barcode and detecting the variations in reflected light. The scanner converts these light variations into electrical variations, which are subsequently digitized and fed into the decoding unit, which is programmed to convert the relative widths of the digitized dark/light spacings into numbers and/or letters.

The concept of barcode scanning for automatic identification purposes was first proposed by N. J. Woodland and B. Silver in a patent application filed in 1949. A patent, titled "Classifying Apparatus and Method," was granted in 1952 as U.S. Patent No. 2,612,994. This patent contained many of the concepts that would later appear in barcode scanning systems designed to read the UPC.

10.1.1 The UPC Code

In the early 1970s, the supermarket industry recognized a need for greater efficiency and productivity in their stores. Representatives of the various grocery manufacturers and

FIGURE 10.1
Typical UPC barcode.

supermarket chains formed a committee to investigate the possibility of applying a coded symbol to all grocery items to allow automatic identification of the product at the checkout counter. This committee, the Uniform Grocery Product Code Council, Inc., established a symbol standardization subcommittee, whose purpose was to solicit and review suggestions from vendors for a standard product code to be applied to all supermarket items.

On April 3, 1973, the Uniform Grocery Product Code Council announced their choice. The code chosen was a linear barcode similar to a design proposed by IBM. The characteristics of this barcode, the now familiar UPC, are described in detail in the article by Savir and Laurer.[1]

The UPC is a fixed-length numeric-only code. It consists of a pair of left guard bars, a pair of right guard bars, and, in the standard version A symbol shown in Figure 10.1, a pair of center guard bars. Each character is represented by two dark bars and two light spaces. The version A symbol contains 12 characters, six in the left half and six in the right half. Thus, a version A UPC symbol will have 30 dark bars and 29 light spaces, counting the six guard bars—left, right, and center. The first character in the left half is always a number system character. For example, grocery items are given the number system 0, which often appears on the left of the symbol. The last character on the right is always a check character. This sometimes appears to the right of the barcode symbol.

The remaining five characters in the left half of the version A UPC symbol identify the manufacturer of the product. For example, the left-half number 20000 represents Green Giant products. This left-half five-digit code is assigned to the various manufacturers by the Uniform Product Code Council.

The remaining five characters in the right half of the version A UPC symbol identify the particular product. This right-half five-digit code is assigned by the product manufacturer at his discretion. For example, Green Giant has assigned the right-half number 10473 to their 17 oz. can of corn. Therefore, the complete UPC code for the Green Giant 17 oz. can of corn, ignoring the number system character and the check character, is 20000–10473.

There are a number of other properties of the UPC code and symbol that are significant relative to the design and use of equipment for reading the code. First, the left and right halves of the version A symbol are independent. That is, each half can be read independently of the other half and then combined with the other half in the logic portion of the reader to yield the full UPC code. Furthermore, as shown in Figure 10.2, each half of the UPC symbol is "over-square." That is, the symbol dimension parallel to the bars is greater than the symbol dimension perpendicular to the bars. The aspect ratio of the barcode is vital in the determination of a minimum scan pattern for reading the code, as we will see in a later discussion.

FIGURE 10.2
Two UPC half-symbols.

It should be noted that the original "over-square" design of the UPC code is not always adhered to by product manufacturers. Often they will truncate the height of the code. These truncated codes, while discouraged by the Uniform Product Code Council, are often used by manufacturers to maximize the space available for product information. That, in turn, has meant that scanner designs must be more complex and the decoding algorithms more sophisticated.

Each character of the code is represented by two dark bars and two light spaces. Individual bars and spaces can vary in width from one module wide to four modules wide. (Note, on all UPC codes, the guard bars are always one module wide and separated by a one-module-wide space.) The total number of modules in each character is always seven. The left-half characters are coded inversely from the right-half characters. As shown in Figure 10.3, for example, if we let white = 0, and black = 1, then the code for the number 2 in the left half is 0010011 while in the right half it is 1101100.

The fact that each character is always seven modules wide leads to a second major property of the UPC code: it is self-clocking. Therefore, absolute time measurements are unimportant. What is measured is the time that is required to go from the leading edge of the first black bar to the leading edge of the third black bar (i.e., the first black bar of the next character). This time interval is then divided into seven equal intervals, and the relative widths of the two black bars and the two white spaces are then determined for decoding purposes. Thus, the total width of a character—black-white-black-white—is measured, and the relative widths of the two black bars and the two white spaces are determined for decoding.

This self-clocking feature is very important in the design of scanners for reading the UPC code. It means that the velocity of a scanning light beam for reading the code does not have to be constant across the full width of the code. The velocity only needs to be reasonably constant across a single character. This means that the UPC code can be read by moderately nonlinear scan patterns, such as sinusoidal or Lissajous patterns. It also means that the code can be read on curved surfaces. Furthermore, the scan lines reading the code do not have to be perpendicular to the bars and spaces in the code. As originally designed, satisfactory reading of a UPC code can be obtained with scan lines that pass through the code at any angle relative to the bars and spaces so long as a single scan line passes completely through a full half symbol, including the center guard bars and one pair of edge guard bars. More sophisticated decoding algorithms now exist that can "stitch" together three smaller, individually scanned pieces of a UPC code (or other symbol), and, as such, the process is commonly referred to as stitching. The use of stitching allows a slightly less thorough scan pattern to do as good a job as a better pattern and makes a good pattern even better; however, the UPC code was designed with only the left and right sides in mind. Stitching was a software adaptation developed later on.[2] One of the factors driving its development was the occurrence of the aforementioned truncated codes, for which stitching is particularly useful.

(a) 0 0 1 0 0 1 1   (b) 1 1 0 1 1 0 0
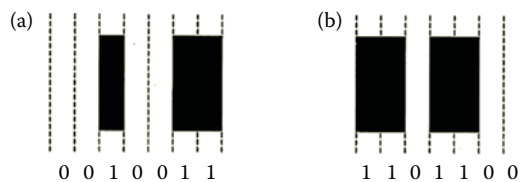
FIGURE 10.3
An example of character encoding for the number 2: (a) left-half character; (b) right-half character.

A third property of the UPC symbol of significance to the scanner designer is the size of the symbol, which is allowed to vary from the nominal size (about 1.0 in × 1.25 in [25.4 mm × 31.75 mm] for the version A symbol) down to 0.8 × nominal and up to 2.0 × nominal. This size variation allows the use of small labels on small packages with good print quality and large labels on larger packages with poorer print quality. From the scanner designer's viewpoint, the small label will establish the minimum bar width to be read and the large label will set a lower limit on the size of the scan pattern.

The minimum bar width established by the UPC specification, including tolerances, is 0.008 in (0.2 mm). This number establishes the maximum attainable depth of field for the optical reader. In practice, the depth of field of the typical laser scanner designed to read the UPC code will easily meet, and exceed, the 1 in (25.4 mm) depth of field required by the early UPC guidelines. However, this 1 in depth of field did not take into consideration the manner in which the scanners would eventually be used. Depths of field of several inches (100 mm+) are required for today's UPC barcode readers.

Finally, the contrast specification for the UPC symbol requires that the contrast be measured using a photomultiplier detector (PMT) with an S-4 photocathode response curve coupled with a Wratten 26 filter. This combination has a peak response at a wavelength of approximately 610 nm (24 μin), falling to zero at approximately 590 nm (23.2 μin) and 650 nm (25.6 μin). This response includes, not coincidentally, the wavelength of the helium–neon laser, 632.8 nm (24.9 μin), which was the preferred laser at the time that laser scanners were first being considered. While several of the inks used for printing UPC labels can provide acceptable contrast out to 700 nm (27.56 μin) there are many other inks in use that do not provide acceptable contrast beyond 650 nm (25.6 μin). These inks would preclude general use of longer-wavelength light sources. Today, nearly all UPC scanners use one or more diode lasers as their light source(s); however, the wavelengths of these lasers fall within the original UPC wavelength specification.

10.1.2 Other Barcodes

The UPC code is not widely used in the industrial environment (the manufacturing, warehouse, and distribution applications). Here, the requirements are different from those of the supermarket, so the codes used are different than the UPC code. The preferred codes for the industrial environment are Bar Code 39, Interleaved 2 of 5, and Codabar.

The most common barcode in the industrial environment is the so-called Bar Code 39, or 3 of 9 barcode. This code is fully alphanumeric and is self-checking. For a full discussion of Bar Code 39, as well as several other codes, and the definition of such terms as "self-checking," see work by Allais.[3] The code shown in Figure 10.4 is an example of Bar Code 39.

Bar Code 39 got its name from the fact that it originally encoded 39 characters: the 26 letters of the alphabet, the numbers from 0 through 9, and the symbols -,., and SPACE, plus a unique start/stop character, the asterisk (*). Today it also encodes the four so-called special
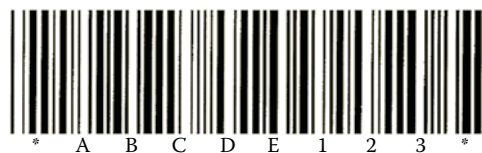
FIGURE 10.4
An example of Bar Code 39.

characters: $, /, +, and %, for a total of 43 characters. However, it is still referred to as Bar Code 39. It is also often called the 3 of 9 code, because each character in the code is represented by nine elements (five dark bars and four light spaces), and three of them are wide with the remaining six narrow. In the primary set of 39 characters, two of the wide elements are dark bars. In the four special characters, the wide elements are all light spaces.

The 2 of 5 code is a subset of the 3 of 9 code. In 2 of 5, only the bars are used for encodation. Two of the five bars are wide, just as in the original 3 of 9 code. The spaces are not used. This code is strictly numeric. The basic 2 of 5 code is not widely used in the industrial environment, but a variation of it, called the Interleaved 2 of 5 code, is used extensively for manufacturing and distribution applications. This code uses the bars to encode one character in the standard 2 of 5 code and then uses the interleaving spaces to encode a second character in 2 of 5 code. This allows more characters to be encoded in a fixed barcode length than either 3 of 9 code or 2 of 5 code. This code is also only numeric, but, due to the interleaving feature, it can encode nearly 80% more characters per unit length than Bar Code 39, assuming both codes have the same minimum bar width. For this reason, the Interleaved 2 of 5 code is often used where space limitations will not permit the use of Bar Code 39.

A third code that is used extensively in medical institutions, and which was adopted as an early standard by the American Blood Commission for use in identifying blood bags, is Codabar. This code is also frequently seen in some transportation and distribution applications.

10.1.3 Barcode Properties

From the standpoint of the scanner, the important properties of any barcode are:

1. Minimum bar width: generally specified in millimeters or mils (thousandths of an inch). And often referred to as the "X" dimension.

2. Contrast: a measure of the reflectance of the bars and spaces. Contrast is generally expressed in terms of the print contrast signal (PCS), defined as

$$\text{PCS} = \frac{r_s - r_b}{r_s} \tag{10.1}$$

where r_s is the reflectance of a space and r_b is the reflectance of a bar. It should be noted that PCS is usually measured for one particular wavelength of light. In the majority of applications this wavelength is 633 nm (24.9 µin), the wavelength of the helium–neon laser, which was the most common light source for most of the early laser scanners. (Some applications allow PCS to be measured at 900 nm [35.4 µin], the wavelength of some infrared light sources used in some readers.) This is an important point to remember since PCS will vary drastically as a function of wavelength if colored inks or backgrounds are used. In practice, the most important reflectance property of the barcode is the absolute contrast, which is simply the space reflectance minus the bar reflectance (the numerator in Equation 10.1).

3. Code length: the physical length of a barcode is determined by the density of the code (which is determined by the minimum bar width) and the number of characters in the code. The physical length of the code determines how long the scan lines must be and, when combined with the code height, will determine how accurately the scan line must be oriented with respect to the barcode.

4. Code height: the height of the barcode (the dimension parallel to the bars) will determine the angular accuracy required in orienting the scan line relative to the barcode.

5. Barcode quality: this includes both the quality of the printing or etching of the code itself and the quality of the surface on which the code is printed. Obviously, the better the quality of both, the easier it will be for the scanner to successfully scan and decode the barcode.

There is a great deal more that could be said about the barcodes themselves. However, a more detailed analysis of the fundamental properties of the barcodes is beyond the scope of this review and is not really necessary for the purposes of our discussion of barcode scanning.

10.2 NONHOLOGRAPHIC UPC SCANNERS

A block diagram of a typical laser scanner system for reading the UPC code is shown in Figure 10.5. The focused laser beam scans the UPC symbol on a package as the package passes over the read window of the scanner. The laser beam is reflected from the symbol as it passes over the dark bars and light spaces. The diffuse portion of the reflected light is modulated by the reflectivity variations of the symbol (bars and spaces). This light modulation is detected by the photodetector, which converts the light modulation into electrical modulation. The electrical "signal" is then amplified, digitized, and transmitted to the "candidate select" block. This block acts as a filter, allowing only valid UPC half-symbols to pass to the decoder. The decoder converts the signals for each half-symbol into characters and then combines the characters for the two half-symbols together to yield a complete UPC product identification code. The computer then searches its memory for a description and price of the item identified by this UPC code. This information is transmitted back to the checkout terminal where it appears on the display and the customer receipt. Simultaneously, the store inventory is updated to reflect the sale of the identified item. All of this takes place in a few milliseconds.

The focused spot size of the scanning laser beam must be about 0.2 mm in order to be able to read the labels with the smallest bar widths while still yielding adequate depth of
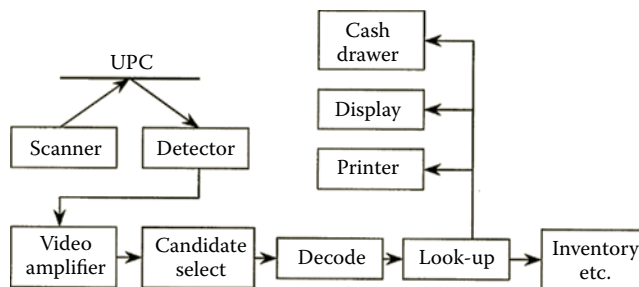
FIGURE 10.5
Block diagram of a UPC scanner.

field. This requires an optical *f*-number of approximately 250, which, when combined with scanner geometry, establishes the focusing optics' requirements.

A number of technologies are available for deflecting the focused laser beam in a conventional, nonholographic, UPC scanner. Cost and performance requirements limit the choice to mechanical deflectors—generally either rotating or oscillating mirrors, or a combination of these. The scan pattern created by the laser deflection mechanism must be capable of reading a full-size version A UPC symbol regardless of the orientation of the symbol in the scan window. In other words, the scanner must be omnidirectional. An omnidirectional scanner will allow maximum freedom for the scanner operator when bringing the item across the scan window.

We have already seen that the two halves of the UPC symbol can be read independently of each other. We have also seen that each half-symbol is over-square. Therefore, the minimum scan pattern that will allow omnidirectional scanning is a pair of perpendicular scan lines in the form of an X (see Figure 10.6a). As the UPC symbol passes over the scan window, at least one of the legs of the X will pass through the entire half-symbol at some point in the window. Figure 10.6a shows two extreme orientations of the symbol as it is passed over the window. These are the worst-case orientations in that they allow the minimum time for scanning the symbol satisfactorily.

The amount that the half-symbol is over-square, when combined with the maximum item velocity of 2.54 m/s (100 in/s), determines the minimum pattern repetition rate to guarantee at least one good scan through the symbol as it passes across the scan window, regardless of its orientation. The pattern repetition rate, the total scan length, and the width of the smallest UPC module establishes the maximum video signal rate seen by the photodetector.

Although the pattern in Figure 10.6a is the minimum scan pattern required to yield an omnidirectional scanner for the UPC symbol, it is not an "optimum" pattern. The pattern repetition rate required to guarantee one scan through a UPC half-symbol moving across this pattern at 2.54 m/s (100 in/s), at the worst-case orientation relative to the scan pattern, is very high. This results in high scanning spot velocities and subsequently high video signal rates. A "better" scan pattern can guarantee one good scan at lower pattern repetition rates and lower scan velocities. An optimum scan pattern will minimize scan velocity, thereby minimizing video signal rates.

If one could increase the amount that a symbol is over-square (i.e., improve the aspect ratio), then one could reduce the pattern repetition rate, and the scan velocity, and still guarantee one good scan through the UPC half-symbol at maximum symbol velocity and worst-case symbol orientation. While the symbol itself cannot be changed, one can effectively improve the aspect ratio of the symbol by using a scan pattern where the scan lines are separated by angles less than 90°. Thus, a scan pattern consisting of three scan lines,
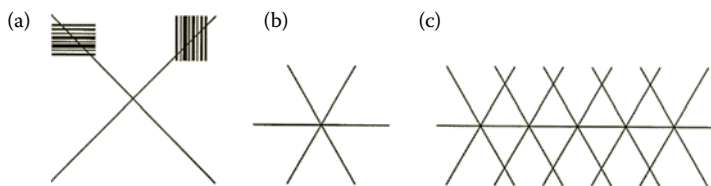
FIGURE 10.6
Omnidirectional scan patterns in the plane of the window: (a) the minimum scan pattern; (b) basic pattern of an optimum scan pattern; (c) optimum pattern for a rectangular window.

for example, instead of two, could be repeated less often and still be able to read the UPC symbol under the worst-case conditions mentioned above. Increasing the number of scan lines has the effect of increasing the total linear distance scanned by the scan pattern, which, by itself, would increase the scan velocity of the scanning spot. However, the reduction in the pattern repetition rate is greater than the increase in the scan length. The net result is a better scan pattern, in the sense described above.

Can we continue to improve the scan pattern by adding still more lines? Unfortunately, the answer is no. The reduction in the required pattern repetition rate realized by using four scan lines is nearly offset by the increased scan velocity resulting from the greater distance traveled by the scanning spot. The small amount of gain is not enough to justify the increased cost and complexity required to generate the four-line pattern. Beyond four lines, there is no gain when scanning UPC symbols. (We will, however, see later that in certain industrial applications four- and five-line patterns can be quite effective, especially when scanning symbologies with extreme aspect ratios.) It appears, then, that for the UPC code the optimum scan pattern in the plane of the window would be one based on the three-line pattern shown in Figure 10.6b. This fundamental three-line pattern, which is still the basic criterion used in the design of UPC scanners in 2004, formed the basis of the first scanner designed to read the UPC code, the IBM 3666 scanner. The linear equivalent of the Lissajous scan pattern used in the IBM 3666 scanner is shown in Figure 10.6c.

10.2.1 Forward-Looking Scanners

Initially, all UPC scanners were conceived as "bottom scanners." That is, the scanning laser beam pointed directly upward to read the UPC symbols on the bottoms of packages as they passed over the scan window. A major problem was encountered in the design of this type of scanner. UPC symbols printed on shiny surfaces were difficult to read because the specular reflection from the shiny surfaces contained no bar-space modulation in the specularly reflected light. In addition, the specular reflection created saturation problems in the photodetector because the specularly reflected light was so much more intense than the diffusely reflected light. In most scanners, the photodetector was located back along the general direction of the outgoing laser beam. Therefore, some solution had to be found that would keep the specularly reflected light from being directed back along the laser beam path.

One solution to this problem is shown in Figure 10.7a. The laser beam is tilted at an angle of approximately 45° relative to the scanner window. In this configuration, the specularly reflected light is reflected away from the photodetector, thereby eliminating the specular reflection problem.
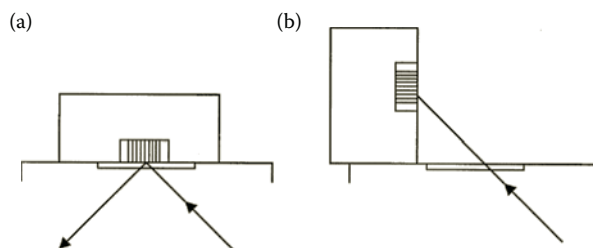
FIGURE 10.7
Tilted-beam scanning: (a) removing the specular reflection problem; (b) side reading with a forward-looking tilted beam.

A fringe benefit occurs when this scanner geometry is used. The tilted beam can be used to read UPC symbols on the front of packages without tilting the packages forward (see Figure 10.7b). Of course, this increases the depth of field required to read these upright labels, but the laser scanner has the capability to provide a depth of field of several inches (100 mm+), which is usually sufficient for side reading with the tilted beam geometry. Nearly all UPC scanners today employ some form of forward-looking, tilted-beam reading geometry.

10.2.2 Scan Pattern Wraparound

The next development in the evolution of the scan pattern was the introduction of scan pattern wraparound. Several scan patterns were introduced that took the basic three-line optimum scan element shown above and created it in such a way that the scan lines were directed at the items from points within the scanner that were slightly off to the sides of the package. In these scanners, horizontal lines were projected forward from immediately in front of the item and from directions slightly off to each side. Vertical lines were also projected from slightly off to each side. The pattern projected from the two sides was essentially a cross pattern, while the scan pattern projected from immediately in front of the item was a horizontal line, as shown in Figure 10.8. The overall pattern was created by using a rotating mirror deflector and an array of fixed folding mirrors.

This type of scan pattern was effective in reading the UPC symbol on the scan window because it employed the basic three-line optimum scan element. It was also effective in reading upright items because it projected a pattern of perpendicular horizontal and vertical lines on the front of the items. Such a pattern is effective for upright reading because the bars in the UPC symbol will usually be either vertical or horizontal when the package is presented to the scanner in this manner.

The major advantage of this type of scan pattern is that it "wraps around" to the sides of the packages to some degree. This means that the operator does not have to align the item as carefully when he brings the item across the scan window. The UPC symbol can be on the bottom of the package, on the front of the package, or on the side of the package and still be readable by the scanner. This has a positive effect on operator productivity.

This concept of wraparound was further exploited in the development of "bi-optic" scanners, exemplified by the NCR 7875, the PSC Magellan SL, and the Metrologie Stratos.
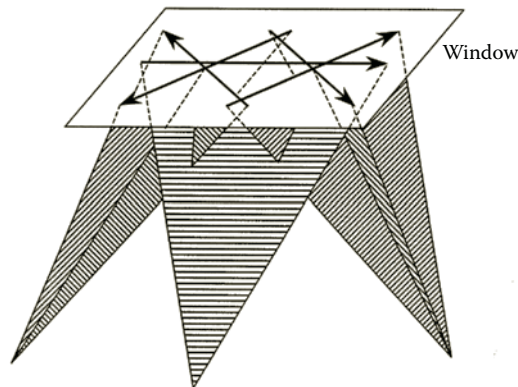
FIGURE 10.8
Projected wraparound scan pattern.

The bi-optic-type scanner is currently becoming the most used type of scanner in super-markets and other large-volume, point-of-sale applications. The bi-optic nature of the scanner refers to the two separate scan windows it possesses, as shown in Figure 10.9. In this configuration there is a horizontal scan window and a second window at or near vertical orientation, depending on the specific scanner in question.

Improved performance is gained by employing the wraparound concept from both windows. In the preferred package-presentation orientation, two of the six faces of an item directly face a scanner window and are seen by the primary three-line pattern. Two other faces, typically the faces in the direction of item motion (see Figure 10.9), are targeted by the wraparound patterns of both scanner windows. The package surface that faces away from the vertical window still potentially sees the wraparound pattern out of the horizontal window. Finally the top surface of the package, which has the least exposure to the scan pattern, may still have some chance of being scanned by the wraparound pattern of the vertical window, depending on the package height and code orientation. The net result is effective scanning throughout the majority of the 360° horizontal orientation range and equally effective scanning through nearly 270° of vertical orientation. Such a wide range of acceptable presentation orientations means very little of an operator's time needs to be spent paying attention to the position and orientation of a code on an item.

10.2.3 Depth of Field

The multidirectional, three-line scan pattern forms the basis for nearly all present-day UPC scanners, holographic and nonholographic. Unfortunately, the forward-looking feature increases the depth of field requirement considerably. As much as 150-mm (6-in) depth of field may be required to read some barcodes on upright items. Because the codes must, in many cases, be read by a tilted beam, the resultant spot ellipticity on the barcode will increase the effective scanning spot diameter. This will reduce the depth of field of the scanner.

Providing satisfactory scanning performance over such a large depth of field, with a tilted scanning beam, is a significant challenge to the scanner designer. Significant improvements in signal processing over the last decade have allowed smaller bar widths to be read without reducing the actual size of the spot, helping to solve this problem and increase depth of field. Another means of easing the problem, without relying on electronic improvements, is to design a scanner that can provide more than one focal plane. Such a scanner could focus
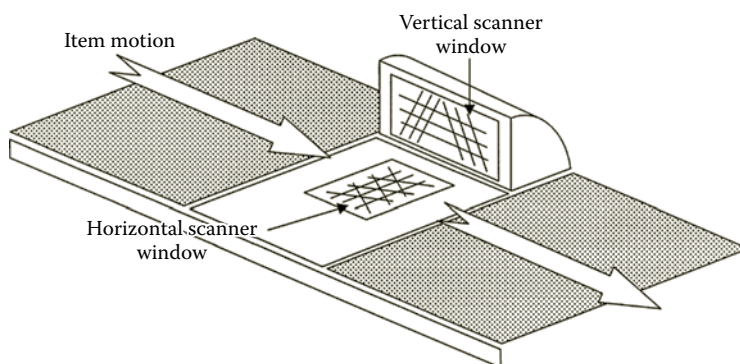
FIGURE 10.9
A "bi-optic" supermarket scanner.

some of its scan lines close to the scan window and some of its lines further from the scan window, thereby increasing the effective depth of field of the scanner.

Holographic scanning allows the scanner designer to add this additional degree of flexibility to the scanner. The holographic scanning element essentially allows each scan line to be optimally focused to provide increased depth of field and increased flexibility in the placement of beam-folding mirrors for the creation of the scan pattern. This also allows for a more complex, and more effective, scan pattern. The need for greater depth of field and the desire for a more effective scan pattern led to the development of the holographic barcode scanner.

10.3 HOLOGRAPHIC BARCODE SCANNERS

The concept of holographic scanning has been around for three decades,[4] and during this time many different applications have been suggested,[5,6] but few have been demonstrated and an even smaller number have made it into the marketplace. A general review of holographic scanning and various applications can be found in the book by Beiser[7] as well as in previous volumes of optical engineering.[8]

Holographic barcode scanners first appeared commercially in 1980 with the introduction of holographic UPC scanners by IBM and Fujitsu. Today, Metrologie manufactures holographic scanners primarily for industrial applications. Such applications range from large depth of field (greater than 1 m [40 in]) overhead scanners, to high-density, high-resolution scanners for large aspect-ratio codes, to completely automated hands-off scanning tunnels for bulk-mail centers.

10.3.1 What is a Holographic Deflector?

Photography is a light-recording process in which a two-dimensional light-intensity distribution incident on a light-sensitive medium is recorded by that medium. In contrast, holography is a light-recording process in which both the amplitude and phase distribution of a complex wavefront incident on the recording medium can be recorded by that medium.

Holography, therefore, differs from photography in that it is able to record all of the information that is needed by the eye, or any other optical system, to interpret the full three-dimensional nature of the object.[9,10] This information is accessed when the recording (the "hologram") is illuminated by the proper light source—usually, but not always, a laser.

The most common form of hologram creates, when viewed, a three-dimensional image of a complex, three-dimensional object. The reduction of the three-dimensional object to a single point-source produces a special case of particular importance to deflection—a hologram that acts as a lens to focus an incident laser beam. This type of hologram is referred to as a holographic optical element (HOE).

The concept of holographic recording and reconstruction—more importantly, how it deflects light—can best be understood with reference to Figures 10.10 and 10.11. In Figure 10.10a, two wavefronts of equal intensity created from a laser are directed to overlap in some region of space where the recording is to be made. If the optical path difference from the point of beam separation to the region of overlap is within the coherence length of the

(a)

(b)   (c)
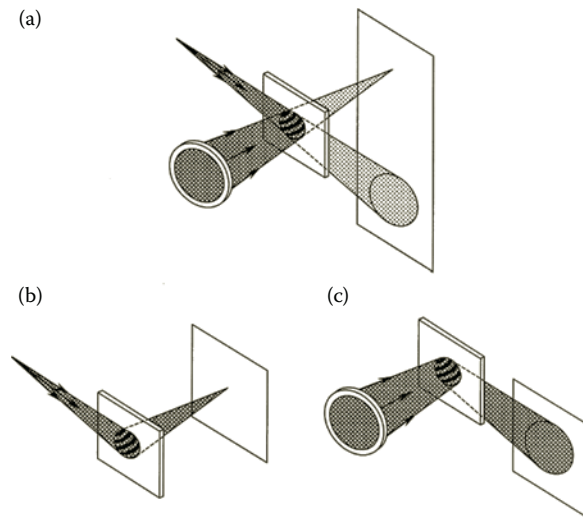
FIGURE 10.10
Simple holography: (a) recording the hologram; (b) reconstruction of the convergent wavefront (positive lens);
(c) reconstruction of the divergent wavefront (negative lens).
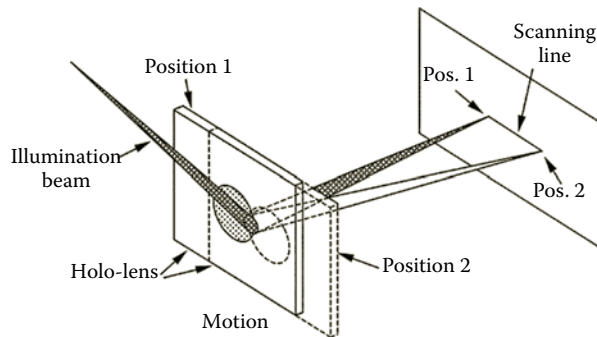
FIGURE 10.11
Principle of holographic deflection.

source, the resulting interference pattern will be stationary in both space and time and will
have high fringe contrast. The intensity distribution in these fringes can be exposed onto,
or more properly into, a suitable photosensitive medium such as a photographic emulsion.
After processing, the recording contains a variation in optical density, refractive index, or
optical thickness—sometimes a combination of all three—and is the hologram. When this
recording is repositioned and illuminated by one of the wavefronts, such as the diverg-
ing wavefront in Figure 10.10b, the structure at each point within the hologram diffracts
the illuminating light and creates a new wavefront that is identical to the original second
wavefront. In the case of Figure 10.10b the result is a converging and deflected wavefront.
This simple HOE is the equivalent of a positive or converging lens in combination with
a prism converting and deflecting light from a point object to a point image. The effi-
ciency and quality with which this wavefront conversion occurs are directly related to the
recording configuration and selection of recording material. In Figure 10.10c we see that

the equivalent of a negative lens is realized by illuminating the same HOE with the converging wavefront, thereby reproducing the original diverging wavefront. Holographic recording of complex multidimensional objects can be treated as the recording of a superposition of individual spherical waves from all the points in the object field.

By using a combination of small area illumination and HOE translation, the reconstruction geometry of Figure 10.10 can be used to create simultaneous light deflection and focusing. This is shown in Figure 10.11, where the HOE is initially located at position 1 and has a small subarea, on the right side, illuminated by a diverging wavefront that corresponds to the diverging reference-construction-beam of Figure 10.10a. The light is focused by this subarea of the HOE to the point in the image plane labeled "position 1" (Figure 10.11). This point in the image plane corresponds to the location, with respect to the displaced HOE, of the original convergence point of the converging construction wavefront in Figure 10.10a. As the HOE is translated, different subareas of the HOE are passed under the illuminating beam, and the reconstructed image point is caused to translate by the same distance, in this case to position 2. This is completely analogous to the deflection and focusing that would occur if a conventional lens were illuminated off axis with a collimated beam and the lens displaced normal to its optic axis. A continuous back and forth motion in either case produces the same motion in the focused spot.

In practice, however, a continuous rotary motion rather than a reciprocating motion is easier to implement. Higher scan speeds can be realized, and different holograms can be easily accessed. Consequently, most holographic deflectors consist of a number of unique HOEs placed circumferentially as sectors on a glass disc, as shown in Figure 10.12. Other materials can be used, and other geometries besides the disc geometry can be used,[8] but, for simplicity, we will restrict our discussion to the glass disc, which is the most common medium and geometry used in holographic scanners today. It should be noted that plane linear gratings, producing prismatic deviation without focal power, must be rotated to generate scanning. Translation of a plane grating in one direction will not produce scanning.

A holographic deflector disc, when properly illuminated by a laser and rotated about its axis of symmetry, can produce a complex variety of scanning laser beams. The optical and geometrical properties of each of these beams can be distinctly different from all of the others. This is the most important feature of holographic scanning. It is the major feature
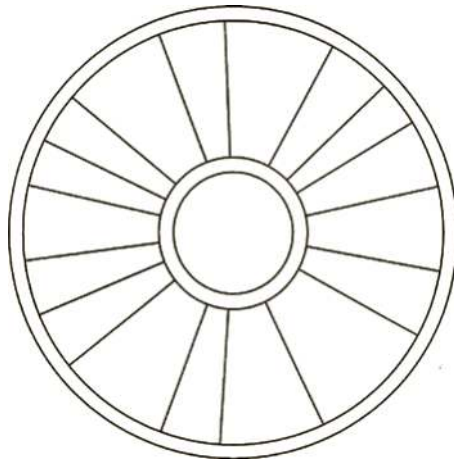
FIGURE 10.12
A holographic deflector disc.

that distinguishes it from conventional laser scanning technology and, in a barcode scanner, allows the introduction of capabilities that could not be readily achieved with conventional technology.[11]

The holographic disc works in the following manner. Each sector, or facet, of the holographic disc is a unique HOE of the type previously described—the holographic equivalent of a prism and lens combined. When a facet is illuminated by a laser beam, the beam is diffracted, or bent, by the facet and focused to some point in space (see Figure 10.13). The focal length and deflection angle are established during the holographic construction of the facet and may vary from facet to facet.

As the disc rotates, the deflected, focused laser beam scans. When the beam scans across a barcode, some of the diffusely reflected light will return to the facet that generated the scanning beam. The facet now acts as a light-collection lens, combined with a prism, to collect a portion of the reflected light and direct it toward a photodetector.[12]

10.3.2 Novel Properties of Holographic Barcode Scanning

The use of holography in barcode scanning allows the introduction of scanning concepts that are not available to the designer of conventional barcode scanners, at least not in any economically practical design. Such concepts as multiple focal planes, overlapping focal zones, variable light-collection aperture, facet identification, and scan-angle magnification allow holographic scanning to bring to barcode scanning some significant design and performance capabilities.

A conventional barcode scanner contains a lens for focusing the laser beam, a device for deflecting the laser beam, and some optics for collecting a portion of the laser light reflected from the barcode and focusing it onto the photodetector. In a holographic scanner, all of these properties—focusing, deflecting, and light collection—are contained in the holographic disc. As indicated earlier, these properties may be different in each sector of the holographic disc; thus, a 16-sector holographic disc, for example, would contain 16 unique optical systems. Each of these systems would have its own focal length, scan angle, and light-collection aperture. One revolution of such a holographic disc would produce the equivalent of scanning with 16 different scanners.

Because each facet of the holographic disc may be different, with its own combination of focal length, deflection angle, and facet area, then one complete rotation of the disc will create multiple scan lines with multiple deflection angles, multiple focal lengths, and multiple light-collection systems. This enables the holographic scanner to introduce some novel operational characteristics.
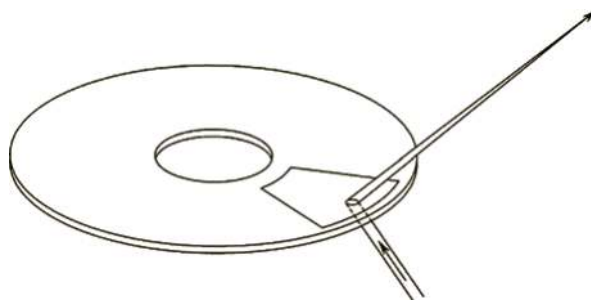
FIGURE 10.13
Light deflection and focusing by a holographic disc.

One of the major advantages of using a holographic disc in a barcode scanner is that it can provide a much larger depth of field than would be attainable with a conventional, single-focal-length, barcode scanner. In order to understand this point, we need to briefly review the subject of depth of field.

10.3.3 Depth of Field for a Conventional Optics Barcode Scanner

In barcode scanning, depth of field is the distance along the laser beam, centered around the focal point of the scanner, over which the barcode can be successfully scanned. The spot size profile of a laser beam along its direction of propagation is established by the beam waist diameter and the wavelength of the laser light source. The depth of field of a barcode scanner employing such a beam depends on the size of the minimum bar width in the barcode being read and also on the resolving ability of the scanner electronics. For a given resolving ability there will be a normalized spot-size definition, or resolution criterion (C in Equations 10.3 through 10.5), based on the assumed Gaussian intensity profile of the laser beam. This resolution criterion simply defines the relative intensity level at which the beam diameter is measured. A commonly used criterion is the $1/e^2$ beam width (13.5% intensity level, $C = 0.135$). Typical resolving abilities of scanners range from the 50% to 70% intensity levels, or possibly even higher. Once all these beam, code, and electronic parameters have been established, there is little that can be done in a conventional barcode scanner to increase the depth of field.

Figure 10.14 illustrates the concept of depth of field. The lens in the scanner focuses the laser beam to a relatively small spot size at the focal point. The diameter of the beam at the focal point is determined by the focal length of the lens, the diameter of the beam at the lens, and the wavelength of the laser being used. When the optical system of the scanner is properly designed, the minimum spot size will be somewhat smaller than the minimum bar width and will therefore be able to successfully scan the barcodes. As one moves the barcode to either side of the focal point, toward or away from the scanner, the spot size increases as the beam becomes out of focus. Eventually, a point is reached where the out-of-focus spot size is larger than the minimum bar width in the barcode. When this occurs, the barcode can no longer be successfully scanned by the beam. The distance between the two points to either side of the focal point where the limit of scanning capability occurs is, by definition, the depth of field.
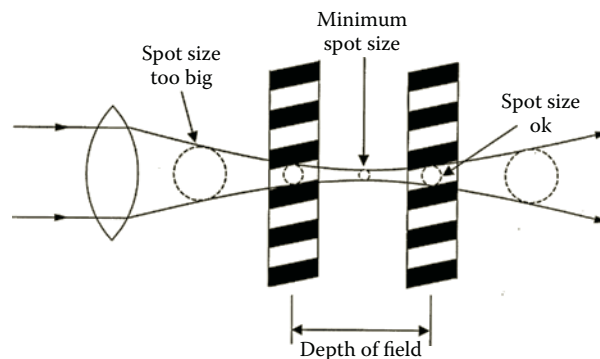
FIGURE 10.14
Depth of field for a conventional optical system.

The major factors determining the depth of field are the spot size at the focal point, the wavelength of the laser, and the minimum bar width. (For convenience, we will assume throughout this discussion that the minimum space width is the same as the minimum bar width.) Using the notation of Dickson[13] for the variation in beam radius of a propagating Gaussian beam, the $1/e^2$ beam radius, r, at a distance, z, from a beam waist of $1/e^2$ radius r_0 is given by

$$r = r_0 \left[1 + \left(\frac{1z}{pr_0^2} \right)^2 \right]^{1/2} \tag{10.2}$$

The beam radius, r_c, for a different resolution criterion, C, is given by

$$r_c = rK = r\sqrt{\frac{-\ln(C)}{2}} \tag{10.3}$$

Similarly, the waist size at that resolution criterion is scaled down by $r_{0c} = r_0 K$. Equations 10.2 and 10.3 can be combined and rearranged to express the depth of field for a given beam radius, r_c, resolution criterion, C, and waist radius, r_{0c}, as

$$\text{DOF} = \Delta z = 2\,|z| = \frac{-4p}{1\ln(C)} \sqrt{r_{0c}^2 r_c^2 - r_{0c}^4} \tag{10.4}$$

It can be shown that, for any resolution criterion, the depth of field given by Equation 10.4 is maximized for a given minimum bar width when the focused spot size, as measured by the particular resolution criterion being used, is equal to the minimum bar width divided by $\sqrt{2}$. Applying that condition requires the substitutions of $2r_{0c} = w_{min}/\sqrt{2}$ and $2r_c = w_{min}$ into Equation 10.4, which then reduces to

$$\Delta z = \frac{-pw_{min}^2}{2\,1\ln(C)} \tag{10.5}$$

At a wavelength of 650 nm, typical of VLDs being used in scanners today (2004), and assuming a reasonable resolution criterion of $C = 0.6$ (60%), Equation 10.5 can be approximated as

$$\Delta z = \frac{w_{min}^2}{8.3} \tag{10.6}$$

where Δz is the depth of field in inches when the minimum bar width is in mils or as

$$\Delta z = \frac{w_{min}^2}{210} \tag{10.7}$$

where Δz is the depth of field in millimeters when the minimum bar width is in microns.

For example, if the scanner is optimally designed to read a barcode that has a minimum bar width of 8 mils (200 μm), the depth of field will be 7.7 in (200 mm). Note that the above equation tells us that the depth of field is strongly dependent on the size of the minimum bar width to be read. Therefore, a small minimum bar width is always accompanied by a small depth of field.

Furthermore, if the scanner is not optimally designed for the minimum bar width to be read, the depth of field will be smaller than it could be. This will be true whether the scanner is optimally designed to read either a higher density barcode or a lower density barcode. In addition, if the scanner uses a laser with a different wavelength, the depth of field will be multiplied by the ratio of the wavelength for which the design is optimized to the wavelength of the laser being used. (This assumes that the new laser has its minimum spot size focused to the same size as the original design.)

There is very little that can be done in a conventional barcode scanner to increase the depth of field. It is possible to use an autofocus scanner. Autofocus scanners have been designed and built for the industrial scanning market by Accu-Sort Systems (Telford, PA). However, a problem with such scanners is that the reaction time of the autofocus system has to be very fast to accommodate fast-moving items on a conveyor system. Because all autofocus systems today require mechanical movement of some of the optics of the scanning system, the reaction time may not be fast enough, depending on the application. In addition, such systems will add cost and complexity to the scanner.

One could also add a supplemental optical element to the scanner that could move into position in the laser beam path to change the net focal length of the scanner. This moving element would allow, for example, two different focal lengths to be selected. In practice, this approach would allow only two or three different focal lengths to be selected, giving only a slight increase in the depth of field.

Furthermore, in either an autofocus system or a dual or triple-focal-length system, only the focal lengths can be easily changed. Ideally, one should also change the aperture of the scanner as the focal length is changed. This would maintain a constant level of light collection over the full range of focus, thereby optimizing the performance of the scanner over the full range of readability. However, rapid variation of the light-collection aperture is difficult to accomplish in a conventional barcode scanner.

10.3.4 Depth of Field for a Holographic Barcode Scanner

The use of holography in a barcode scanner would allow the introduction of a true multifocal-plane scanner with a variable light-collection aperture. The way this would be accomplished in a holographic scanner is illustrated in Figure 10.15.

Figure 10.15 shows focusing of the laser beam by two consecutive facets on the holographic disc. Each facet will exhibit a conventional depth of field as established by the focal length of the facet, the beam diameter at the disc, and the wavelength of the laser. Notice, however, that the two facets are focused at different distances from the disc. Therefore, while each facet has only a conventional depth of field, the combined depth of field of the two facets is twice as great as for either facet alone, assuming that the focal lengths are chosen so that the end of the depth of field for facet 1 coincides with the beginning of the depth of field for facet 2.

Thus, with only two focal planes, the holographic disc can double the depth of field of a conventional, nonholographic scanner.

If the disc is designed so that all of the facets are focused at different distances, then a much larger overall depth of field can be achieved. For example, if the minimum bar width
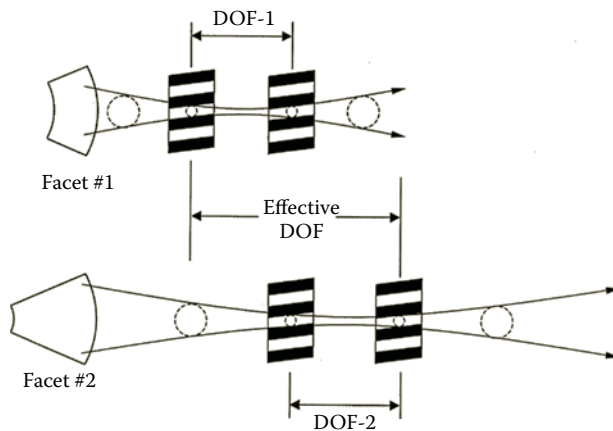
FIGURE 10.15
Combined depth of field for two holographic facets.

to be scanned is 0.2 mm (8 mils), a conventional, single-focal-length scanner would have a depth of field of approximately 200 mm (8 in). However, a properly designed holographic scanner could provide as much as 800 mm (32 in) depth of field for the same 0.2 mm minimum bar width code by using only four facets.

Even greater depth of field can be achieved with more facets; however, there are diminishing returns to this route to greater depth of field. Ideally, we would like to be able to control the focused spot size in each focal plane; however, this would require automatic aperture adjustment of the outgoing beam (similar to autofocus), which would prove mechanically difficult and very costly. As a compromise, typically, the centermost focal plane is optimized for maximum depth of field, which causes the other focal planes to be close to optimum as well.

Another limit to the continuing growth of depth of field is the very large distances that are eventually encountered as facet focal lengths increase. While theoretically this is not a problem for the resolution of the outgoing beam profile, it does present a greater challenge to light collection, which will be discussed later.

10.4 OTHER FEATURES OF HOLOGRAPHIC SCANNING

There are other novel features of holographic scanning that are not as obvious as the ability to provide a large depth of field. The major features of holographic scanning are:

1. Multiple focal planes
2. Overlapping focal zones
3. Variable light-collection aperture
4. Facet identification and scan tracking
5. Scan-angle magnification

We have already discussed the multiple-focal-plane feature and the large depth of field that it provides. Let us now examine the other features to see what they are, how they are produced by the holographic disc, and what capability they provide.

10.4.1 Overlapping Focal Zones

We showed in the previous section how holographic scanning can provide a large depth of field by designing the holographic disc so that the depth-of-field region of each successive facet was contiguous with the depth-of-field regions of the facets immediately preceding it and following it. This may not, in practice, be the best disc design. It may, in fact, be better to design the holographic disc so that the focal point of one facet coincides with the limit of the depth of field for the preceding and following facets, resulting in an overlapping focal zone design. The reason why this design may be superior is explained in the following paragraphs.

One of the major contributors to decoding problems in a barcode scanner is the existence, or creation, of noise in the so-called quiet zone, the white, or clear, region immediately preceding and following the barcode. One of the contributing factors to noise in this region, and throughout the barcode, is substrate noise, or paper noise. Paper noise occurs when the size of the focused spot of the scanning laser beam is about the same as the size of the granularity of the substrate material. Paper fibers, for example, can be as large as 0.1 mm (4 mils). For very coarse paper or cardboard, the fiber size can be even greater. For nonpaper substrates, such as for barcodes etched into plastic or metal, the granularity can be greater still.

If a barcode on a noisy substrate is scanned at the focal point of a scanning laser beam, the small, in-focus spot will "see" the granularity of the substrate material. This will introduce paper noise on the return light signal that will, in turn, lower the probability of achieving a successful read.

While noise could, in general, be reduced with low-pass electrical filtering, the filter properties would have to be altered for each facet to correct for the differences in spot velocity. That is, a low-pass filter designed to remove noise from the short-focal-length facets would, at the same time, filter out the barcode signals from long-focal-length facets. Electrical filtering does not appear to be a practical solution for large depth-of-field scanners.

The solution to this problem is to scan the noisy barcode with a slightly out-of-focus spot. This spot will be larger than the in-focus spot, but still small enough to read the barcode. This larger spot will, while scanning the barcode, act as a filter to smooth out the surface roughness, effectively lowering the paper noise and increasing the probability of achieving a successful read.

Figure 10.16 shows the analog photodetector signal for a noisy barcode when scanned (a) by an in-focus spot, and (b) by a slightly out-of-focus spot. The noise on the signal from the in-focus scanning spot, is apparent. The resultant reduction in the noise level due to scanning the same barcode with a slightly out-of-focus scanning spot is equally apparent.

By overlapping the focal zones of the individual holographic facets, as shown in Figure 10.17, we can guarantee that all barcodes will be scanned by both an in-focus scanning spot and one or more slightly out-of-focus scanning spots. The slightly out-of-focus spots will be small enough to read the barcodes, but large enough to smooth out the substrate noise.

This in-focus/out-of-focus capability, which would be difficult to implement with conventional scanning technology, is relatively simple to introduce with holographic scanning. One merely selects, during the master holographic disc design phase, the focal length for each of the facets that guarantees the desired amount of focal zone overlap.

Scanning with in-focus spot

Scanning with out-of-focused spot
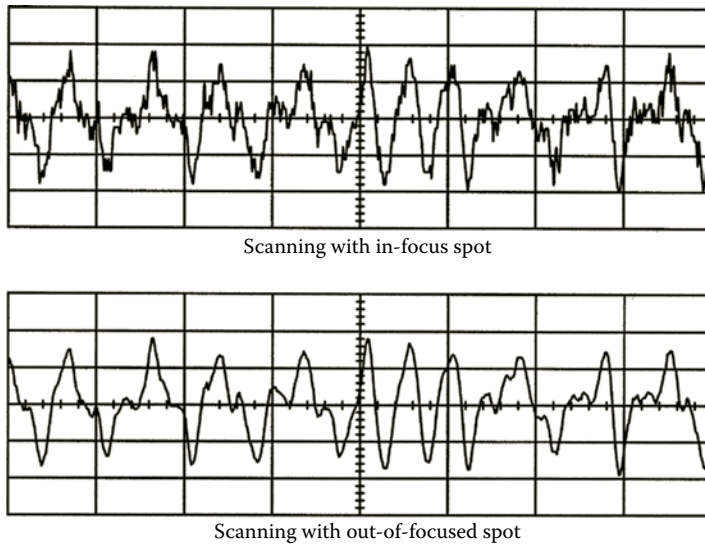
FIGURE 10.16
Photodetector signals for in-focus and out-of-focus scanning laser beams.
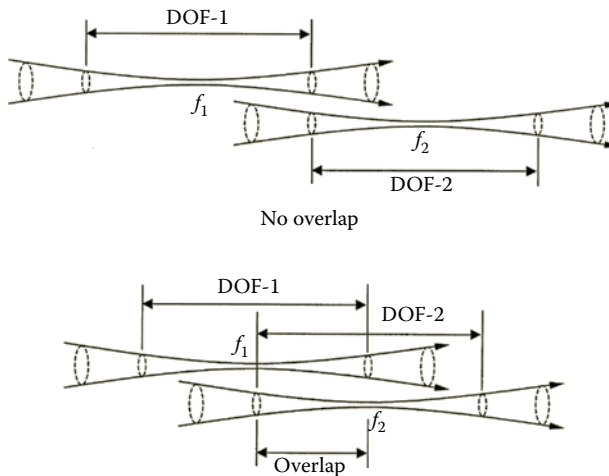
No overlap

Overlap

FIGURE 10.17
Overlapping focal zones of two holographic facets.

10.4.2 Variable Light-Collection Aperture

There is more to successfully achieving a large depth-of-field scanner than simply providing multiple focal planes. If, for example, one designs a scanner with a 1-m depth of field where the optical throw (closest reading distance) is 200 mm (8 in) and the range (farthest reading distance) is 1200 mm (47 in), then the variation in the light level returned to the detector, for barcodes with identical reflection characteristics, will be 36:1, the square of the ratio of the far and near distances. This places a severe dynamic range requirement on the analog electronics in the scanner. The problem is worse in practice since other factors, such as label skew and label reflectivity variations, also affect the amount of light returned.

In order to reduce the variation in the light level of the return light in a multiple-focal-plane scanning system, it would be desirable to vary the light-collection aperture to compensate for changes in the distance to the barcode being scanned. One could then use a relatively small aperture for a near-focus scan line and a relatively larger aperture for a far-focus scan line.

Holographic scanning allows one to do exactly that. Figure 10.18 shows a holographic scanning disc designed for focus distances ranging from 1000 to 1680 mm (39.5 to 66 in). Notice that the light collection area of each facet is different. The facet with the shortest focal length has the smallest light collection area while the facet with the longest focal length has the largest light collection area. The light collection area of each of the remaining intermediate facets is a direct function of its focal length.

This difference in light collection area for the near and far facets of the holographic disc allows the light collection to be approximately uniform over the total depth of field of the scanner. This is a major advantage in obtaining decoding accuracy over a large depth of field for a barcode scanning system.

10.4.3 Facet Identification and Scan Tracking

Note that the disc design shown in Figure 10.18 incorporates a gap in the outer annulus between two of the holographic facets. This gap may be transparent or opaque, depending on the application, and is referred to as the home-pulse gap. Because the outgoing laser beam is incident on the disc in this outer section, a detector placed in the proper location above the disc can sense this gap by measuring the laser power incident on the detector. As the gap passes over the laser the change in power recognized by the detector generates a home pulse in the analog signal. With this information embedded in the signal we can determine the rotational speed of the disc and, thereby, where on the disc the laser beam is currently incident. This knowledge can then be used to determine what facet is currently scanning, and even where in that facet the beam is located. This method of facet
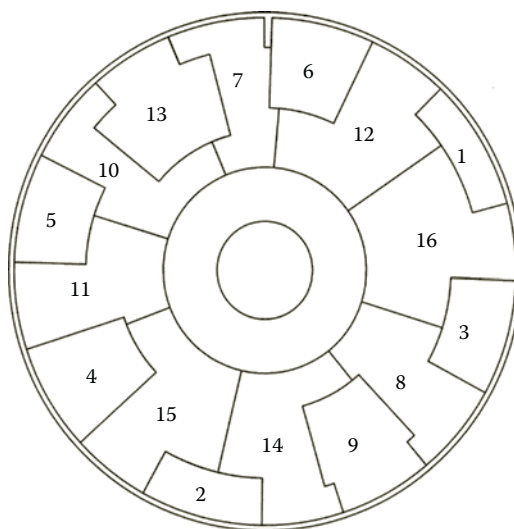
FIGURE 10.18
The Metrologie Penta holographic scanning disc showing large variation in facet areas.

identification can be used in several ways to improve the decoding accuracy of the scanner, as well as provide for additional features.

If we knew, for example, that we were on a short-focus facet, we could decrease the electrical gain in the analog electronics. If we were on a long-focus facet, we could increase the gain of the analog electronics. This electronic automatic gain control (AGC) would add to the already existing optical AGC, introduced by the variable light-collection aperture, to further improve the decoding accuracy.

We could also vary the internal clock rate from facet to facet to improve resolution. Because the scanner is an angular scanner, the linear velocity of the scanning beam will vary directly with the distance from the scanner. The bit rate seen by the detector while scanning with a long-focus facet would be greater than the bit rate seen when scanning with a short-focus facet, assuming that the code density is the same in both cases. By making the clock rate vary from facet to facet to maintain the optimum clock rate for a given bit rate, we could, once again, improve the decoding accuracy of the scanner.

An additional capability made possible by this facet identification feature is scan tracking, for which several patents have been filed by Metrologie Instruments, Blackwood, NJ (U.S. patent numbers 6,382,515 B1; 6,457,642 B1; 6,517,004 B2; and 6,554,189 B1). Knowing precisely where the incident beam is striking each facet at each instant in time allows indirect determination of where the item being scanned is located in a given three-dimensional, spatial reference system. This knowledge is extremely useful in fully automated systems where little human interaction is desired, such as in scan tunnels in bulk-shipping centers. With the location of a scanned package identified, other automated mechanical systems can then redirect packages to their intended destination.

The reason holography lends itself so well to scan tracking is that the facets on a holographic disc are easily repeatable in the manufacturing process with high precision. Obtaining good repeatability of the deflector with mirrored scanning systems is more difficult.

10.4.4 Scan-Angle Multiplication

Holographic scanning discs used in barcode scanners are frequently designed to be illuminated with a collimated beam incident normal to the surface of the holographic disc. This illumination geometry provides a scanning spot that is free from aberrations across the entire scan line because the relative incidence geometry with respect to the hologram does not change as the disc rotates. This is a special case of the more general aberration-free illumination geometry for rotationally symmetric systems in which the designed illumination beam is a converging spherical wavefront, converging toward a point located on the rotational axis of the disc. (A normally incident collimated beam has a point of convergence on the axis at infinity.) Under these conditions, the illuminating wavefront remains unchanged with respect to the hologram, always converging to the hologram's design convergence point even as the disc is rotated. Because the HOE was designed to produce an aberration-free diffracted beam when the incidence wavefront converges to that point, the diffracted beam remains aberration-free throughout the motion of the hologram. In essence, the playback (illumination) beam remains identical to the reference (recording) beam, which is the condition for zero aberrations.

If the holographic disc is designed to be illuminated with a collimated beam inclined at a non-normal angle, then some amount of aberrations will be introduced in the scanning beam. Each facet of the disc can be designed to still provide zero aberrations at the center of its corresponding scan line, but there will always be aberrations introduced as the disc

rotates because of the resulting mismatch between the recording and playback wavefronts. The amount of the aberrations will be dependent on the amount of rotation away from the center of the facet and the amount of tilt of the collimated incident beam.

There is, however, one advantage to tilting the incident beam. It can be shown that a tilted, collimated reference beam will provide a greater scan-angle multiplication factor relative to an untilted, collimated reference beam geometry. A precise determination of the multiplication factor requires the use of a computer program because of the interdependence of the diffracted beam elevation angle (β in Figure 10.19) and the rotation angle (ϕ_{rot}). A first-order approximation, accurate to a few percent for holograms whose construction geometry is not too extreme, can be obtained from the following simple relationship. The variable f represents the focal length of the holographic lens (facet). The other terms and the geometry are defined in Figure 10.19.

$$f_{\text{scan}} = f_{\text{rot}}\left(\frac{r}{f}\cos\gamma + \cos\alpha + \cos\beta\right) = f_{\text{rot}}\left(\frac{r}{f}\cos\gamma + \frac{1}{d}\right) \tag{10.8}$$

where d represents the grating spacing of the hologram and the angle γ is determined from Equation 10.9:

$$\sin\gamma = \cos\beta\sin\theta_{\text{skew}} \tag{10.9}$$

The multiplication effect of the tilted reference beam is due to the $\cos\alpha$ term in Equation 10.8; for normal incidence this term goes to zero.
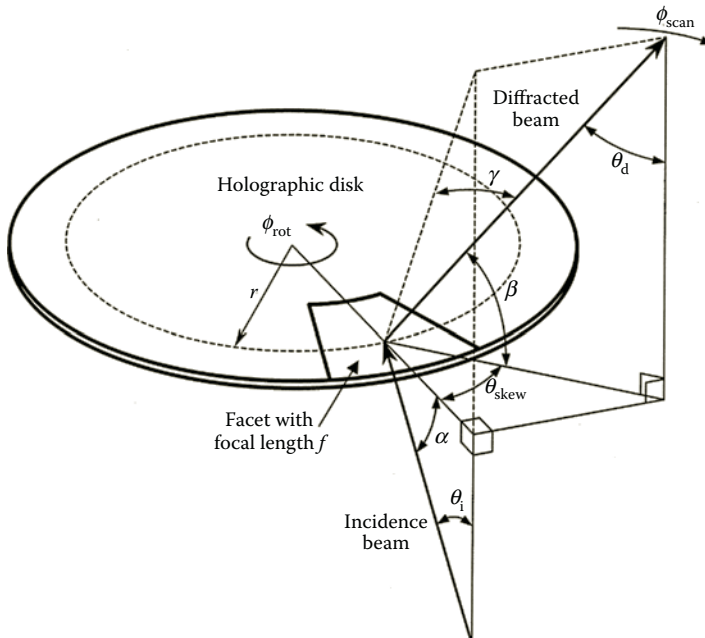
FIGURE 10.19
Scan-angle magnification parameters.

As an example of the effect of the tilted reference beam, consider a holographic facet with f = 350 mm (13.8 in), θ_{skew} = 40°, β = 66°, and r = 72 mm (2.8 in). For a normal reference (and incidence) beam (α = 90°), ϕ_{scan} = 0.605 ϕ_{rot}, while for a tilted reference resulting in an incident beam tilted at an angle of 22° relative to the normal (α = 68° in Figure 10.19), ϕ_{scan} = 0.980 ϕ_{rot}. The relative multiplication factor obtained by using the tilted incident beam is 1.62. This is a significant amount of scan-angle multiplication.

Scan-angle multiplication factors of this magnitude provide significant design flexibility. For a given scan length one could: make the disc smaller to produce a more compact unit; keep the disc the same size and add facets to provide more scan lines or to generate a more complex scan pattern; move the disc closer to the window to increase the light collection efficiency of the individual facets; or some combination of all three options. One may, of course, elect to use scan-angle magnification to just generate a longer scan line.

The incident-beam tilt angle, the aberrations, and the scan-angle multiplication factor are interrelated: the smaller the amount of tilt of a collimated incident beam, the smaller the aberrations; conversely, the larger the tilt, the larger the scan-angle multiplication factor. By careful selection of the tilt angle of the incident beam, one can get a significant amount of scan-angle multiplication while maintaining an acceptable amount of aberrations. The amount of aberrations introduced can be determined with a suitable ray tracing program.[11] Acceptable levels of aberrations are established by the individual application.

If, on the other hand, the incidence angle is dictated by some other mechanical constraint in the scanner system, the aberrations can be controlled, as alluded to before, by controlling the designed convergence of the incidence beam. The closer the convergence point is to lying on the axis of disc rotation, the less severe the aberrations will be. This, however, has the same trade-off as manipulating the incidence angle. As the convergence of the incidence beam becomes shorter, the focal length, f, of the facet will become longer, perhaps even reaching or "passing" infinity and becoming negative. As this happens, the first term in Equation 10.8 gets smaller or, in the case of negative f, becomes negative, thereby canceling out some of the scan multiplication factor of the other two terms. Whether or not this is a disadvantage depends on what is more important to the application, scan-line length or spot quality.

10.5 HOLOGRAPHIC DEFLECTOR MEDIA FOR HOLOGRAPHIC BARCODE SCANNERS

All holographic barcode scanners today (2004) use a rotating circular disc as the substrate for the recording medium. Other geometries have been considered, but, for barcode reading applications, the disc geometry offers a number of manufacturing advantages and is generally less expensive.

All of today's holographic barcode readers also operate only in the transmission mode. It would not be impossible to develop a reflective holographic barcode scanner, but the transmission mode provides a simpler design, an easier manufacturing process, and less susceptibility to disc wobble.[8]

There are two general types of media suitable for recording HOEs on a disc surface for use in a holographic barcode scanner: surface-relief phase media, such as photoresist, and volume-phase media, such as bleached silver halide and dichromated gelatin (DCG). There are advantages and disadvantages associated with both types of media. For a

more general review of the wide variety of holographic recording materials, see work by Bartolini[14] and Smith.[15]

The major factors influencing the selection of the type of holographic medium are manufacturing cost, diffraction efficiency, and, as will be described, the scan pattern density.

10.5.1 Surface Relief Phase Media

There are only two significant surface-relief phase media presently being used for holographic scanning discs—photoresist, and plastic copies made either directly from photoresist or from intermediate copies of the photoresist. As will be discussed later, this latter type is probably the least expensive of all to manufacture in a high-volume process and, from a purely cost consideration, this medium would appear to be the best choice for a holographic deflector disc.

The major disadvantage of the low-cost surface-relief material is that when using a simple mechanical replication process the resulting diffraction efficiency will be relatively low, on the order of 30%. This is because a much larger grating aspect ratio (depth vs. spacing) is required to produce high efficiency. Mechanically releasing such a high-efficiency copy from the master can be very difficult, often damaging the master and the copy. This is a major drawback in barcode scanner applications. The low efficiency means that a higher power laser must be used to get sufficient laser power onto the barcode symbol to obtain a good reading. The greater cost of the higher power laser may offset the lower cost of the disc.

It is possible to get high diffraction efficiency using a surface-relief medium for light that is incident on the disk in an S-polarized orientation.[16,17] It is not possible at the present time, however, to mechanically replicate these high-efficiency surface-relief holograms because of the aforementioned aspect ratio of the relief profile. (An example of such a high aspect ratio is shown in Figure 10.20.) This means that original holograms, not inexpensive copies, would have to be used. In some holographic deflector applications, this is acceptable. In general, however, this is not an acceptable alternative for a holographic barcode scanner due to the higher cost of the discs and their greater susceptibility to physical damage. (Surface relief holographic discs cannot be protected by a cover glass since an index-matching adhesive would effectively eliminate the surface relief structure.)

The low diffraction efficiency of the mechanically replicated surface-relief material also means that the collected light will be low in a system employing the holographic facets of the disc in a retroreflective mode. This loss in collected light cannot be compensated by increasing the laser power further because of the limitations established by the federal laser safety standards. The only means left to compensate for the low diffraction efficiency in the collected light is to increase the size of the facets on the holographic disc. However, this reduces the total number of facets, hence the total number of independent scan lines, and the subsequent scan pattern density. In some barcode scanning applications, this may be an acceptable trade-off. In other applications, such a trade-off may be unacceptable.

For example, in supermarket/retail barcode scanning applications, the depth-of-field requirement is relatively moderate, so that a holographic scanner with as few as two focal planes can provide adequate performance. However, in industrial barcode scanning applications, the required depth of field may be as large as 1 m (40 in), or more, for medium density barcodes (barcodes with a minimum bar width on the order of 0.3 mm). This kind of depth-of-field requirement can only be met with a scanner that can provide a large number of focal planes. A holographic scanner can be designed to provide this capability, but the number of independent facets on the scanning disc must be as large as possible.
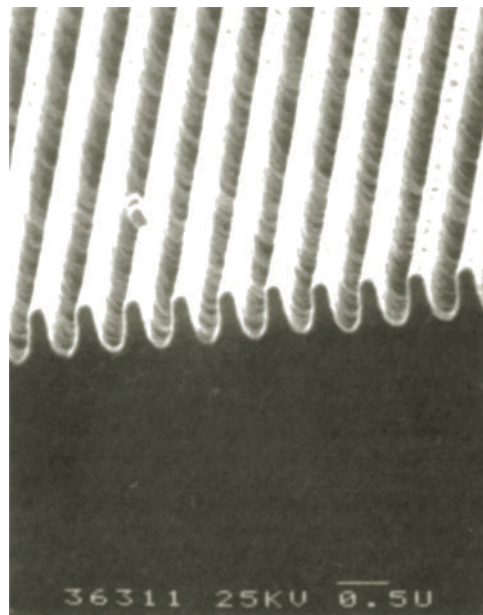
FIGURE 10.20
Surface-relief hologram in positive photoresist.

Therefore, any reduction in the number of facets imposed by a low diffraction efficiency recording medium will reduce the depth of field.

For many retrocollection scanning applications even the "high-efficiency" holograms mentioned above do not have a high enough diffraction efficiency for good light collection. In order to eliminate specular reflection noise in a barcode scanner, a polarizer is often placed in front of the light detector. This is done so that one linear light polarization can be used for the outgoing, scanning beam while the other, perpendicular light polarization will return to the scanner, pass through the polarizer, and be detected. This presents a problem for surface-relief holograms in that an even higher aspect-ratio relief profile is necessary to diffract P-polarized light than is necessary to diffract S-polarized light. As a result, while the outgoing efficiency could produce a strong scanning beam, the light collection ability would be relatively weak (or vice versa) requiring the same kind of trade-offs as discussed above.

Even so, despite the relatively low diffraction efficiency of the holographic discs produced by mechanical replication, which is a surface relief process, the low cost of such discs make them very attractive in supermarket and retail barcode scanners in which component cost is a major factor.

10.5.2 Volume Phase Media

Volume phase materials are capable of very high diffraction efficiencies, on the order of 90% or more. Such high efficiencies means that the individual facets on the disc can be relatively small, even when the disc is used in a retroreflective mode. This means that there can be more facets on the disc, which, in turn, means that the scanner can generate more independent scan lines, resulting in a larger depth of field and/or a more complex

scan pattern. The higher diffraction efficiency also means that a lower power laser can be used to generate the scan lines.

There are a number of materials that are suitable for use as volume phase materials in holographic scanners. The first material that comes to mind is bleached silver halide. In this process, the absorptive structure in a photographic emulsion hologram is chemically converted from metallic silver to a material having a refractive index different than the surrounding gelatin matrix.[18–21] For example, the silver may be rehalogenated by exposure to bromine vapor. Holograms created with this material can have high diffraction efficiencies, on the order of 80% or more.[22,23] Processing is relatively simple, and the holograms are reasonably stable. There are a few bleaches, however, that leave reaction products behind in the emulsion. Some of these products are photosensitive and exhibit printout effects, particularly when subjected to intense ultraviolet irradiation. Nevertheless, moderately efficient holograms can be realized, and the advantages of photographic emulsions, such as extended spectral response and speed, may be exploited. One practical disadvantage associated with this material is that the discs must be coated and sensitized by one of the major companies producing general photographic materials. Such companies are usually reluctant to stock odd substrate shapes (like discs) and coat them to a user's specifications, in quantities that are, for them, relatively small. This creates a very real sourcing problem.

The next most attractive volume phase material is photopolymer.[24,25] Cross-linking in these materials is produced when they are exposed to light of relatively short wavelength, blue to ultraviolet. When a photopolymer is exposed to a holographic fringe pattern at the proper wavelength, the periodic variation in light intensity of the fringe pattern produces a corresponding periodic variation in cross-linking in the polymer. When developed, the photopolymer will exhibit a periodic variation in refractive index corresponding to this periodic variation in cross-linking. These materials are relatively stable when exposed to normal levels of ambient light, heat, and humidity.

The main drawback to these materials has been their relatively small change in refractive index, Δn, produced by exposure to light and subsequent processing. This means that, in order to get high diffraction efficiency, the thickness of the photopolymer coating has to be on the order of 50 microns (2 mils). Such a large thickness would make the holographic deflector disc very sensitive to the Bragg angle. That is, very slight deviations in the angle of incidence of the reconstruction beam in the scanner would cause severe reductions in the disc diffraction efficiency. Deviations on the order of 1/4° (4.4 mrad) could cut the diffraction efficiency in half.[26] This is generally unacceptable in a product where the total angular manufacturing tolerances could easily be this large. Furthermore, the anticipated mode of operation could cause the effective angle of incidence to vary by 1/4° (4.4 mrad) during disc rotation.

The DuPont[27] and Polaroid[28] photopolymers have exhibited refractive index changes that are much greater than those of previous photopolymers. Δn values approaching the values obtainable with DCG (nearly ten times as great as earlier photopolymer Δn values) have been obtained. These materials have great potential for use as a recording medium for holographic deflectors used in barcode scanners, since high diffraction efficiency should be achievable in relatively thin coatings, on the order of 5 microns (200 μin).

The volume phase material that has, up to now, been the most successful material for use in holographic deflectors for barcode scanners is DCG.[29–31] The major advantage of this material, as a medium for holographic deflector discs, is that its diffraction efficiency can be very high (>90%) in a relatively thin (3–5 microns or 120–200 μin) coating because of its high Δn (0.10–0.15 or greater). This means that DCG can have, simultaneously, high diffraction efficiency and very low Bragg-angle sensitivity. This is a significant advantage from both a manufacturing and an application standpoint.

The major disadvantage of DCG is that it is extremely sensitive to moisture. Holograms made with DCG must be sealed to protect them from environmental moisture.

From the standpoint of the development of a barcode scanner, there is one other disadvantage to DCG. Although it has been around a long time, DCG is the least understood of all the holographic recording media. There are at least three theories that claim to explain the mechanism of image formation,[32–34] and there are as many recipes for processing DCG as there are authors writing on the subject. Many of them start with gelatin that is already coated on photographic plates,[35] a procedure which is unacceptable for the same reasons that bleached silver halide is unacceptable: the sourcing problem.

In most large corporations, one will also find considerable resistance to the use of DCG. Most chemists feel comfortable with well-understood inorganic materials, such as silicon, and the more traditional organics, such as photoresist, photopolymer, and so on. DCG is an organic material whose properties are poorly understood and relatively unpredictable. Gelatin is, after all, made from the skins, bones, and connective tissues of animals. Its properties can vary depending on what the animals ate or where they were raised.

Nevertheless, because of its excellent holographic qualities, DCG is one of the best recording materials for holographic deflectors used in barcode scanners. It is relatively stable when exposed to normal ambient temperatures. However, it is extremely moisture sensitive and must be sealed to protect it from normal ambient humidity.

DCG is generally sensitive only to the short wavelength portion of the visible spectrum, $\lambda < 520$ nm (20.5 µin), and although it is possible to sensitize it to the red end of the spectrum[36–40] only moderate success has been achieved. The primary problem has been removal of the residual sensitizing dye to give a complete phase structure. For barcode scanning applications, where the light source in the scanner is generally a visible laser diode (VLD) with a wavelength somewhere in the red region of the spectrum, unsensitized DCG cannot be used to make the master holograms at the operating wavelength. Because of this, the DCG holographic disc must be made as a copy of masters formed in one of two ways.

In the first method, the wavelength that will be used in the scanner is used to construct the masters with a material that is sensitive to that region of the spectrum, such as silver halide. This allows a relatively simple optical setup that will produce an aberration-free hologram. Typically, DCG submasters (in the form of a submaster disc, or individual submasters) are made from the masters, providing greater efficiency to the production process.

In the second method, the masters are made directly in DCG using a wavelength within its spectral sensitivity range, such as 488 nm (19.2 µin), one of the high-power lines of an argon laser. The difference of the exposure wavelength from that which the scanner will employ requires some kind of aberration-compensating optics in the master-exposure setup. This makes the setup more complicated, but the result can be a higher-quality hologram since the submaster step of the process is not needed, and the use of additional aberration-correcting optics will maintain the essentially aberration-free performance of the final hologram.

Whichever method is used, the DCG copy disc can then be made using any wavelength to which the DCG is sensitive. There will be no aberrations introduced in the copy process, regardless of the wavelength used in the copy process. We will have more to say about this in the section on disc fabrication.

DCG is processed in a sequence of alcohol/water baths of varying concentrations of alcohol and varying temperatures. Times, temperatures, and concentrations vary, depending on whose process is used.

Diffraction efficiencies obtained with DCG approach the theoretical limits for volume phase materials. Efficiencies greater than 90% can be readily obtained. The only things limiting the diffraction efficiencies of a sealed DCG holographic disc are reflections off the glass surfaces and absorption and scattering losses in the gelatin. If antireflection (AR) coatings are not employed, the Fresnel reflections at the air/glass interfaces will cause the primary losses. This may limit maximum efficiency to about 70%, depending on polarization. If, however, good AR coatings can be applied, there will still be some minor Fresnel losses at the internal gelatin/glass interface and some small amount of scattering and absorption in the gelatin. If the film properties can be controlled enough to keep scattering to a minimum, then efficiencies of 95% are achievable. This control, however, is the key, and is sometimes easier said than done. Even so, efficiencies in excess of 85% are relatively easy to maintain.

10.6 FABRICATION OF HOLOGRAPHIC DEFLECTORS

10.6.1 The DCG Holographic Disc

One possible DCG holographic disc fabrication process using DCG masters (no submasters) is shown schematically in Figure 10.21. Each facet of the holographic master disc is recorded individually using an argon laser and a vibration-isolated table. The facets are recorded on rectangular DCG holographic plates, which are then processed in the water/alcohol baths. When a suitable efficiency is achieved, resulting in an optimum intensity ratio of diffracted to transmitted light, the facets are "capped" (sealed with another piece of glass using an index-matching optical adhesive) for moisture protection then cut to a size appropriate to an automated copy-exposure machine. They are then masked to provide the designed size and shape when exposed by the copy beam.

The DCG copy disc fabrication process resembles, but is not identical to, a photographic contact-copy process. All of the DCG masters are placed on a computer-controlled wheel in the appropriate sequence. When the exposure cycle is started, a DCG disc is brought in under the master wheel with a slight air gap between the disc and the master. Because of this gap the process is not a true contact-copy process and some settling time is required after the motion stops before the exposure can begin. In situations where relative motion between the master and the disc is not required (such as with a submaster disc), index matching fluid can be used between the master and the disc. This limits relative motion and also greatly reduces reflections at the interface. For air gap copying, AR-coated caps are recommended for the masters. During the exposure sequence each master facet swings into position and is then sequentially illuminated in a step-and-repeat exposure process, using an expanded beam from an argon laser. That beam may be collimated, divergent, or convergent, depending on the desired characteristics of the copy HOE. The angle of illumination of the laser beam is modified from facet to facet considering each facet's construction and the difference between the copy-exposure wavelength and the scanner wavelength.

Exposure of each master holographic facet creates an optically identical holographic copy facet in the DCG through the interference of the diffracted beam with the undiffracted zero-order beam. As long as the copy process is reasonably close to a contact-copy process there will be no aberrations introduced by copying, regardless of copy wavelength or exposure angle. The reason for this is that the recorded interference pattern in the
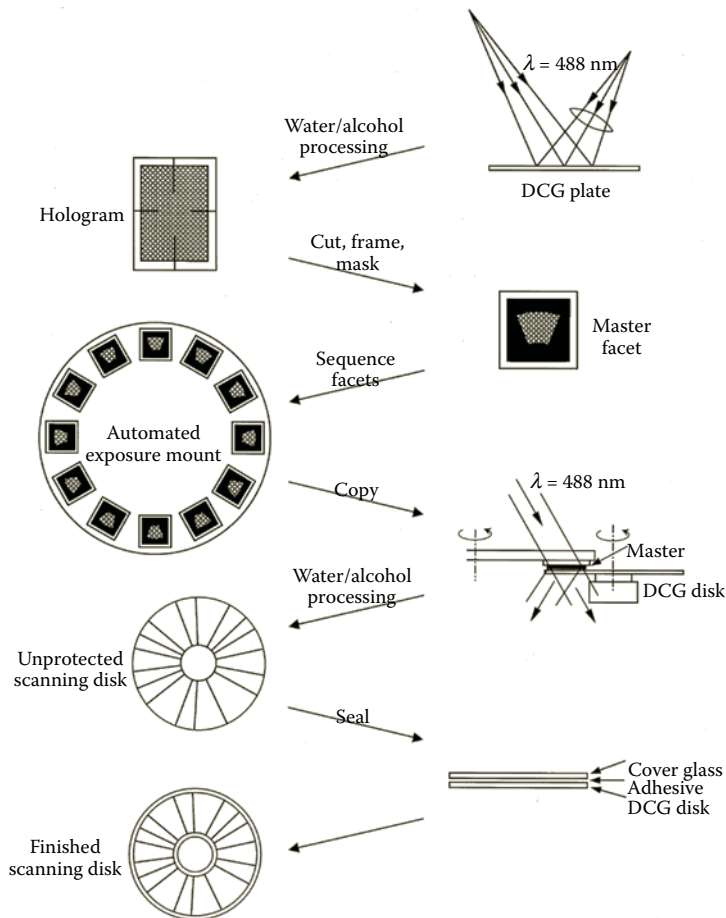
FIGURE 10.21
Holographie disc fabrication process.

hologram will accept any configuration of incidence beam and produce a corresponding conjugate diffraction beam. Because those two beams were created from that interference pattern, those two beams are the exact beams that will recreate that same interference pattern, provided the recording medium is located in the same place as the master hologram. The more removed the recording medium is from the master hologram, the more the properties of the two will differ. For a very small space, the difference is insignificant.

When all of the facets of the DCG copy disc are exposed, the disc is processed in a sequence of water and alcohol baths. The process is essentially the same as the process for the DCG masters, although some of the timing may be different, and precise control and consistency are, ironically, more important in the production process than they are in the master process. This is because DCG can be reprocessed if the initial results are not satisfactory; however, reprocessing leads to inconsistency and inefficiency, both of which are undesirable in a production environment. The details of the baths, the relative times, and the temperatures of the liquids, are proprietary.

A major objective in any DCG disc manufacturing operation is to establish a total exposure and development process that provides consistent results and high yield. One of the more difficult problems encountered in attempting to do this is the problem of "gel

swell"—the tendency for the exposed and processed gelatin to be thicker than the unexposed, unprocessed gelatin. This residual gelatin swell causes a shifting of the Bragg planes within the thickness of the gelatin so that the angle of the Bragg planes, relative to the surface of the gelatin, is not the same after processing as it was during exposure (see Figure 10.22). This results in a decrease in diffraction efficiency when the reconstruction beam is at the designed incidence angle. Any attempt to increase the diffraction efficiency by changing the reconstruction beam angle will introduce undesirable aberrations.

A similar effect also occurs if the postprocessing bulk refractive index is different from that anticipated. A different refractive index causes the incident reconstruction beam to refract into the gelatin at an angle different from that expected and, thereby, meet the Bragg planes at the wrong angle. Because the bulk index in DCG is dependent on the amount of microscopic air voids present in the processed gelatin, the bulk index is sensitive to changes in both the film preparation process and the water/alcohol process. In preparation, if the film is excessively hardened it will be more difficult to form voids and the bulk index will be higher. In general this also limits the range of the index modulation, Δn. During wet processing, if the hologram is left too long in hot water the gel can be oversoftened, creating excessive voids and lowering the bulk index. This also can cause the problem of excess scattering losses since the voids can become larger and, therefore, make the gel appear less homogenous to the scanner wavelength.

Gel swell and bulk index changes are separate effects that can be separately measured; however, symptomatically they produce the same effect—reduction in efficiency due to "Bragg error." Several methods of eliminating these undesirable effects have been described in the literature.[35] Generally, these involve either some sort of postprocessing chemical treatment or some form of postprocessing baking of the hologram. None of these methods is predictable enough to be suitable in a manufacturing process. If, however, the gel swell is predictable and consistent, it can be compensated for in the copy process by reducing the calculated angle of incidence of the argon laser copy beam. This increases the Bragg plane tilt angle. After processing, the gel swell will raise the Bragg planes, decreasing the tilt angle until it equals the original value. This process is described in greater detail by Dickson.[42]

If the processing methods are well controlled, so that the gel swell and bulk index are both predictable and consistent, then one can eliminate the Bragg error problem by altering the copy beam angle. Altering the copy beam angle, incidentally, has no effect on the optical properties of the holographic copy disc since these are fixed in the surface fringe
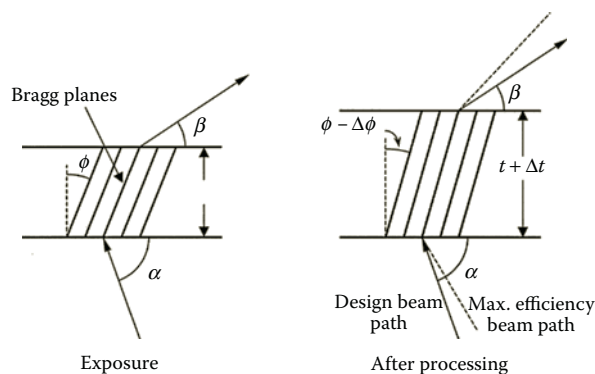
FIGURE 10.22
Effect of gel swell on the angle of the Bragg planes.

structure of the master. The copy process will always faithfully reproduce this fringe structure.

After the DCG disc is exposed and processed, several millimeters of the gelatin are stripped from the outer and inner edges to inhibit wicking-in of moisture in the sealed disc. The disc is then sealed with a glass cover disc for protection from moisture, using an index-matching optical adhesive. A metal hub is then bonded to the inner diameter of the disc and the disc is dynamically balanced.

The optical properties of each and every DCG copy disc will be identical to those of the master holographic disc. The optical characteristics of the holographic scanner are essentially established at the time of the construction of the holographic master. While it is possible to modify these characteristics somewhat through variations in predisc or postdisc optics, this is generally not done in holographic barcode scanners.

10.6.2 The Mechanically Replicated Surface-Relief Holographic Disc

The other primary holographic recording material, photoresist, has a diffracting structure in the form of surface deformation or relief as shown in Figure 10.20.[16,17] Consideration can therefore be given to using mechanical replication for mass production. This is not to imply that optical copying techniques cannot be used with surface relief, because they certainly can, either as master or copy or both.

Mechanical replication of surface relief is not new, having been developed decades ago for low-cost replication of mechanically ruled gratings. Today, it has become one of the primary manufacturing techniques of the production of high-quality gratings from holographic masters.[43]

Although replication of a surface relief hologram can be performed directly from the photoresist master, there is some danger that the photoresist may not stand up to repeated mechanical pressures, elevated temperatures, and/or the copy–master release process. Considering the difficulty and potential expense associated with master fabrication, a replication process that permits maximum replication volume is required. This is accomplished by the fabrication of a more durable submaster and usually results in the master itself being sacrificed. A very early technique, borrowed from the audiorecording industry uses a metal "stamper" to emboss or compression mold the relief into a vinyl thermoplastic.[44,45]

In the adaptation of this process,[46,47] a very fine grain layer of nickel or gold is deposited by evaporation or sputtering onto the relief to form a conductive conformal coating, typically a few hundred Angstroms thick. Nickel formation is then continued by other methods, such as electrochemical deposition, until a thickness of several hundred microns is achieved. At this point, the outer nickel surface has no significant relief and it can be attached to a rigid substrate. The sandwich is separated at the nickel/resist interface and residual photoresist dissolved away leaving behind a rigid metal replication of the relief. This structure is a negative of the original and can be used in hot pressing, injection molding, or epoxy replication processes, which will be discussed later.

An alternative method of submaster preparation is to transfer the resist relief downward into its own substrate by radio frequency (RF) sputter etching or reactive ion etching (RIE) techniques.[48,49] In these methods, the relief surface is removed at a uniform rate by bombarding the relief with accelerated ions or, in the case of RIE, with reactive atoms that react with the substrate molecules to form a volatile gas. The valleys of the resist pattern disappear first and the underlying substrate is exposed and etching occurs. By the time the resist peaks have disappeared, the valley areas are deeply cut into the substrate.

Proper choice of photoresist, substrate materials (such as silicon or quartz), and plasma parameters allows the surface relief to be accurately transferred into the substrate and the cross-sectional shape to be preserved.[50] These processes result in a submaster having a positive replication of the original master in contrast to the negative shape of the previous nickel submaster.

Once fabricated, these more durable submasters may be used to generate multiple copies. One such method, thermal mechanical embossing or compression molding, is accomplished by pressing the relief into a heated and softened thermoplastic film such as polymethyl methacrylate (PMMA) or polyvinyl chloride (PVC). Bartolini and colleagues[46] rolled the submaster together with a vinyl strip between two heated cylinders. Gale et al.[47] used a conventional hot stamping press at 150 °C and 3 atm. to emboss into PVC sheets. A similar pressing technique was used by Iwata and Tsujiuchi[51] with separation of the copy from the mold performed by sudden cooling and differential contraction.

Replication by pressing tends to introduce considerable strain and other inhomogeneities into the new substrate. These problems can be overcome by using injection molding techniques that have been developed for high-volume, high-quality fabrication of plastic lenses.[52] In this case, the submaster is one surface and an optically polished stainless steel flat is used as the facing, parallel surface. The appropriate polymer is plasticized to a more fluid state than used by compression molding and introduced into the temperature-controlled mold under high pressure.[53] Most of the materials in use are copolymers of PVC, polyvinyl acetate (PVA), and acrylic (PMMA) compounds. The acrylic material has an advantage over vinyl due to the lack of birefringence in the finished substrate, and the stability and ease with which it can be machined and polished.

The final alternative is to use a polymer that can be cross-linked by ultraviolet illumination.[54] This technique eliminates the need for high-temperature processing and reduces the possibility of induced stresses and dimensional changes upon cooling/curing. An injection mold apparatus can be adapted for these purposes as long as one plate is sufficiently transparent for the UV illumination. Depending on the use of release agents and relative adhesion, the replica can also be attached directly to a rigid substrate in the same operation.

10.7 AN EXAMPLE OF A HOLOGRAPHIC BARCODE SCANNER: THE METROLOGIC PENTA SCANNER

As an example of a holographic scanner that uses most of the design techniques and methods discussed thus far we will now discuss the Metrologie Penta Scanner, an industrial application scanner that exploits many of the advantages of holography. First we will discuss the design of the scan pattern and then the means by which it is produced.

10.7.1 The Penta Scan Pattern

The Penta scanner was designed as a large-scale "pass-through" scanner. In general, it was designed to create an aggressive scan volume at some distance from the scanner through which packages would pass in a roughly uniform manner. Ideally, a large range of package sizes with barcodes of moderate resolution would be able to pass through the scan volume, either manually or automatically, and as long as the package was roughly facing the scanner it would be successfully scanned throughout a large depth of field.

This broad application definition places several requirements on the design. A large depth of field will require multiple focal planes. With no specification given to package orientation the scanner will have to read essentially all orientations. Also, the variety of package sizes requires a large scan pattern size coming out of the scanner. All of these requirements can be met, quite easily, with holography.

The name of the "Penta" scanner comes from the pentagonal configuration of the scan pattern, shown in Figure 10.23. The basic pattern is formed by combining five simple rasters (groups of parallel lines) of different angular orientation. The rasters are evenly spaced through an entire 360°, making this pattern omnidirectional. Omnidirectionality is an essential characteristic for a barcode scanner if it is desired that the operator not waste time trying to identify the proper orientation of the code for a good scan. In automated applications, where the highest efficiency is desired, packages often are presented in truly random orientations as they pass under the scanner on a conveyor belt.

Assuming a direction of package flow, as shown in Figure 10.23, the primary parameters that define the pattern are the scan line length, the line separation, and the number of lines per orientation group. These parameters must be combined in such a way as to provide the desired total pattern omniwidth (width over which full omnidirectional scanning is possible). At the same time enough overlap in the near-vertical fields must be maintained such that codes in the "ladder" orientation (i.e., codes travelling in the direction perpendicular to its own bars and spaces) cannot slip through the pattern unscanned.

This pentagonal pattern not only provides for omnidirectional scanning, but it is also very aggressive on codes with higher aspect ratios, which are more difficult to scan than square codes like the original UPC code. The smaller angle between adjacent scan-line groups means each group has a smaller angular range that it must cover, and, therefore, less code height is required. This also means there will be less reliance on software stitching algorithms, but such algorithms can still be employed to make the scanner's performance even better.

Once the optimum scan pattern is established, the task still remains to provide for a large depth of field. Since the optimum performance pattern has been determined, the logical conclusion is to repeat that pattern several times at different distances from the scanner.
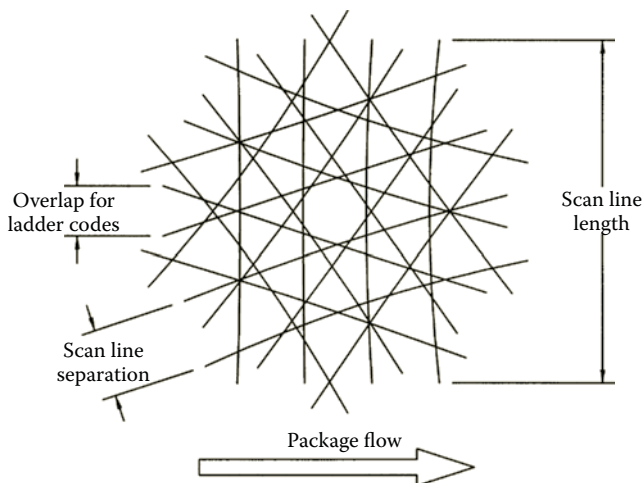
FIGURE 10.23
The two-dimensional representation of the Penta scan pattern.
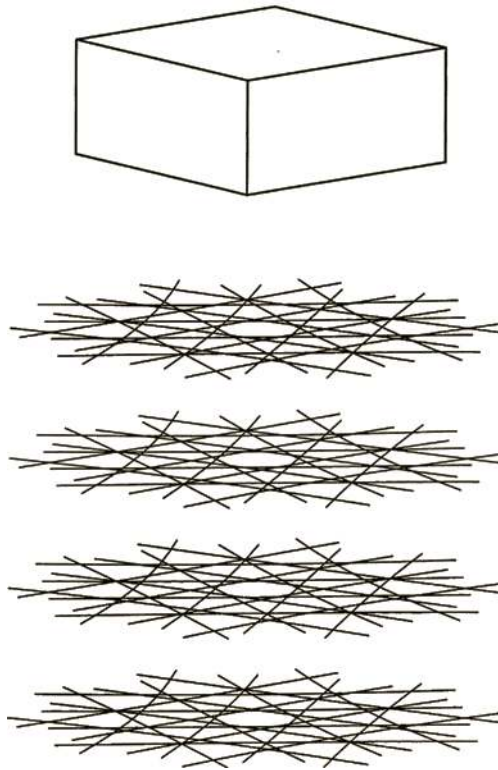
FIGURE 10.24
The three-dimensional scan volume of the Penta scanner.

Different focal planes are established providing acceptable overlap and producing a full, contiguous depth of field. For Penta four focal planes were chosen. A three-dimensional representation of the Penta scan pattern is shown in Figure 10.24.

10.7.2 The Penta Scanning Mechanism

The heart of the Penta scanning mechanism is, of course, a holographic disc. Also included are five scanning stations located around the periphery of the disc, each comprised of a VLD module prior to the disc and a folding mirror after the disc to direct the beams out the scanner window. A top view of the scanner with the cover removed, showing the five scanning stations, is shown in Figure 10.25.

A clearer view of the optics of a single scanning station can be seen from the side view in Figure 10.26. The dotted line in Figure 10.26 represents the outgoing beam path. The path starts at the VLD and is first roughly collimated by a conventional, aspheric lens. From there it reflects off a folding mirror, which directs the beam to the multifunction plate (MFP). The MFP is a multipurpose hologram, which, along with the VLD, lens, and mirror, finishes the subassembly of the "optics module."

The VLDs used in barcode scanners today have certain inherent properties, some of which are undesirable. With the use of the MFP, however, some of the undesirable effects can be alleviated. The functions performed by the MFP include beam aspect-ratio modification, astigmatism reduction, and dispersion minimization. In fact, VLDs inherently
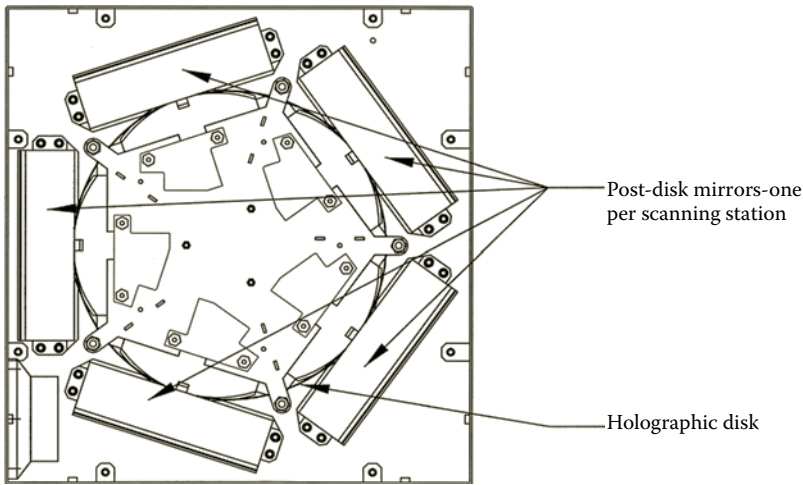
FIGURE 10.25
Penta scanner top view showing the different scanning stations.
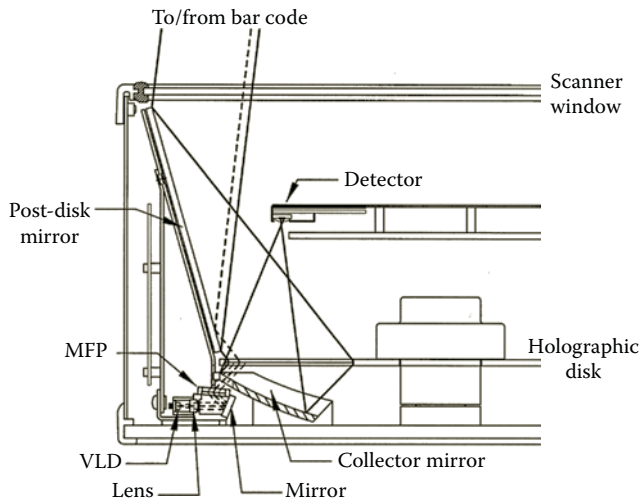
FIGURE 10.26
Penta scanner side view showing the scanning and light-collection optics.

produce beams with an elliptical shape and a characteristic astigmatism. Both of these properties of the beam can be manipulated by a single MFP (or by more than one if a greater range of control is desired) by simply choosing the incidence and diffraction angles of the MFP properly. At the same time, the dispersion produced by the facets of the disc (natural to all diffraction gratings) can be minimized by the same MFP by countering it with its own dispersion.

Following the MFP the beam heads directly to the holographic disc, incident on it at a specified angle. This is where the primary pattern formation occurs. Each of the five rasters in the pattern in Figure 10.23 contains four lines, and that pattern is repeated in four focal planes. This requires 16 unique facets on the disc, which were shown previously in

Scan lines are coming out of the page

Disk axis of rotation

Focal planes
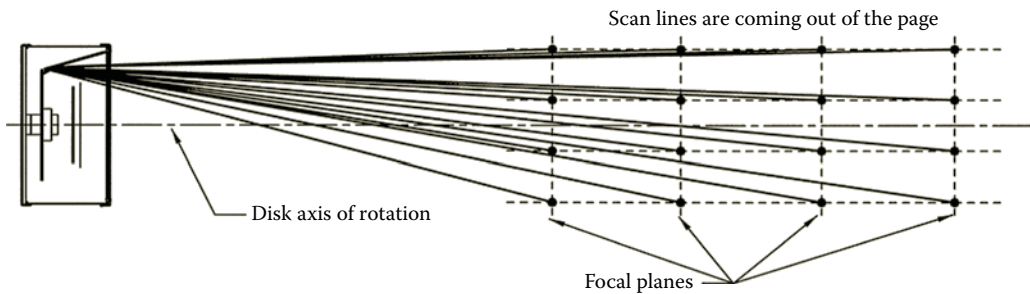
FIGURE 10.27
Side view of the scan lines produced by one scanning station.

Figure 10.18. Each one of these facets has a different diffraction angle, θ_d, and a different focal length, f, which results in the beam being focused at a different distance, s. The result of combining all the different focal lengths and diffraction angles is shown in Figure 10.27. Choosing the right combination of focal lengths and diffraction angles results in scan lines that are laterally equidistant from the disc rotation axis, thereby reproducing the desired pattern at the different distances.

The solid lines in Figure 10.26 represent the path of the light reflected off a barcode as it makes its way to the signal detector. The light that is collected returns essentially along the same path by which it left the scanner, making the scanner a retrocollective system. On the way back the light has diffusely spread out to completely fill the aperture of the facet, after first being reflected off the large pattern-folding mirror. The light is then diffracted back towards the module; however, a collector mirror is in the path everywhere except for a small hole through which the outgoing beam passed.

This collector mirror generally has a parabolic or elliptical shape. The light incident on the collector is focused and directed upward to the signal detector; however, in order to get there it must, once more, pass through the disc. On this third pass, however, the desire is for the disc to not affect the beam in any way. In reality it is impossible not to have some losses occur, but if the disc is correctly designed and manufactured those losses need not be much worse than that of a plain piece of glass. This is due to the angular sensitivity of the diffraction efficiency of the holograms. A properly manufactured disc will only have high efficiency at the designed incidence and diffraction angles. Because the rays proceeding from the collector to the detector are incident on the disc at angles far enough removed from the design incidence and diffraction angles, the resulting transmission of the disc is relatively high.

REFERENCES

1. Savir, D.; Laurer, D.J. *IBM Systems J.* 1975, *14*, 16.
2. Broockman, E. U.S. Patent 4,717,818, assigned to IBM, January 5, 1988.
3. Allais, D.C. *Bar Code Symbology;* Intermec Corporation: Everett WA, 1984.
4. Cindrich, I. *Appl. Opt.* 1967, *6*, 531.
5. Beiser, L. *Proc. 1975 Electro-Opt. Syst. Des. Conf.* 1975, 333.
6. Beiser, L.; Darcey, E.; Kleinschmitt, D. *Proc. 1973 Electro-Opt. Syst. Des. Conf.* 1973, 75.

7. Beiser, L. *Holographic Scanning;* Wiley: New York, 1988.
8. Sincerbox, G.T. *Laser Beam Scanning;* Marshall, G., Ed.; Marcel Dekker: New York, 1985; 1.
9. Gabor, D. *Nature* 1948, *161,* 777.
10. Leith, E.; Upatnieks, J. *J. Opt. Soc. Am.* 1962, *52,* 1123.
11. Dickson, L.D.; Sincerbox, G.T.; Wolfheimer, A.D. *IBM J. Res. Dev.* 1982, *26,* 228.
12. Pole, R.V.; Werlich, H.W.; Krusche, R. *Appl. Opt.* 1978, *17,* 3294.
13. Dickson, L.D. *Appl. Opt.* 1970, *9,* 1854.
14. Bartolini, R.A. *Proc. SPIE* 1977, *123,* 2.
15. Smith, H.M., Ed. *Holographie Recording Materials;* Springer-Verlag: New York, 1977.
16. Werlich, H.; Sincerbox, G.; Yung, B. Dig. *1983 Conf. Lasers Electro-Opt.* 1983, 224.
17. Werlich, H.; Sincerbox, G.; Yung, B. *J. Imaging Tech.* 1984, *10*(3); 105.
18. Rogers, G. *J. Opt. Soc. Amer.* 1965, *55,* 1185.
19. Upatnieks, J.; Leonard, C. *Appl. Opt.* 1969, *8,* 85.
20. Pennington, K.; Harper, J. *Appl. Opt.* 1970, *9,* 1643.
21. Graube, A. *Appl. Opt.* 1974, *13,* 2942.
22. Phillips, N.; Porter, D. *J. Phys. E.* 1976, *9,* 631.
23. Phillips, N.; Cullen, R.; Ward, *A.*; Porter, D. *Photogr. Sei. Eng.* 1980, *24,* 120.
24. Booth, B. *J. Appl. Phot. Eng.* 1977, *3,* 24.
25. Chandross, E.; Tomlinson, W.; Aumiller, G. *Appl. Opt.* 1978, *17,* 566.
26. Kogelnik, H. *Bell. Sys. Tech. J.* 1969, *48,* 2909.
27. Gambogi, W.J.; Gerstadt, W.A.; Mackara, S.R.; Weber, A.M. *Proc. SPIE* 1991, *1555,* 256.
28. Ingwall, R. *Proc. SPIE* 1986, *615,* 81.
29. Shankoff, T. *Appl. Opt.* 1968, *7,* 2101.
30. Lin, L. *Appl. Opt.* 1969, *8,* 903.
31. Chang, B.J. *Opt. Eng.* 1980, *19,* 642.
32. Meyerhofer, D. *RCA Rev.* 1972, *33,* 111.
33. Samoilovich, D.; Zeichner, *A.*; Freisem, A. *Photogr. Sei. Eng.* 1980, *24,* 161.
34. Sjolinder, S. *Photogr. Sei. Eng.* 1981, *25,* 112.
35. Chang, B.J.; Leonard, C.D. *Appl. Opt.* 1979, *18,* 2407.
36. Graube, A. *Opt. Commun.* 1973, *8*, 251.
37. Graube, A. *Photogr. Sei. Eng.* 1978, *22,* 37.
38. Kubota, T.; Ose, T. *Appl. Opt.* 1979, *18,* 2538.
39. Akagi, M. *Photogr. Sei. Eng.* 1974, *18,* 248.
40. Kubota, T.; Ose, T.; Sasaki, M.; Honda, M. *Appl. Opt.* 1976, *15,* 556.
41. Dickson, L.D. U.S. Patent 4,416,505, assigned to IBM, November 22, 1983.
42. Lerner, J.; Flamand, J.; Thevenon, A. *Proc. SPIE* 1982, *353,* 68.
43. Ruda, J.C. *J. Audio Eng. Soc.* 1977, *25,* 702.
44. Roys, W.E., Ed. *Disc Recording and Reproduction;* Dowden, Hutchinson & Ross: Stroudsburg, PA, 1978.
45. Bartolini, R.; Feldstein, N.; Ryan, R.J. *J. Electrochem. Soc.* 1973, *120,* 1408.
46. Gale, M.T.; Kane, J.; Knop, K. *J. Appl. Phot. Eng.* 1978, *4,* 41.
47. Hanak, J.J.; Russell, J.P. *RCA Rev.* 1971, *32,* 319.
48. Lehman, H.W.; Widner, R. *J. Vac. Sci. Tech.* 1980, *17,* 1177.
49. Matsui, S.; Moriwaki, K.; Aritome, H.; Namba, S.; Shin, S.; Suga, S. *Appl. Opt.* 1982, *21,* 2787.
50. Iwata, F.; Tsujiuchi, J. *Appl. Opt.* 1974, *13,* 1327.
51. Wolpert, H.D. *Photonics spectra* 1983, *17*(2–3), 68.
52. Ryan, R.J. *RCA Rev.* 1978, *39,* 87.
53. Okino, Y.; Sano, K.; Kashihara, T. *Proc. SPIE* 1982, *329,* 236.

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

11

Acousto-Optic Scanners and Modulators

Reeder N. Ward

Noah Industries, Inc.
Melbourne, Florida, USA

Mark T. Montgomery

SkyCross, Inc.
Viera, Florida, USA

Milton Gottlieb

Consultant
Carnegie Mellon University
Pittsburgh, Pennsylvania, USA

CONTENTS

11.1 INTRODUCTION

It will be apparent to the reader of this book that there are a great variety of applications of lasers for which scanning devices are required, and that these applications include a wide range of performance requirements on the scanner. The basic specifications include speed, resolution, and random access time, and the choice of a scanner will be determined by these parameters. Acousto-optic (AO) scanners are best suited to those systems that are of moderate cost, since the cost of AO Bragg cells and the associated drive electronics are by no means trivial, and for which the resolution requirement is about 1000 spots. In addition, AO technology is most appropriate where random access times on the order of 10 μs are needed, or where it may be desired to perform intensity modulation on the laser beam, as in image recording. There are currently many systems employing AO scanners, perhaps the most familiar being laser printers, in which the scanner capability is an excellent match to the system requirements. Large-area television display was one of the first applications considered for AO scanners, and it performs this function very well, although such display systems are relatively uncommon. These, as well as other applications of AO scanners, will be described in detail in a later section.

The interaction of light waves with sound waves has in recent years been the basis of a large number of devices related to various laser systems for display, information handling, optical signal processing, and numerous other applications requiring the spatial or temporal modulation of coherent light. The phenomena underlying these interactions were largely understood as long ago as the mid-1930s, but remained as scientific curiosities, having no practical significance, until the 1960s. During this period several technologies were developing rapidly, at the same time that many applications of the laser were being suggested that require high-speed, high-resolution scanning methods. These new technologies gave rise to high-efficiency, wideband acoustic transducers capable of operation to several gigahertz, high-power wideband solid-state amplifiers to drive such transducers, and the development of a number of new, synthetic AO crystals with very large figure of merit (low-drive-power requirements) and low acoustic losses at high frequencies. This combination of properties makes acousto-optics the method of choice for many systems, and is very often the only approach to satisfy demanding requirements. In this chapter,

the underlying principles of AO interactions will be reviewed, and this will be followed by a description of the materials considerations and the relevant acoustic technology. AO scanning devices will be described in some detail, including the important features of optical design for various types of systems.

11.2 ACOUSTO-OPTIC INTERACTIONS

11.2.1 The Photoelastic Effect

The underlying mechanism of all AO interactions is very simply the change induced in the refractive index of an optical medium due to the presence of an acoustic wave. An acoustic wave is a traveling pressure disturbance that produces regions of compression and rarefaction in the material, and the refractive index is related to the density, for the case of an ideal gas, by the Lorentz–Lorenz relation

$$\frac{n^2 - 1}{n^2 + 1} = \text{constant} \times r \tag{11.1}$$

where η is the refractive index and ρ is the density. In fact, this relation is adhered to remarkably well for most simple solid materials as well. The elasto-optic coefficient is obtained directly by differentiation of Equation 11.1:

$$r\frac{\partial n}{\partial r} = \frac{(n^2 - 1)(n^2 + 1)}{6n} \tag{11.2}$$

where it is understood that the derivative is taken under isentropic conditions. This is generally the case for ultrasonic waves, in which the flow of energy by thermal conduction is slow compared with the rate at which density changes within a volume smaller than an acoustic wavelength. The fundamental quantity given by Equation 11.2, also known as the photoelastic constant p, can be easily related to the pressure applied, with the result

$$p = \frac{1}{b}\frac{\partial n}{\partial P} \tag{11.3}$$

where P is the applied pressure and β is the compressibility of the material. The photoelastic constant of an ideal material with a refractive index of 1.5 is 0.59. It will be seen later that the photoelastic constants of a wide variety of materials lie in the range from about 0.1 to 0.6, so that this simple theory gives a reasonably good approximation to measured values.

The relation in Equation 11.3 follows from the usual definition of the photoelastic constant:

$$\Delta\left(\frac{1}{e}\right) = \Delta\left(\frac{1}{n^2}\right) = pe \tag{11.4}$$

where ε is the dielectric constant ($\varepsilon = n^2$) and e is the strain amplitude produced by the acoustic wave. From Equation 11.4 it is easily seen that the change in refractive index, Δn, produced by the strain is

$$\Delta n = -\frac{1}{2}n^3 pe \qquad (11.5)$$

where e is of the form $e_0\, exp(i\Omega t)$ for an acoustic wave of frequency Ω. The magnitude of refractive index change typical for AO devices is not large. Strain amplitudes lie in the range 10^{-8} to 10^{-5}, so that using the above expressions for Δn *and* p gives for Δn about 10^{-8} to 10^{-5} (for $n = 1.5$). It may be somewhat surprising, then, that devices based upon such a small change in refractive index are capable of generating large effects, but it will be seen that this comes about because these devices are configured in a way that can produce large phase changes at optical wavelengths.

The relation defining the photoelastic interaction has been written in Equation 11.5 as a scalar relation, in which the photoelastic constant is independent of the directional properties of the material. In fact, even for an isotropic material such as glass, longitudinal acoustic waves and transverse (shear) acoustic waves cause the photoelastic interaction to assume different parameters. A complete description of the interaction, particularly for anisotropic materials, requires a tensor relation between the dielectric properties, the elastic strain, and the photoelastic coefficient. This may be represented by the tensor equation

$$\Delta\left(\frac{1}{n^2}\right)_{ij} = \sum_{kl} p_{ijkl} e_{kl} \qquad (11.6)$$

where $(1/n^2)_{ij}$ is a component of the optical index ellipsoid, e_{kl} are the cartesian strain components, *and* p_{ijkl} are the components of the photoelastic tensor. The crystal symmetry of any particular material determines which of the components of the photoelastic tensor may be nonzero, and also which components are related to others. This may be useful in determining whether some crystal, based only upon its symmetry, may even be considered for certain applications.

11.2.2 Isotropic AO Interaction

The most useful photoelastic effect is the ability of acoustic waves to diffract a light beam. There are several ways to understand how diffraction comes about; the acoustic wave may be thought of as a diffraction grating, made up of periodic changes in optical phase, rather than transparency, and moving at sonic velocity rather than being stationary. Thus, it is possible to analyze the diffraction as resulting from a moving phase grating. Alternatively, the light and sound may be thought of as particles, photons, and phonons, undergoing collisions in which energy and momentum are conserved. Either of these descriptions may be used to obtain all the important diffraction effects, but some are more easily understood on the basis of one or the other. It will be useful, then, to outline both of these approaches.

To examine the simplest case of plane acoustic waves interacting with plane light waves, consider Figure 11.1. Suppose the light wave, of frequency ω and wavelength λ, is incident from the left into a delay line with an acoustic wave of frequency Ω and wavelength Λ. If
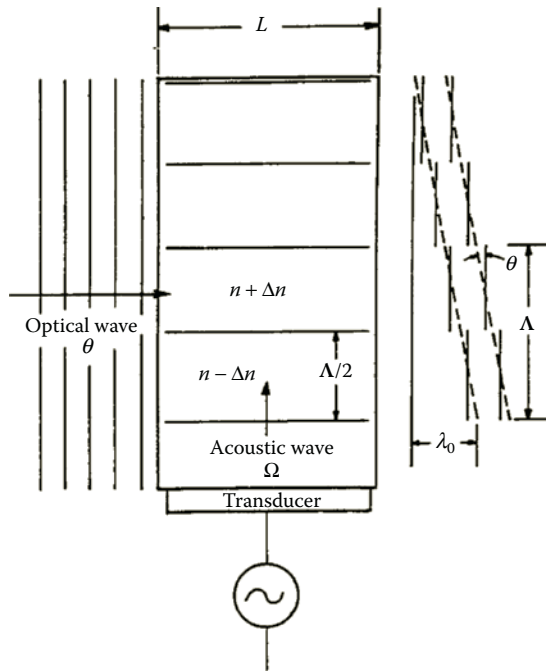
FIGURE 11.1
Tilting of optical wavefronts caused by upward-traveling acoustic wave.

the refractive index of the delay medium is $n + \Delta n$ in the presence of the acoustic wave, the phase of the optical wave will be changed by an amount

$$\Delta f = 2p \frac{L}{1} \Delta n \tag{11.7}$$

if the length of the delay line is L. Some typical values of $\Delta\phi$ can be obtained by assuming $L = 2.5$ cm (1 in) and $\lambda = 0.5$ μm, with Δn reaching a peak value of 10^{-5}. This yields a phase change of π rad, which is, of course, quite large. It is large because L/λ, the number of optical wavelengths, is 50,000, so that a very small Δn can still produce a sizable $\Delta\phi$. If the electric field incident on the delay line is represented by

$$E = E_0 e^{iwt} \tag{11.8}$$

then the field of the phase-modulated emerging light will be

$$E = E_0 e^{i(wt + \Delta f)} = e^{iwt} e^{i2p(L/1)(a_0 \sin \Omega t)} \tag{11.9}$$

We shall not give a detailed derivation of the resulting temporal and spatial distribution of the light field, but we can use intuition and analogy with radiowave modulation to arrive at the resultant fields. It is well known from radio frequency (RF) engineering that the spectrum of a phase-modulated carrier of frequency ω consists of components separated by multiples of the modulation frequency Ω, as shown in Figure 11.2. There is
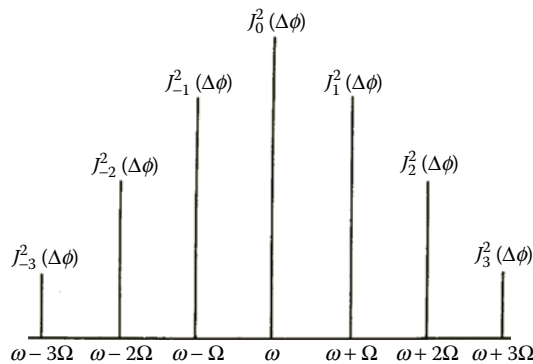
FIGURE 11.2
Intensity of diffracted orders due to Raman–Nath interaction.

a multiplicity of sidebands about the carrier frequency, such that the frequency of the nth sideband is $\omega + n\Omega$, where n is both positive and negative. The amplitude of each sideband is proportional to the Bessel function of order equal to the sideband number, and whose argument is the modulation index $\Delta\phi$. Although not shown by Figure 11.2, note that the odd-numbered negative orders are 180° out of phase with the others. The light emerging from the delay line is composed of a number of light waves whose frequencies have been shifted by $n\Omega$ from the frequency ω of the incident light. The relative amplitudes will be determined by the peak change in the refractive index.

In order to understand the diffraction of the light by the acoustic wave, consider the optical wavefronts in Figure 11.1. Because the velocity of light is about five orders of magnitude greater than the velocity of sound, it is a good approximation to assume that the acoustic wave is stationary in the time that it takes the optical wave to traverse the delay line. Suppose that during this instant the half-wavelength region labeled $\eta + \Delta n$ is under compression and $n - \Delta n$ is under rarefaction. Then the part of the optical plane wave passing through the compression will be slowed (relative to the undisturbed material of index n) while the part passing through the rarefaction will be speeded up. In this rough picture, the emerging wavefront will be "corrugated," so that if the corrugations are joined by a continuous plane its direction is tilted relative to that of the incident light wavefronts. Because the optical phase changes by 2π for each acoustic wavelength Λ along the acoustic beam direction, the tilt angle will be given by $\theta \cong \lambda/\Lambda$.

The direction normal to the tilted plane is the direction of optical power flow and represents the diffracted light beam. Note that the corrugated wavefront could just as well have been connected by a tilted wavefront at an angle given by $\theta \cong -\lambda/\Lambda$. This corresponds to the first negative order, the other to the first positive order. At this point we will note that an important consideration in the operation of AO Bragg cells, or for that matter of most ultrasonic devices, is the ratio of the acoustic wavelength Λ to the transducer length L. The assumption that the acoustic energy propagates as a plane wave is valid when this ratio is very small or when there is little diffraction of the wave. However, when this ratio is not large, the acoustic propagation is more properly described in terms of the sum of plane waves, the angular spectrum of such plane waves increasing as the ratio increases. If we consider that partial wave which is propagating at an angle λ/Λ to the forward direction, then we see that the light that has been diffracted into the first order may be diffracted a second time by this partial wave into an angle $2\theta = 2\lambda/\Lambda$, and that the frequency of this light will once again be upshifted, for a total frequency shift of 2ω. If the angular spectrum

of acoustic waves contains sufficient power of still higher orders, then this process can be repeated again, so that light will be multiply diffracted into higher order angles, $\eta\theta = n\lambda/\Lambda$ each with a frequency shift $n\omega$. A similar argument holds for the negative orders, so that a complete set of diffracted light beams will appear as shown in Figure 11.3, where the angular deflection corresponding to the nth order is given by $\theta n \cong \pm\, n\lambda/\Lambda$ and the frequency of the light deflected into the nth order is $\omega \pm n\Omega$. The intensity of the carrier wave, or zeroth order, will be zero when the modulation index $\Delta\phi$ is equal to 2.4. The generally important first order will have a maximum value of 34% of the input for $\Delta\phi = 1.8$, decreasing for higher modulation. These phenomena were described by Debye and Sears[1] and are often referred to as Debye–Sears diffraction. Similar observations were published almost simultaneously by Lucas and Biquard.[2] An extensive theoretical analysis of the effect was given by Raman and Nath,[3] and so it is alternatively referred to as Raman–Nath diffraction. A distinctive feature of this type of diffraction is that it is limited to low acoustic frequencies (or relatively long wavelengths). The origin of this limitation lies in the diffraction spreading of the light beam as it traverses apertures formed by the columns of compression and rarefaction in the acoustic beam. If the length of the acoustic beam along the light propagation direction is large enough, the diffraction spread of the light between adjacent compression and rarefaction regions will overlap, so that the Debye–Sears model is no longer valid. To estimate the characteristic length L_0 that bounds the Debye–Sears model, suppose the compression and rarefaction apertures are one-half an acoustic wavelength, $\Lambda/2$, so that the angular diffraction spread of the light is $\delta\phi \approx 2\lambda/\Lambda$. Then L_0 can be defined as that interaction length for which the aperture diffraction spreads the light by one-half an acoustic wavelength,

$$L_0\, df = \frac{\Lambda}{2} \tag{11.10}$$

or

$$L_0 = \frac{\Lambda^2}{4\lambda} \tag{11.11}$$
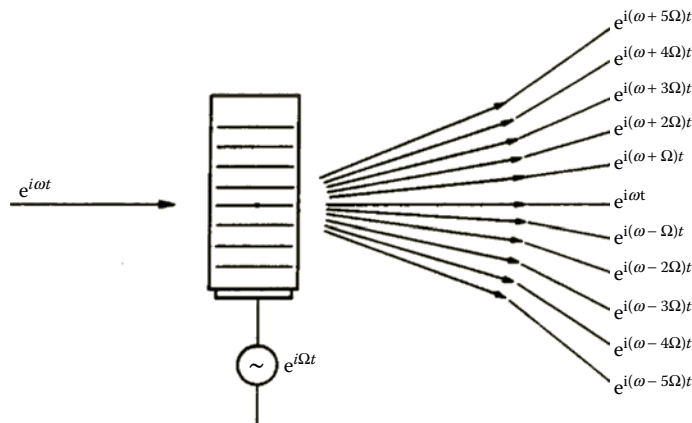
FIGURE 11.3
Raman–Nath diffraction of light into multiple orders.

The interaction length is sometimes expressed in terms of a ratio known as the Raman–Nath parameter (often referred to as the Klein–Cook parameter) Q, as

$$Q = \frac{2pL\mathit{1}}{\Lambda^2}$$

Devices with interaction length $Q < \pi$ are said to operate in the Raman–Nath regime, while devices with $Q > 4\pi$ are said to operate in the Bragg regime. For typical values of $L = 1$ cm and $\lambda = 6.33 \times 10^{-5}$ cm, $Q = 1$ for $\Lambda = 0.0159$ cm (0.006 in), which corresponds to a frequency of 31.4 MHz for a material whose acoustic velocity is 5×10^5 cm/s (2×10^5 in/s).

In the Bragg regime, the thin grating approximation no longer holds. If the incident light beam is normal to the sound beam propagation direction, the higher diffraction orders interfere destructively beyond L_0, eventually completely wiping out the diffraction pattern. In order for constructive interference to take place, the angle of incidence must be tilted with respect to the acoustic beam direction. To better understand what conditions must be satisfied for this, it is easier to think of the light and sound waves as colliding particles, photons, and phonons. In this description, the light and sound take on the attributes of particles, and the dynamics of their collisions are governed by the laws of conservation of energy and conservation of momentum. The magnitudes of the momenta of the light and sound waves are given by the well-known expressions

$$|k| = \frac{wn}{c} = \frac{2pn}{\mathit{1}_0} \tag{11.12}$$

and

$$|K| = \frac{\Omega}{u} = \frac{2p}{\Lambda} \tag{11.13}$$

respectively. In the latter equation, ν is the velocity of sound in the delay medium, $\upsilon = 2\pi\Omega\Lambda$. Conservation of momentum is expressed by the vector relation

$$k_i + K = k_d \tag{11.14}$$

the diagram for which is shown in Figure 11.4a, where k_i and k_d represent the momentum of the incident photon and the diffracted photon, respectively. The process may be thought of as one in which the acoustic phonon is absorbed by the incident photon to form the diffracted photon. Thus, conservation of energy requires that

$$hw_0 = hw_1 + h\Omega \tag{11.15}$$

or

$$w_d = w_1 + \Omega$$

in which h is Planck's constant.
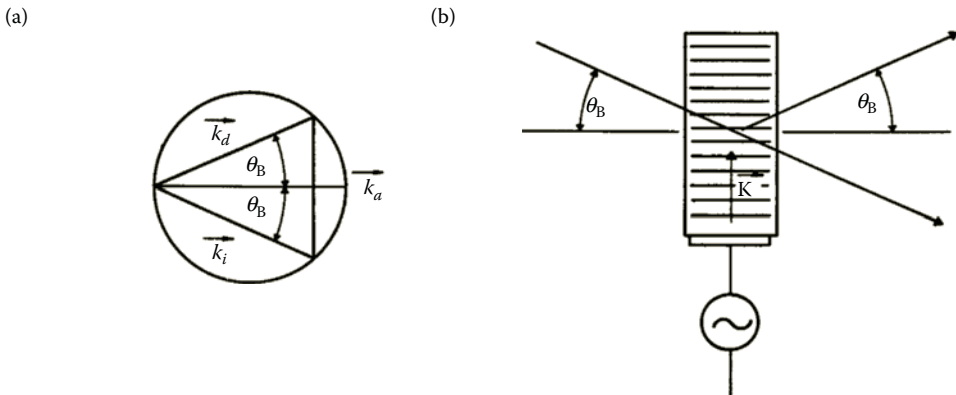
FIGURE 11.4
Bragg diffraction in isotropic medium: (a) phasor diagram; (b) Bragg cell deflection.

Because ω_i lies in the optical frequency range, and Ω will typically lie in the RF or microwave range, $\Omega \ll \omega_i$ so that $\omega_d \approx \omega_i$. This results in the magnitudes of k_i and k_d being almost equal, so that the momentum triangle of Figure 11.4a is isosceles, and the angle of incidence (with respect to the normal to K) is equal to the angle of diffraction. This angle of incidence is easily obtained from Figure 11.4a as

$$\sin q_B = \frac{1}{2}\frac{K}{k} = \frac{1}{2}\frac{\lambda}{\Lambda}$$

(11.16)

It is called the Bragg angle because of its similarity to the angle of diffraction of X-rays from the regularly spaced planes of atoms in crystals. The configuration of these vectors in relation to the delay line is shown in Figure 11.4b. In order for diffraction to take place, the light must be incident at the angle θ_B, and the diffracted beam will appear only at this same angle. In contrast to the Debye–Sears regime, there are no higher-order diffracted beams. In the full mathematical treatment of the Bragg limit $(Q \gg 1)$, light energy may appear at the higher orders, but the probability of its doing so is extremely small so that the intensity at higher orders is essentially zero. The diagrams in Figure 11.4 show the interaction in which the diffracted photon is higher in energy than that of the incident photon, but the reverse can also take place. If the sense of the vector K is reversed with respect to k_i, then $\omega_d = \omega_i - \Omega$, and the diffracted negative first-order results.

It is important to understand that the Debye–Sears effect and Bragg diffraction are not different phenomena, but are the limits of the same mechanism. The Raman–Nath parameter Q determines which is the appropriate limit for a given set of values λ, Λ, and L. Quite commonly in practice, these values will be chosen such that neither limit applies, and $Q \approx 1$. In this case, the mathematical treatment is quite complex, and experimentally it is found that one of the two first-order diffracted beams may be favored, but that higher orders will be present.

Having obtained the angular behavior of light diffracted by acoustic waves, the next most important characteristic is the intensity of the diffracted beam. Again, the full mathematical treatment is beyond the scope of this book, but a very good intuitive calculation leads to results that are useful. Referring to the spectrum of a phase-modulated wave

shown in Figure 11.2, we can see that the ratio of the intensity in the first order to that in the zero order is

$$\frac{I_1}{I_0} = \left[\frac{J_1(\Delta f)}{J_0(\Delta f)}\right]^2 \tag{11.17}$$

We shall now show in detail how this result comes about for acousto-optically diffracted light. The acoustic power flow is given by

$$p = \frac{1}{2}cue^2 \tag{11.18}$$

where c is the elastic stiffness constant. The elastic stiffness constant is related to the bulk modulus β and the density and acoustic velocity through the well-known expression

$$c = \frac{1}{b}ru^2 \tag{11.19}$$

Thus, the acoustic power density is

$$P_A = \frac{1}{2}ru^3e^2 \tag{11.20}$$

We can express the phase modulation depth in terms of the acoustic power density, using Equation 11.5 for Δn and Equation 11.7 for $\Delta\phi$, with the result

$$\Delta f = 2p\frac{L}{l}\Delta n = -p\frac{L}{l}n^3p\left(\frac{2P_A}{ru^3}\right)^{1/2} \tag{11.21}$$

For small modulation index, the zero-order and first-order Bessel functions can be approximated by

$$J_0(\Delta f) \approx \cos(\Delta f) \approx 1 - \Delta f \tag{11.22}$$

and

$$J_1(\Delta f) \approx \sin(\Delta f) \approx \Delta f$$

so that the small-signal approximation to the diffracted light is, from Equation 11.17,

$$\frac{I_1}{I_0} \approx (\Delta f)^2 = \frac{p^2}{2}\left(\frac{L}{l}\right)^2\left(\frac{n^6p^2}{ru^3}\right)P_A \tag{11.23}$$

This efficiency may be expressed in terms of the total acoustic power P,

$$P = P_A(LH) \tag{11.24}$$

where H is the height of the transducer, and

$$\frac{I_1}{I_0} = \frac{p^2}{2} \frac{L}{H} \left(\frac{n^6 p^2}{ru^3} \right) \frac{P}{1^2} \tag{11.25}$$

The quantity in parentheses depends only upon the intrinsic properties of the AO material, while the other parameters depend upon external factors. It is therefore defined as the figure of merit of the material,

$$M_2 = \left(\frac{n^6 p^2}{ru^3} \right) \tag{11.26}$$

from which it can be seen that, in general, the most important factors leading to high AO efficiency will be a high refractive index and a low acoustic velocity. This does not guarantee a large figure of merit, since the photoelastic constant may be very small, or even zero.

The other factors in Equation 11.25 have the following effect on the diffraction efficiency. The efficiency decreases quadratically with increasing wavelength, so that the power requirements for operation in the infrared (IR) may be hundreds of times that required for the visible. For high efficiency, it will be desirable to have a large aspect ratio, L/H, leading to a configuration as shown in Figure 11.5. It is difficult to make conventional bulk devices with H much less than 1 mm, so that aspect ratios up to about 50 can be achieved. Much higher aspect ratios can be reached in guided optical wave devices. A more exact calculation of the diffraction efficiency in the Bragg regime[4] yields the result

$$\frac{I_1}{I_0} = \sin^2 \left[\frac{p^2}{2} \frac{L}{H} M_2 \frac{P}{1^2} \right]^{1/2} \tag{11.27}$$

For low signal levels Equation 11.27 reduces to the same expression as in Equation 11.25. To obtain an order of magnitude for the power requirements of an acousto-optic deflector
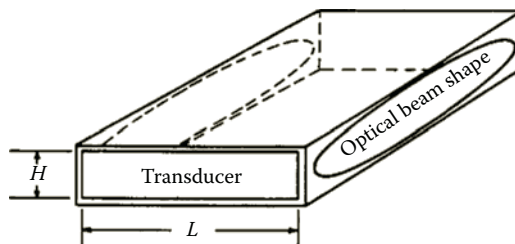
FIGURE 11.5
Transducer and optical beam shapes for optimization of acousto-optic diffraction.

(AOD), let us assume a material with $n = 1.5$, $\rho = 3$, $\nu = 5 \times 10^5$ cm/s, and a photoelastic constant calculated from the $\simeq$ Lorentz–Lorenz expression, $p \cong 0.6$, so that $M \simeq 1.1 \times 10^{-17} \, s^3/g$. If the remaining parameters are $L = 1$ cm and $\lambda = 0.6$ µm, then by assuming a maximum acoustic power density for CW operation of 1 W/cm² (10^7 ergs/cm² s), the maximum obtainable efficiency is 15%. We shall see later, however, that materials and designs are available that are capable of realizing higher efficiencies with lower power levels.

11.2.3 Anisotropic Diffraction

Optical materials such as glass, or crystals with cubic structure, are isotropic with respect to their optical properties; that is, they do not vary with direction. Many crystals, on the other hand, are of such structure, or symmetry, that their optical properties depend on the direction of polarization of the light in relation to the crystal axes. They are birefringent; that is, the refractive index is different for different direction of light polarization.

The theory of diffraction of light thus far presented has assumed that the optical medium is isotropic, or at least that it is not birefringent. A number of important AO devices make use of the properties of birefringent materials, so a brief description of the important characteristics of anisotropic diffraction will be given here. The essential difference from diffraction in isotropic media is that the momentum of the light,

$$k = \frac{2p}{l} = \frac{2pn}{l_0} \tag{11.28}$$

will in general, be different for different light polarization directions. Thus, the vector diagram representing conservation of momentum will no longer be the simple isosceles triangle of Figure 11.4a. The momentum vectors for light that is ordinary polarized will terminate on a circle, as shown in Figure 11.6, while those for light that is extraordinary polarized will terminate on the ellipse of Figure 11.6.

To understand the effect of anisotropy on diffraction, it is necessary to mention another phenomenon that occurs when light interacts with shear acoustic waves, that is, waves in which the displacement of matter is perpendicular to the direction of propagation of the acoustic wave. A shear acoustic wave may cause the direction of polarization of the diffracted light to be rotated by 90°. The underlying reason for this is that the shear disturbance induces a birefringence which acts upon the incident light as does a birefringent plate; in other words, it causes the plane of polarization to be rotated. This phenomenon occurs in isotropic materials as well as in anisotropic materials; however, in isotropic materials the momentum vector, $k = 2\pi n/\lambda_0$, will be the same for both polarizations, so there is no effect on the diffraction relations. Suppose, instead, that the interaction occurs in a birefringent crystal in a plane containing the optic axis. Let us choose the example as shown in the index surfaces in Figure 11.6, in which the incident light is an extraordinary ray and the diffracted light is an ordinary ray. For this example,

$$k_i = \frac{2pn_e}{l_0} \quad \text{and} \quad k_d = \frac{2pn_0}{l_0} \tag{11.29}$$

and the angles of incidence, θ_i, and diffraction, θ_d, are in general not equal. The theory of anisotropic diffraction was developed by Dixon,[5] in whose work the expressions for the
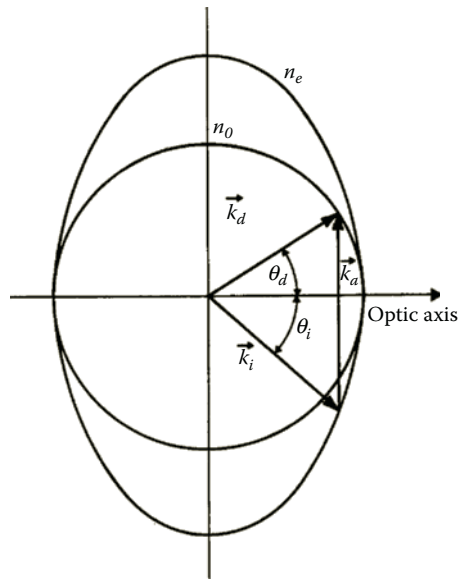
FIGURE 11.6
Vector diagram for diffraction in birefringent medium.

anisotropic Bragg angles were derived as

$$\sin q_i = \frac{1}{2n_i} \frac{1_0 f}{n} \left[1 + \left(\frac{n}{1_0 f} \right)^2 (n_i^2 - n_d^2) \right] \tag{11.30}$$

$$\sin q_d = \frac{1}{2n_d} \frac{1_0 f}{n} \left[1 - \left(\frac{n}{1_0 f} \right)^2 (n_i^2 - n_d^2) \right] \tag{11.31}$$

where n_i and n_d are the refractive indices corresponding to the incident and the diffracted light polarizations, and f is the acoustic frequency,

$$f = \frac{n}{\Lambda} \tag{11.32}$$

These angles are plotted in Figure 11.7 about that frequency f_m, for which there is a minimum in the angle of incidence. These curves, the general shapes of which are similar for all birefringent crystals, have a number of interesting characteristics that are useful for several types of AO devices. The minimum frequency for which an interaction may take place corresponds to $\theta_i = 90°$ and $\theta_d = -90°$, for which all three vectors will be collinear, as shown in Figure 11.8. It is easily shown that for this case, since the vector equation for conservation of momentum can be written as a scalar equation,

$$|k_i| + |K| = |k_d| \tag{11.33}$$
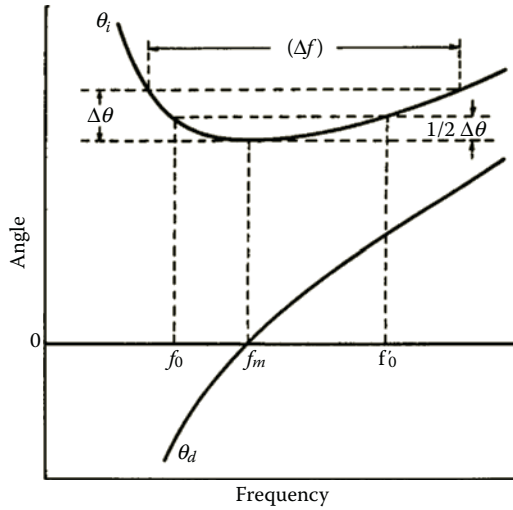
FIGURE 11.7
Angles of incidence and diffraction for anisotropic biréfringent diffraction.
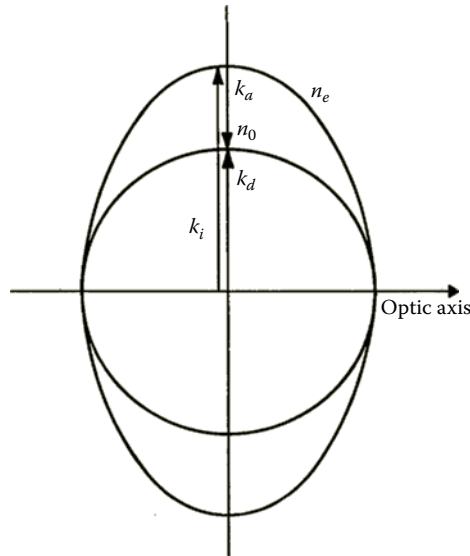
FIGURE 11.8
Vector diagram for collinear diffraction in biréfringent medium.

the frequency for which collinear diffraction takes place is

$$f = \frac{n(n_i - n_d)}{\lambda_0} \tag{11.34}$$

Such collinear phase matching has been used as the basis of an important device, the electronically tunable AO filter.[6] Note that if the incident light had been chosen as ordinary rather than extraordinary polarized, the sense of the acoustic vector K would be reversed.

In fact, the roles of the two curves in Figure 11.7 would be reversed by interchanging n_i and n_d.

Another interesting region of anisotropic diffraction occurs at the minimum value in the curve representing θ_i, at which frequency $\theta_d = 0$. This frequency, f_m, is obtained by setting the quantity in brackets in Equation 11.31 equal to zero:

$$f = \frac{u}{\lambda_0}\sqrt{n_i^2 - n_d^2} \qquad (11.35)$$

The significance of this point is that the angle of incidence of a scanned beam is relatively insensitive to change over a very broad range of frequencies. This frequency has important implications for the design of scanners because the bandwidth can be much greater than for a comparable isotropic scanner. Because the incident beam angle reaches a minimum value while the diffracted beam angle passes through zero at this point, and increases approximately linearly with frequency, the Bragg angle matching can be maintained over a large range, as will be described later. It will be seen that it is very difficult to achieve an interaction bandwidth this large by any other method.

The description of the interaction of light with sound we have given above is perhaps the simplest in terms of giving an intuitive understanding of the phenomena. Other descriptions, with totally different mathematical formalisms, have been carried out, and these lead to many details and subtleties in the behavior of AO systems that are beyond the scope of this book. Exact calculations have been carried out to extend the range of validity[7] from the limits allowed by the Raman–Nath theory,[8] and this has been experimentally investigated.[9] Other studies have also been carried out to give accurate numerical results for the intensity distribution of light in the various diffraction orders.[10] The diffraction process has been reviewed and analyzed by Klein and Cook,[11] using a coupled mode formulation, and there is continuing recent interest in refining the plane-wave scattering theory to give explicit results for intermediate cases.[12,13] Finally, the AO interaction can be viewed as a parametric process in which the incident optic wave mixes with the acoustic wave to generate polarization waves at sum and difference frequencies, leading to new optical frequencies; this approach has been reviewed by Chang.[14]

One of the most remarkable materials to have appeared recently for AO applications is paratellurite (TeO_2).[15] It has a unique combination of properties, which leads to an extraordinarily high figure of merit for a shear-wave interaction in a convenient RF range. It will be recalled that the anisotropic Bragg relations of Equations 11.30 and 11.31 led to a particular frequency, given by Equation 11.35, for which the angle of incidence is a minimum and therefore satisfies the Bragg condition over a wide frequency range. However, typical values of birefringence place this frequency around 1 GHz or higher. Of particular interest in TeO_2 is its optical activity for light propagating along the c-axis, or (001) direction; the indices of refraction for left- and right-hand circularly polarized light are different, so that plane polarized light undergoes a rotation of its plane of polarization by an amount

$$R = \frac{2n_0}{\lambda}\delta d \qquad (11.36)$$

where δ is the index splitting between left- and right-hand polarized light,

$$\delta = \frac{n_1 - n_r}{2n_0} \qquad (11.37)$$

Just as acoustic shear waves can phase-match two linearly polarized light waves, they can also phase-match two oppositely circularly polarized light waves. Thus, shear waves propagating in the (110) direction, with shear polarization in the (110) direction, will diffract left- or right-hand polarized light propagating along the (001) direction, one into the other. The anisotropic Bragg relations apply to crystals with optical activity, where the birefringence is interpreted as

$$\Delta n = n_1 - n_r = 2n_0 d \tag{11.38}$$

and the value of δ obtained from specific rotation is wavelength dependent. For the light and sound wave propagation directions described above, the acoustic velocity is 0.62×10^5 cm/s (0.24×10^5 in/s) and the figure of merit, M_2, is 515 relative to fused quartz. The frequency for which the Bragg angle of incidence is a minimum, as evaluated from Equation 11.35 for $\lambda = 0.633$ μm, is $f = 42$ MHz, a very convenient frequency. For other important wavelengths, the minima occur at 36 MHz for $\lambda = 0.85$ μm and at 22 MHz for $\lambda = 1.15$ μm.

The application of the anisotropic Bragg equations to optically active crystals was discussed in detail by Warner et al.[16] They showed that near the optic axis the indices of refraction are approximated by the relations (for right-handed crystals, $n_r < n_1$)

$$\frac{n_r^2(q)\cos^2 q}{n_0^2(1-d)^2} + \frac{n_r^2(q)\sin^2 q}{n_1^2} = 1 \tag{11.39}$$

and

$$\frac{n_1^2(q)\cos^2 q}{n_0^2(1+d)^2} + \frac{n_1^2(q)\sin^2 q}{n_0^2} = 1 \tag{11.40}$$

For incident angles near zero with respect to the optic axis and for small values of δ,

$$n_r^2 = n_0^2 \left(1 - 2d + \frac{n_1^2 - n_0^2}{n_1^2} \sin^2 q \right) \tag{11.41}$$

and

$$n_1^2 = n_0^2 \left(1 + 2d \cos^2 q \right) \tag{11.42}$$

For light incident exactly along the optic axis the two refractive indices are simply

$$n_r = n_0 (1 - d) \tag{11.43}$$

and

$$n_1 = n_0 (1 + d) \tag{11.44}$$

The anisotropic Bragg equations for optically active crystals are obtained by substitution of Equations 11.41 and 11.42 into Equation 11.30 and 11.31 for n_i and n_d. By ignoring the higher-order terms, this results in

$$\sin q_i \cong \frac{1f}{2n_0 n}\left[1 + \frac{4n_0^2 n^2}{1^2 f^2}d + \frac{\sin^2 q_r n_0^2}{1^2 f^2}\left(\frac{n_1^2 - n_0^2}{n_e^2}\right)\right] \tag{11.45}$$

and

$$\sin q_d \cong \frac{1f}{2n_0 n}\left[1 - \frac{4n_0^2 n^2}{1^2 f^2}d - \frac{\sin^2 q_e n_0^2}{1^2 f^2}\left(\frac{n_1^2 - n_0^2}{n_0^2}\right)\right] \tag{11.46}$$

The anisotropic Bragg angles (as measured external to the crystal) are shown in Figure 11.9 for TeO$_2$ at $\lambda = 0.6328$ μm. It is obvious that for frequencies around the minimum, it will be possible to achieve a much larger bandwidth for a given interaction length than is possible with normal Bragg diffraction; a one-octave bandwidth corresponds to a variation in angle of incidence for perfect phase matching of only 0.16°. A useful advantage of such operation is that large bandwidths are compatible with large interaction lengths, which avoids higher diffraction orders from Raman–Nath effects. For normal Bragg diffraction, on the other hand, large bandwidths can only be reached with interaction lengths that are so small that significant higher-order diffraction occurs. This decreases the efficiency with which light can be directed to the desired first order, and also limits the bandwidth to less than one octave in order to avoid overlapping low-frequency second-order with higher-frequency first-order diffracted light. However, tellurium dioxide operating in the
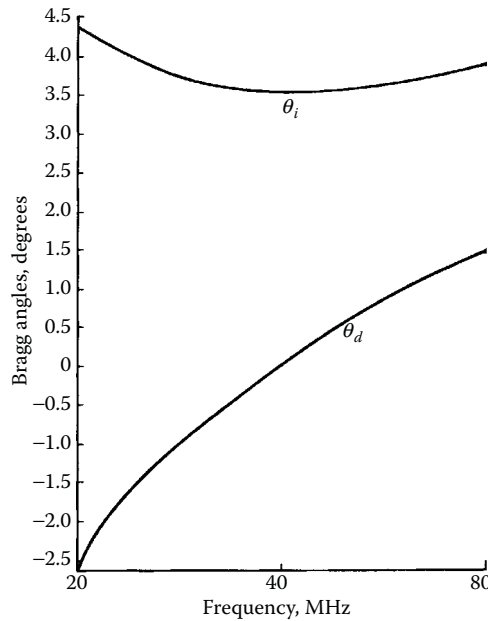
FIGURE 11.9
Bragg angles of incidence and diffraction (external) for anisotropic TeO$_2$ scanner, $\lambda = 0.6328$ μm.

anisotropic mode can always be made to diffract in the Bragg mode with no bandwidth limitation on length, so that this feature combined with the extraordinary high figure of merit leads to deflector operation with very low drive powers.

An important degeneracy occurs for anisotropic Bragg diffraction, which causes a pronounced dip in the diffracted light intensity at the midband frequency, where θ_i has its minimum. This degeneracy was explained by Warner et al.,[16] and is easily understood by referring to the diagram in Figure 11.10. Two sets of curves are shown in the figure; the solid pair represents θ_i and θ_d when the incident light momentum vector has a positive component along the acoustic momentum vector, and the dotted pair represents these angles when the incident light momentum vector has a negative component along the acoustic vector. In the former case, the frequency of the diffracted light is upshifted, and in the latter it is downshifted. The vector diagram for this process is shown in Figure 11.11. Light is incident to the acoustic wave of frequency f_0 at an angle θ_0, and is diffracted as a frequency upshifted beam, $(v + f_0)$, normal to the acoustic wave. This light, in turn, may be rediffracted; referring to Figure 11,.10, it can be seen that for a frequency f_0 light that is incident at $\theta = 0°$ can be rediffracted to either θ_0 or $-\theta_0$, In the former case, the light will be downshifted to the original incident light frequency v, and in the latter case it is upshifted to $v + 2f_0$. Note that this degeneracy can only occur at the frequency f_0 where light incident normal to the acoustic wave is phase matched for diffraction into both θ_0 and $-\theta_0$. How the
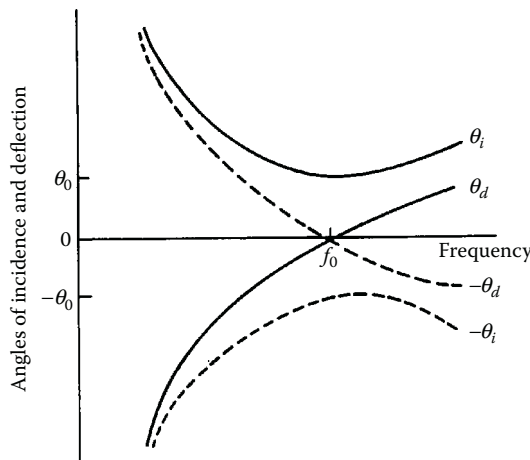
FIGURE 11.10
Angles of incidence and diffraction for anisotropic diffraction. Solid curves are for incident light having a component in the same direction as the acoustic wave, and dotted curves are for incident light having a component in the opposite direction.
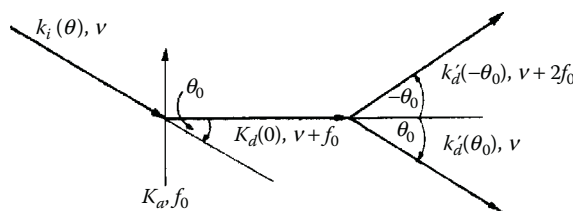
FIGURE 11.11
Vector diagram for midband degeneracy of Bragg diffraction in birefringent medium.
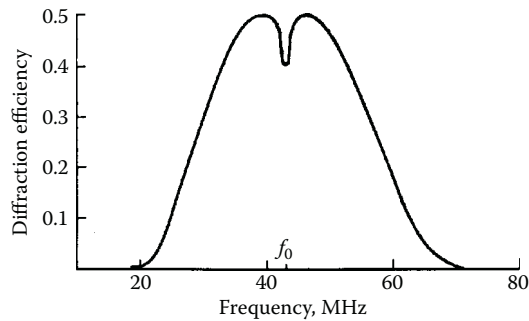
FIGURE 11.12
Effect of midband degeneracy on diffraction efficiency, for a maximum efficiency of 50%.

light is distributed in intensity between the three modes will depend upon the interaction length and the acoustic power level. The exact solution to this is found by setting up the coupled mode propagation equations under phase-matched conditions. The result of this is that maximum efficiency for deflection into the desired mode at f_0 is 50%. At low acoustic power, the deflection of light into the undesired mode is negligible; at high powers the unwanted deflection increases so that, for example, if the efficiency is 50% for frequencies away from f_0, it will be 40% at f_0. The theoretical response of such a deflector is shown in Figure 11.12, and is in excellent agreement with experimental results.

11.3 ACOUSTO-OPTIC MODULATOR AND DEFLECTOR DESIGN

11.3.1 Resolution and Bandwidth Considerations

Resolution, bandwidth, and speed are the important characteristics of AO scanners, shared by all types of scanning devices. Depending upon the application, only one, or all of these characteristics may have to be optimized; in this section, we will examine which AO design parameters are involved in the determination of resolution, bandwidth, and speed. Consider an AO scanner with a collimated incident beam of width D, diffracted to an angle θ_0 at the center of its bandwidth Δf. If the diffracted beam is focused onto a plane by a lens, or lens combination, at the scanner, the diffraction spread of the optical beam will be

$$dx = F df \cong \frac{F \lambda}{D} \tag{11.47}$$

where F is the focal length of the lens. The light intensity will be distributed in the focal plane as illustrated in Figure 11.13. As an example for diffraction-limited optics, the spot size for a 25 mm (1 in) wide light beam of wavelength 6.33 µm at a distance of 30 cm (12 in) from the delay line is 7.6 µm (3×10^{-4} in). There are, however, aberrations that prevent this from being fully realized, as will be discussed later.

The number of resolvable spots will be the angular scan range divided by the angular diffraction spread,

$$N = \frac{\Delta q}{df} \tag{11.48}$$
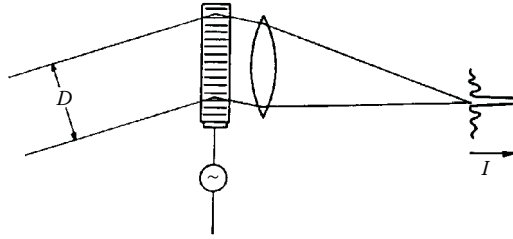
FIGURE 11.13
Distribution of light intensity due to diffraction by acoustic field.

where $\Delta\theta$ is the range of the angular scan. Differentiating the Bragg angle formula yields

$$\Delta q = \frac{1}{n\cos q_0}\Delta f$$

(11.49)

and

$$N = (\Delta f)\left(\frac{D}{n\cos q_0}\right) = \Delta f\, t$$

(11.50)

where τ is the time that it takes the acoustic wave to cross the optical aperture. The resulting expression is the time–bandwidth product of the AO scanner, a concept applied to a variety of electronic devices as a measure of information capacity. The time–bandwidth product of an AO Bragg cell is equivalent to the number of bits of information that may be instantaneously processed by the system. For an acousto-optic modulator (AOM) that is strictly a temporal modulator, a time–bandwidth product near unity is generally desired as the goal is often to modulate as fast as possible and therefore to minimize the aperture delay time. In contrast, for an AOD the time–bandwidth product is generally desired to be as large as possible to produce a large number of resolution elements.

There are two factors limiting the bandwidth of an AO device: the bandwidth of the transducer structure and the acoustic absorption in the delay medium. The acoustic absorption increases with increasing frequency; for high-purity single crystals the increase generally goes with the square of the frequency. For glassy materials, on the other hand, the attenuation will increase more slowly with frequency, often approaching a linear function. The maximum frequency is generally taken as that for which the attenuation of the acoustic wave across the optical aperture is equal to 3 dB. A reasonable approximation of the maximum attainable bandwidth is $\Delta f = 0.7 f_{max}$, so that we may derive some relationships for the maximum number of resolution elements.

For a material with a quadratic dependence of attenuation on frequency,

$$a(f) = \Gamma f^2$$

(11.51)

and the maximum aperture for 3 dB loss is

$$D = \frac{3}{\Gamma f^2}$$

(11.52)

Using these results, the maximum number of resolution elements is

$$N_{\max} \simeq q\sqrt{\frac{1.5D}{u^2\Gamma}}$$ (11.53)

from which it can be seen that, in principle, it is always advantageous to make the delay line as long as possible. In practice, the aperture will be limited by the largest crystals that can be prepared, or ultimately by the size of the optical system. For a glassy material for which the attenuation increases linearly with frequency,

$$a(f) = \Gamma'f$$ (11.54)

and the maximum number of resolvable spots will be

$$N_{\max} \simeq \frac{2}{\Gamma'u}$$ (11.55)

which is independent of the size of the aperture, being determined only by the material attenuation constant and the acoustic velocity.

In the next section we will review material considerations in some detail, and see what the performance limits are of currently available AO materials. As a numerical example, however, the highest-quality fused quartz has an attenuation of about 3 dB/cm at 500 MHz and an acoustic velocity of 5.96×10^5 cm/s (2.35×10^5 in/s) (for longitudinal waves), leading to $N_{\max} = 560$.

11.3.2 Interaction Bandwidth

The number of resolution elements will be determined by the frequency bandwidth of the transducer and delay line, but a number of other bandwidth considerations are also of importance for the operation of a scanning system. While a large value of τ leads to a large value *of N,* the speed of the device is just equal to l/τ. That is, the position of a spot cannot be changed randomly in a time less than τ. If the acoustic cell is being used to temporally modulate the light as well as to scan, then obviously the modulation bandwidth will similarly be limited by the travel time of the acoustic wave across the optical aperture. In order to increase the modulation bandwidth, the light beam must be focused to a small width, ω, in the acoustic field. The 3 dB modulation bandwidth is approximately

$$\Delta f = \frac{0.75}{t} = \frac{0.75u}{w}$$ (11.56)

and the diffraction-limited beam waist (the $1/e^2$ power points) of a Gaussian beam is

$$w_0 = \frac{2 l_0 F}{pD}$$ (11.57)

where D is the incident beam diameter and F is the focal length of the lens. With this value of beam waist, the maximum modulation bandwidth is

$$\Delta f = 0.36p\frac{uD}{l_0 F}$$ (11.58)

It can be seen from Equation 11.58 that the modulation bandwidth for a diffraction-limited focused Gaussian beam can be very high; for example, for a material of acoustic velocity 5×10^5 cm/s (2×10^5 in/s) the bandwidth of a 0.633 μm light beam focused with an $f/10$ lens is about 1 GHz. Such a system, however, is practically useless, because the diffraction efficiency would be extremely small.

In order for the Bragg interaction bandwidth to be large, there must be a large spread of either the acoustic or the optical beam directions, $\delta\theta_a$ and $\delta\theta_0$, respectively, or both. This spread may occur either by focusing, which in the case of the acoustic beam is achieved by curving the plane of the transducer, or it may be due simply to the aperture diffraction for both beams. It follows from fairly simple arguments that the optimum configuration for the most efficient utilization of optical and acoustic energy corresponds to approximately equal angular spreading, $\delta\theta_0 \cong \delta\theta_a$, as illustrated in Figure 11.14.

For an AOD, the angular spread of the acoustic beam should be made large enough to match Bragg diffraction over the frequency range of the transducer-driving circuit bandwidth. As mentioned previously, this will result in some reduction in efficiency. To examine the relationship between bandwidth and the efficiency, we must first state another well-known result of acoustically diffracted light. This is, as shown by Cohen and Gordon,[17] that the angular distribution of the diffracted light will represent the Fourier transform of the spatial distribution of the acoustic beam. This Fourier transform pair is illustrated in Figure 11.15 for the usual case of the rectangular acoustic beam profile. It seems intuitively obvious for this simple case, in which the diffraction spread of the incident optical beam is ignored, that there will be components in the diffracted light corresponding to the acoustic field side lobes. It is shown in Reference 17 that the Fourier transform relationship holds for an arbitrary acoustic beam profile.

For the rectangular profile, the angular dependence of the diffracted light, illustrated in Figure 11.15, is

$$\frac{I(q)}{I_0} \propto \left[\frac{\sin \frac{1}{2} KL(q - q_B)}{\frac{1}{2} KL(q - q_B)} \right]^2 \tag{11.59}$$
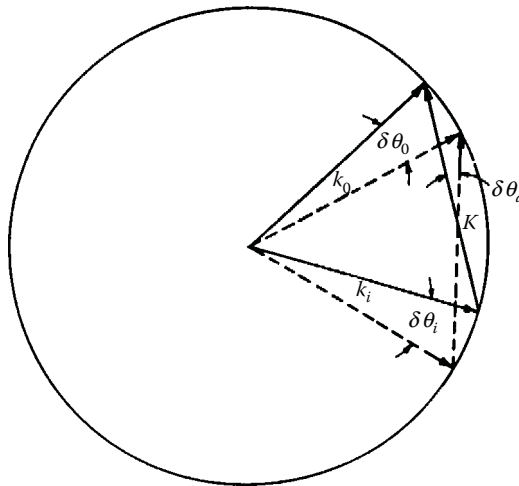
FIGURE 11.14
Vector diagram for Bragg diffraction in isotropic medium with angular spread of acoustic beam direction.
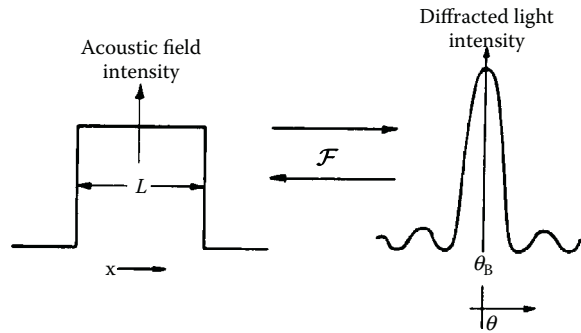
FIGURE 11.15
Fourier transform relationship between acoustic field intensity and diffracted light intensity.

for which the −3 dB points occur at

$$\frac{1}{2}KL(\Delta q)_{1/2} \quad \pm 0.45p \tag{11.60}$$

where $(\Delta\theta)_{1/2}$ is the value of $\theta - \theta_B$ at the half-power points. This yields a value for the angular width of the optical beam, just equal to the diffraction spread of the acoustic beam, namely

$$2(\Delta q)_{1/2} \quad \frac{1.8p}{KL} \tag{11.61}$$

The frequency bandwidth is obtained by equating this result to the differential of the Bragg condition:

$$dq = \frac{\lambda_0 \Delta f}{n u \cos q_B} \tag{11.62}$$

This result is

$$\Delta f = \frac{1.8 n u^2 \cos q_B}{L f_0 \lambda_0} \tag{11.63}$$

For AO scanning devices in which the bandwidth as well as the diffraction efficiency is of importance, a more relevant figure of merit may be the product of the bandwidth with the efficiency. By combining Equations 11.63 and 11.25, this product is

$$2 f_0 \Delta f \cdot \frac{I_1}{I_0} = \frac{1.8 p^2}{\lambda_0^3 H \cos q_B} \left[\frac{n^7 p^2}{r u}\right] P \tag{11.64}$$

The quantity in brackets can be regarded as the figure of merit of the material when the efficiency–bandwidth product is the important criterion, and is designated as

$$M_1 = \frac{n^7 p^2}{ru} \tag{11.65}$$

Other methods of achieving a large interaction bandwidth include transducer designs that steer the acoustic beam in direction in order to track the Bragg angle as it changes with frequency. A description of beam steering will be included in the section on transducers. Still another figure of merit was introduced by Dixon[18] in connection with wideband AO devices. Because the power requirements decrease as the transducer height H decreases, it is advantageous to make H as small as possible. If there are no limitations on the minimum size of H, it can be as small as the optical beam waist in the region of the interaction, h_{min}. The modulation bandwidth is determined by the travel time of the acoustic wave across this beam waist,

$$t \approx \frac{1}{\Delta f} = \frac{h_{min}}{n} \tag{11.66}$$

so that

$$h_{min} = \frac{n}{\Delta f} \tag{11.67}$$

Substitution of this value for H in Equation 11.62 results in the relation

$$2f_0 \frac{I_1}{I_0} = \frac{1.8 p^2}{l_0^3 \cos q_B} \left[\frac{n^7 p^2}{ru^2} \right] \tag{11.68}$$

and the appropriate figure of merit for this situation is the quantity in brackets:

$$M_3 = \left[\frac{n^7 p^2}{ru^2} \right] \tag{11.69}$$

Note that the optical wavelength appears as l_0^3 in both Equations 11.64 and 11.68, so that operation at long wavelengths is relatively more difficult in terms of power requirements for configurations optimizing bandwidth as well as efficiency.

11.3.3 Deflector Design Procedure

The useful optical aperture of a Bragg cell is usually considered to be that length across which the difference in acoustic attenuation between the highest and the lowest frequencies within the operating bandwidth of the cell is 3 dB. A particular application may dictate either a bandwidth or a resolution, that is, a time–bandwidth product. In general, an optimized Bragg cell design will maximize the number of resolvable spots, as well as other transducer structure parameters.

The number of resolvable spots, or the time–bandwidth product, will be determined by three key factors:[19] the acoustic attenuation Γ, the optical aperture of the AO crystal D, the angular beam spreading of the acoustic wave, which is determined by the transducer length L, and the acoustic wavelength. The constraints placed upon the number of resolvable spots N by these three factors is given by the relations[19]

$$N \leq \frac{1.5\Lambda_c}{\Gamma\Lambda_1^2} \tag{11.70}$$

$$N \leq \frac{D}{2\Lambda_c} \tag{11.71}$$

$$N \leq \left(\frac{L}{2\Lambda_c}\right)^2 \tag{11.72}$$

where Λ_c is the acoustic wavelength at the center frequency, Λ_1 is the acoustic wavelength at 1 GHz, and Γ is the acoustic attenuation in dB per unit length, normalized to 1 GHz (under the usual assumption that the attenuation increases quadratically with frequency). Note that Equation 11.70 allows for a 3-dB attenuation. Once the center frequency and the bandwidth of the cell have been determined, the transducer structure must be designed. This will include the electrode length L and height H. The length must be chosen so that it is small enough to allow sufficient beam spread to satisfy the Bragg angle matching requirements over the desired bandwidth (for a fixed angle of incidence of the optical beam). At the same time, the diffraction efficiency will decrease as L decreases, so that we will want L to be as large as possible within the interaction bandwidth constraint.

11.3.4 Modulator Design Procedure

AOD and AOM have very similar design requirements and in some cases, one design may be suitable for either application or both. While the key design parameter for deflectors is typically the number of resolvable spots, the key design parameter for modulators is typically rise time or modulation bandwidth. These differing design parameters lead to the characteristic that, for deflectors, the optical aperture is typically made as large as possible, while for modulators, the optical beam is made as small as possible.

The rise time of an AOM is fundamentally limited by the acoustic velocity of the modulator material. When an acoustic pulse is transmitted from the transducer, diffraction will begin when the leading edge of the pulse reaches the optical beam. Full, diffracted beam power will not be obtained until the acoustic wavefront reaches the opposite end of the optical beam. The shape of the rising optical pulse will depend on the shape of the optical beam.

For a Gaussian optical beam, the time required for the acoustic wave to cross the $1/e^2$ beam diameter is

$$t = \frac{D_{1/e}^2}{V_a} \tag{11.73}$$

where V_a is the acoustic velocity. This beam width corresponds to a rise time from 2.3% to 97.7%. The more conventional rise time from 10% to 90% is calculated as

$$t_R = 0.64\,t \tag{11.74}$$

For video modulation applications, the rise time limits the frequency response of the modulator. The modulator bandwidth can be expressed as the frequency at which 3 dB roll-off occurs, which is estimated by the standard relation

$$f_0 = \frac{0.35}{t_R}$$

(11.75)

In the case of square pulse video modulation, the modulation speed may be defined by a specific dynamic extinction ratio requirement. For square wave modulation at frequency f_0, the dynamic extinction ratio is approximately 10:1. For high extinction ratios on the order of 1000:1, the maximum square wave modulation frequency is approximately $f_0/2$. For a given rise time, the design beam diameter can be calculated using the above relations. Note that the optical beam cannot be made arbitrarily narrow because the beam must remain relatively collimated over the acoustic interaction length L. If the beam waist is too small, the beam divergence over the length L will result in a longer rise time than predicted based on the beam waist. The minimum value of L is constrained by the need to stay in the Bragg regime (Sec. 2.2) and achieve a specified diffraction efficiency (Sec. 2.3).

The above discussion of dynamic extinction ratio assumes that the static extinction ratio is not limiting. The limit of static extinction ratio is determined by the ability to discriminate the diffracted beam from undiffracted or scattered light. Scattered light is a function of the quality of the material and the surface finish of the AO cell and is typically the limiting parameter for static extinction ratio when the diffracted beam separation is made large enough. The beam separation between the zero- and first-order beams is given by

$$\Delta f = \frac{\lambda}{\Lambda}$$

(11.76)

If the separation angle is made equal to the full divergence angle of the optical beam and a knife edge placed halfway between the zero- and first-order beams, approximately 2.3% of the blocked beam will pass the knife edge. This means the minimum static extinction ratio would be about 40:1. If the beam separation is increased to twice the beam angle, the optical power passing the knife edge is decreased to 0.003%. For most applications, this amount of beam separation is sufficient to make the extinction ratio limitation due to beam separation negligible. Using this condition with the formula of divergence of a Gaussian beam gives

$$\Delta f > \frac{8\lambda}{pD_0}$$

(11.77)

Combining the above two equations gives the following condition for beam separation

$$\Lambda > \frac{pD_0}{8}$$

(11.78)

Another consideration is that the angular acceptance window of the acoustic field should be large enough to allow Bragg interaction over the optical beam. If the angular acoustic field is too narrow, the optical beam will be apodized in angle, resulting in output beam

distortion and decreased diffraction efficiency. The angular acoustic intensity from a rectangular transducer is described by

$$I(q) = \operatorname{sinc}^2\left[\frac{q}{(\Lambda/L)}\right] \tag{11.79}$$

The null-to-null width is therefore $2\pi\Lambda/L$. This width should be much greater than the $1/e^2$ beam angle to maintain good diffraction efficiency and prevent distortion. Using the formula for beam divergence, the condition for L becomes

$$L \ll \frac{p^2 \Lambda D_0 n}{2\lambda} \tag{11.80}$$

Remember that for Bragg interaction L cannot be made arbitrarily small. In practice, the length is chosen to be as short as tolerable based on requirements for efficiency and suppression of higher orders, typically such that Raman–Nath parameter $Q \approx 12$.

11.4 SPECIALIZED ACOUSTO-OPTIC DEVICES FOR SCANNING

11.4.1 Acoustic Traveling Wave Lens

Most AO applications are based on diffraction effects, requiring interaction over at least a few periods sinusoidal index change in the acoustic media. However, it is also possible to use the index change over a fraction of a period to act as a lens to focus light by refraction. In this case, the index change produced by a segment of an acoustic wave can be viewed as a gradient-index cylinder lens moving with the speed and direction of the acoustic wave.

Consider a conventional scanning system with a single axis scan device, in this case an AO Bragg deflector, followed by a scan lens. The total number of resolvable spots in the scan is determined by the aperture size and scan angle of deflector, and can be determined approximately as the product of the bandwidth and the acoustic transit time (Equation 11.50). The scan lens can be altered to change the scan length and spot size, but the total number of spots remains the same. However, the number of spots may be increased by adding a traveling lens after the scan lens as shown in Figure 11.16. In this case, the speed and timing of the deflector is synchronized with the speed and phase of the acoustic traveling wave such that the input beam tracks the acoustic lens as it propagates in the scan direction. The acoustic lens reduces the scanned spot size but has no effect on the scan length, thereby producing spot gain.

11.4.1.1 Design Considerations

The sinusoidal index variation is approximately parabolic near the index minimum and acts as a focusing (positive) lens. A quarter-wave aperture centered on the index minimum will provide the lens with near diffraction-limited performance. The focal length of the traveling lens was derived by Foster:[20]

$$F = \left(\frac{\Lambda}{4}\right)\left(\frac{n_0}{\Delta n}\right)^{1/2} \tag{11.81}$$
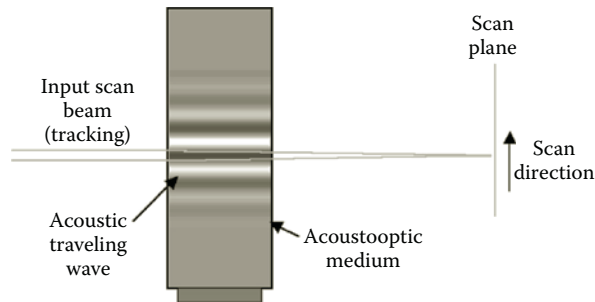
FIGURE 11.16
Application of an acoustic traveling-wave lens.

where Λ is the acoustic wavelength, n_0 is the index of refraction, and Δn is the peak refractive index variation. The focal length *(F)* in Equation 11.81 refers to the focal distance internal to the lens. It can also be viewed as the lens thickness required to produce a quarter-pitch lens. If the thickness of the lens is less than *F*, the effective focal length will be longer.

Foster also derived an expression for the *f*-number of the lens by considering the path of the extreme rays at $\pm\Lambda/8$:

$$f\text{-}number = \frac{F}{D} = \frac{2}{p\left(n_0 / \Delta n\right)^{1/2}} \tag{11.82}$$

Assuming a Gaussian beam of diameter D_0 $(1/e^2)$ is input to the traveling lens, the size of the focused spot is estimated as

$$D_1 = \frac{4\,\mathit{1}F}{pn_0 D_0} \tag{11.83}$$

Combining Equations 11.81, 11.82, and 11.83 gives an expression for the output spot size from the traveling lens:

$$D_1 = \frac{8\,\mathit{1}}{p^2 \left(n_0 / \Delta n\right)^{1/2}} \tag{11.84}$$

The above derivation uses the assumptions that the lens thickness is approximately equal to *F* and the input beam diameter (D_0) is equal to $\Lambda/4$. In an application where the traveling lens is the final scan lens, *F* is typically made less than the lens thickness such that the focus occurs outside the lens and with a back focal distance sufficient to reach the scan plane.

As an example, assume that a prescanner produces a linear scan of 50 mm with a 0.5-mm diameter beam of wavelength 633 nm. A traveling acoustic lens device made from dense flint glass (SF-59) will be added to provide a final spot size of 0.05 mm or spot gain of 10. The required refractive index change (Δn) is determined to be 0.000157 from Equation 11.84. The acoustic wavelength will be $4D_0$ or 2 mm, and therefore *F* is determined to be 55.7 mm from Equation 11.81.

The length of the acoustic transducer (L), corresponding to the thickness of the lens, is chosen to be 45 mm, which is slightly less than F such that the focus can be outside the traveling lens device. The height of the transducer (H) is chosen to be 15 mm, such that the acoustic near field is longer than the scan length to prevent excessive acoustic loss along the scan direction due to diffraction spreading.

The amount of refractive index change is proportional to the square root of the acoustic intensity:

$$\Delta n = \left(\frac{M_2 P_A}{2} \right)^{1/2}$$

(11.85)

where M_2 is the AO figure of merit and P_A is the acoustic intensity. Using Equation 11.85, the acoustic intensity required is 2.6 W per mm². The instantaneous power required is therefore 1800 W.

While the power requirement is very large compared to those for typical Bragg cells, the average power required can be reduced significantly by pulsing the acoustic signal once per scan line. Even so, the large instantaneous and average power requirements for traveling acoustic wave devices are a significant challenge to practical implementation. One way to reduce the power requirement is to narrow the AO cell height to a fraction of an acoustic wavelength to form an acoustic slab waveguide. The waveguide properties eliminate the acoustic diffraction spreading problem and associated need for a tall transducer.

11.4.2 Chirp Lens

In a typical AOD application, the transducer frequency is swept linearly over a range Δf to provide a linear angular scan ($\Delta\theta$) of the output beam. This frequency sweep is referred to as a chirp. The equivalent length of the acoustic chirp is equal to the product of the chirp time and the acoustic velocity. If the aperture were large enough to cover the entire acoustic chirp, then the diffracted angle would vary over the range $\Delta\theta$ along the chirp length as shown in Figure 11.17. Using the small angle approximation, the acoustic chirp acts as a lens with f-number inversely proportional to $\Delta\theta$:

$$f\text{-number} \approx \frac{1}{\Delta q} \quad \text{Acoustic apertute} \geq \text{Chirp length}$$

(11.86)
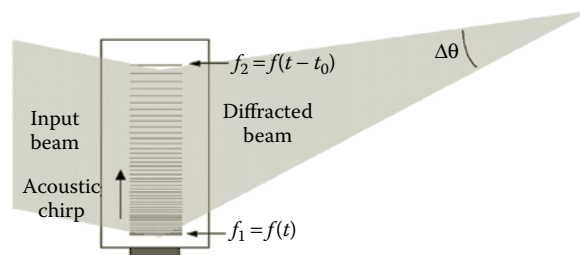
FIGURE 11.17
Focusing effect of an acoustic chirp.

If the aperture is smaller than the chirp length, as is typically the case, then the *f*-number will be inversely proportional to the fraction of the chirp length covered by the aperture:

$$f\text{-number} \approx \frac{1}{\Delta q} \frac{T_{\text{chirp}}}{T} \quad \text{Acoustic apertute} < \text{Chirp length} \tag{11.87}$$

For a typical beam deflector application, the chirp length T_{chirp} is much larger than the acoustic aperture time T. In this case, the focusing power of the chirp will be much less than that of an f–θ lens, and can be neglected. For very fast scanners, where T_{chirp} approaches T, the chirp focusing effect may contribute to the number of scan spots. Note that if $T_{\text{chirp}} = T$, the chirp focusing effect is equivalent to the power of an f–θ lens. However, this arrangement does not make a useful scanner, as the scan time is approximately the same as the access time.

By making the chirp smaller than the aperture length, the chirp can be used as a traveling lens. The chief application of a traveling chirp lens is to employ it as a post scan lens as described in the previous section. The traveling chirp lens is diffractive, and does not require the high instantaneous acoustic power needed to produce the refractive quarter-wave lens with comparable *f*-number. It also allows more flexibility with lens aperture, as the lens size is a function of chirp time and not a function of acoustic frequency. Unlike the refractive lens, the chirp lens is subject to diffractive losses and the diffraction efficiency may not be uniform across the aperture. Another disadvantage is that the quality of the lens is a function of the linearity of the chirp signal, such that phase error in the chirp signal will translate into aberrations in the lens.

11.4.3 Multichannel Acousto-Optic Modulator

AOMs are often used in conjunction with a scan beam to produce a raster scan. The scan line rate may be limited by the response time of the AOM, the exposure time, or the scan rate of the scan generating components. One way to increase throughput for a raster scan is to use multiple beams in a parallel scan arrangement. The beams may pass through a common set of scan elements, such that the final scans are identical except for position offset at the scan plane. Although scanned together, each beam must have its own modulation sequence. This can be accomplished by using a multichannel AOM.

Multichannel AOM can be fabricated using the same number of steps as single-channel modulators, with the main difference being that an array of electrodes is deposited on the transducer substrate instead of a single electrode. In application, a parallel array of beams is registered to the transducer array of the modulator as shown in Figure 11.18.

The effect of one channel upon another is of particular concern for multichannel devices, especially as larger numbers of channels are integrated in a single device. Cross talk between channels can occur through a number of mechanisms including electrical cross talk in the feed circuitry or electrodes and acoustic cross talk between adjacent channels. Another cross-talk mechanism is thermal, where the heat load added from turning on one or more channels causes a thermo-optic or strain-optic effect that alters the output of an independent channel. Electrical cross talk is controlled by using good RF design practices including use of controlled impedance microstrip lines and providing good ground continuity from the feed circuitry to the transducer. Acoustic cross talk is dependent on the location of the optical aperture and the spacing between electrodes. Because the acoustic field in isotropic materials will spread once it propagates beyond the near-field zone
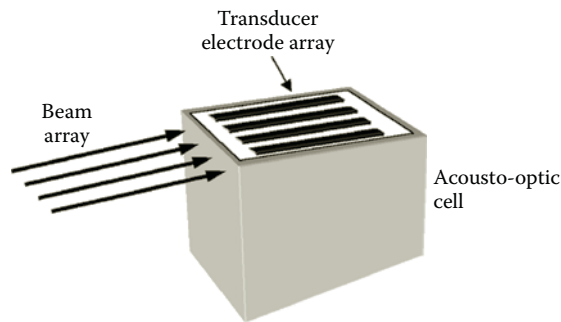
FIGURE 11.18
Multiple transducer channels on a monolithic acousto-optic device.

adjacent to the transducer, the amount of acoustic overlap is generally greater if the optical aperture is placed further from the transducer.

The degree of acoustic overlap between adjacent channels can also be manipulated through the electrode design. For an electrode with a simple rectangular shape, the acoustic intensity will have a $sinc^2$ angular distribution, which has side lobes 13 dB down from the main lobe. Other shapes, such as diamond or Gaussian envelope, can produce much lower side lobe values, although sometimes at the expense of faster spread of the main lobe.

11.5 MATERIALS FOR ACOUSTO-OPTIC DEVICES

11.5.1 General Considerations

We have seen in the preceding section that two important criteria for choosing materials for AO scanning systems are the AO figure of merit and the high-frequency acoustic loss characteristics. Other properties that determine the usefulness of a material are its optical transmission range, optical quality, availability in suitable sizes, mechanical and handling characteristics as they may pertain to polishing and fabrication procedures, and chemical stability under normal conditions. As with most components, cost will be an important factor, even when all the other factors may be positive, if competing techniques are available.

One of the limitations on the use of AO scanners before the late 1960s was the availability of materials with reasonably high figure of merit. As we have seen, fused quartz, which is used as the standard for comparison, has a figure of merit so low that only a few percent diffraction efficiency can be obtained for scanners of typical dimensions, and with RF powers that can be applied without causing damage to transducer structures. Water is a fairly efficient material, with a figure of merit about 100 times larger than fused quartz and has actually found use in some scanning systems. As with most liquids, it cannot be used at frequencies higher than about 50 MHz, so that large numbers of resolution elements cannot be achieved. Since the late 1960s, many new materials have been synthesized and existing ones were found to have excellent properties. Materials can now be found for most scanning applications from the UV through the intermediate IR where high bandwidth is required.

The selection of a material for any particular device will be dictated by the type of operation under consideration. In general, it is desirable to select a material with low-drive-power requirements, suggesting those with large refractive index and low density and acoustic velocity. If, however, high-speed modulation is of paramount importance, then a low acoustic velocity may lead to slower than required speeds. In the following section, we will review the factors and trade-offs involved in the selection of materials for various AO applications. Whatever the particular material requirements may be, there are also a number of practical considerations that dictate several generally important material properties whatever the application: (1) the optical quality must be high so that not only absorption but scattering and large-scale inhomogeneities are small; (2) good chemical stability is required so that protective enclosures are not needed to maintain integrity; (3) good mechanical properties are required so that the device can be cut and polished without extraordinary procedures and can be adjusted and used with normal handling techniques; (4) the availability of crystal growth methods for obtaining suitably large, high-quality boules with reasonable cost is needed; and (5) a low-temperature coefficient of velocity is required to avoid drift of scan properties.

11.5.2 Theoretical Guidelines

There is no simple microscopic theory of the photoelastic effect in crystals. Therefore it is not possible to predict the magnitude of the photoelastic constants from first principles. However, Pinnow[21] has suggested the use of certain empirical relationships between the various physical properties in order to systematize and group AO materials. It is well known that such relations exist, for example, for the refractive index and the acoustic velocity for such groups as the alkali halides, the mineral oxides, and the III–IV compounds.

A large amount of data has been collected on the refractive indices of crystals, and generally good agreement is found with the Gladstone–Dale[22] equation

$$\frac{n-1}{r} = \sum_i q_i E_i \tag{11.88}$$

in which R_i is the specific refraction of the ith component and q_i is percentage by weight. Reliable values of R_i have been determined from mineralogical data over many years. From the expression for the AO figure of merit, it is apparent that a high value of refractive index is desirable for achieving high diffraction efficiency. It is not, however, possible simply to select for consideration those materials with high refractive index, as even a casual survey shows that such materials tend to be opaque at shorter wavelengths. This trend was examined in great detail by Wemple and DiDomenico,[23] who found that the refractive index is simply related to the energy band gap. The semiempirical relation for oxide materials is

$$n^2 = 1 + \frac{15}{E_g} \tag{11.89}$$

where E_g is the energy gap (expressed in electron volts). For other classes of materials the energy gap constant will be different, but the same form holds. It can be seen from Equation 11.89 that the largest refractive index for an oxide material transparent over the entire visible range (cutoff wavelength at 0.4 μm) is 2.44. Higher refractive indices can be chosen only by sacrificing transparency at short wavelengths.

Pinnow[21] has found that a good approximation to the acoustic velocity for a wide range of materials is obtained with the relation

$$\log\left(\frac{n}{r}\right) = -b\bar{M} + d \tag{11.90}$$

where $\bar{M}$ is the mean atomic weight, defined as the total molecular weight divided by the number of atoms per molecule, and b and d are constants. Large values of d are generally associated with harder materials, while b does not vary greatly for oxides. Thus, in general, low acoustic velocities tend to be found in materials of high density, as is intuitively expected. Another useful velocity relationship has been pointed out by Uchida and Niizeki;[24] this is the Lindemann formula relating the melting temperature T_m and the mean acoustic velocity v_m,

$$n_m^2 = \frac{cT_m}{\bar{M}} \tag{11.91}$$

in which c is a constant dependent upon the material class. This relation suggests that high-efficiency materials would likely be found among those with large mean atomic weight and low melting temperature, that is, dense, soft materials.

In order for an AO material to be useful for wideband applications, the ultrasonic attenuation must be small at high frequencies. An attenuation that is often taken as an upper limit is 1 dB/µs (so that the useful aperture will depend upon the velocity). Many materials that might be highly efficient and otherwise suitable are excessively lossy at high frequency. A microscopic treatment of ultrasonic attenuation was carried out by Woodruff and Ehrenreich.[25] Their formula for the ultrasonic attenuation is

$$a = \frac{g^2 \Omega^2 kT}{r n^5} \tag{11.92}$$

where Ω is the radian frequency, γ is the Griinneisen constant, κ is the thermal conductivity, and T is the absolute temperature. This formula would suggest that the requirement of low acoustic velocity and low attenuation conflict with each other, since $\alpha \sim v^{-5}$; it is quite unusual for materials with low acoustic velocity to not also have a high absorption, at least for the low-velocity modes.

The determination of the photoelastic constants of materials is essentially an empirical study, although a microscopic theory of Mueller,[26] developed for cubic and amorphous structures, is still referenced. For both ionic and covalent bonded materials the photoelastic effect derives from two mechanisms: the change of refractive index with density, and the change in index with polarizability under the strain. Both of these effects may have the same or opposite sign under a given strain, and one or the other may be the larger. It is for this reason that the magnitude or even the sign of the photoelastic constant cannot be predicted, since the effects may completely cancel each other. It is possible, however, to estimate the maximum constants for groups of materials. This has been done for three important groups with the result

$$|P_{\max}| = \begin{cases} 0.21 & \text{water-insolube oxides} \\ 0.35 & \text{water-solube oxides} \\ 0.20 & \text{alkali halides} \end{cases}$$

In general, the photoelastic tensor components corresponding to shear strain will be less than those corresponding to compressional strain because there is no change, to first order, of density with shear; only the polarizability effect will be present. It is always possible that exceptionally large values of shear-related photoelastic coefficients may be found, but in no case could they be expected to be larger than the estimated value of $|P_{max}|$. The maximum values of photoelastic constant are shown in Table 11.1 for a number of important oxides and other materials.

11.5.3 Selected Materials for Acousto-Optic Scanners

Among older materials, those that have been shown useful for AO applications are fused quartz, because of its excellent optical quality and low cost for large sizes, and sapphire and lithium niobate, because of their exceptionally low acoustic losses at microwave frequencies. For IR applications germanium[27] has proven very useful, as has arsenic trisulfide glass, where bandwidth requirements are not high. Among the newer crystal materials, very good AO performance has been obtained in the visible with GaP[28] and PbMoO$_4$.[29,30] One of the most interesting new materials to be developed within the past several years is TeO$_2$,[31] which along with PbMoO$_4$ has found wide use in commercially available AO scanners. More design details for devices employing this material will be given later. Among the new materials that have been developed for IR applications, very high performance has been reached with several chalcogenide crystals.[32] Particularly important members of this group of materials include Tl$_3$AsS$_4$[33] and Tl$_3$PSe$_4$.[34] The compound Tl$_3$AsSe$_3$[35] is particularly interesting beyond its possible use as an IR AOM material. Since Tl$_3$AsSe$_3$ belongs to the crystal class 3m, its symmetry permits it to possess a nonzero p_{41} photoelastic coefficient, and it is suitable for use as a collinear tunable AO filter, a device first realized by Harris,[36] using lithium niobate. Tables 11.2 through 11.4 summarize the properties of some

TABLE 11.1

Maximum Photoelastic Coefficients

Material	$\mathbf{|P_{max}|}$ **measured**
LiNbO$_3$	0.20
TiO$_2$	0.17
Al$_2$O$_3$	0.25
PbMoO$_4$	0.28
TeO$_2$	0.23
Sr$_{.5}$Ba$_{.5}$Nb$_2$O$_6$	0.23
SiO$_2$	0.27
YIG	0.07
Ba(NO$_3$)$_2$	0.35
α-HIO$_3$	0.50
Pb(NO$_3$)$_2$	0.60
ADP	0.30
CdS	0.14
GaAs	0.16
As$_2$S$_3$	0.30

Source: Klein W.R.; Hiedemann, E.A. *Physica* 1963, *29,* 981.

TABLE 11.2

Acousto-Optic Properties of Amorphous Materials

Material	Transmission range (μm)	Acoustic mode	ν (cm/s × 10⁵)	Γ (dB/ cm GHz²)	Opt. pol. dir.	n (0.633 μm)	M_1 (cm² s/g × 10⁻⁷)	M_2 (s³/g × 10⁻¹⁸)	M_3 (cm s²/g × 10⁻¹²)
Water	0.2–0.9	L	1.49	2400	∥ or ⊥	1.33	37.2	126	25
Fused quartz	0.2–4.5	L	5.96	12		1.46	8.05	1.56	1.35
SF-4	0.38–1.8	L	3.63	220	⊥	1.62	1.83	4.51	3.97
SF-59	0.46–2.5	L	3.20	1200	∥ or ⊥	1.95	39	19	12
SF-58		L	3.26	1200	∥ *or* ⊥	1.91	18.2	9	5.6
SF-57		L	3.41	500	∥	1.84	19.3	9	5.65
SF-6		L	3.51	500	∥ or ⊥	1.80	15.5	7	4.42
As₂S₃	0.6–11	L	2.6	170	∥	2.61	762	433	293
As₂S₅	0.5–10	L	2.22			2.2	278	256 (est.)	125

TABLE 11.3

Acousto-Optic Properties of Crystals for the Visible

Material	Transmission range (μm)	Acoustic mode & prop. dir.	ν (cm/s × 10⁵)	Γ (dB/ cm GHz²)	Opt. pol. dir.	n (0.633 μm)	M_1 (cm² s/g × 10⁻⁷)	M_2 (s³/g × 10⁻¹⁸)	M_3 (cm s²/g × 10⁻¹²)
LiNbO₃	0.04–4.5	L[100]	6.57	0.15		2.20	66.5	7.0	10.1
		S[001]	3.59	2.6	⊥	2.29	9.2	2.92	2.4
Al₂O₃	0.15–6.5	L[100]	11.0	0.2	∥	1.77	7.7	0.36	0.7
YAG	0.3–5.5	L[100]	8.60	0.25	⊥	1.83	0.98	0.073	0.114
		S[100]	5.03	1.1	∥ or ⊥	1.83	1.1	0.25	0.23
TiO₂	0.45–6	L[001]	10.3	0.55	⊥	2.58	44	1.52	4
SiO₂	0.12–4.5	L[001]	6.32	2.1	⊥	1.54	9.11	1.48	1.44
		L[100]	5.72	3.0	[001]	1.55	12.1	2.38	2.11
α-HIO₃	0.3–1.8	L[001]	2.44	10	[100]	1.99	103	86	42
PbMoO₄		L[001]	3.63	15	∥	2.62	108	36.3	29.8
TeO₂	0.35–5	L[001]	4.20	15	⊥	2.26	138	34.5	32.8
		S[110]	0.616	90	Circ [001]	2.26	68.0	793	110
Pb₂MoO₅	0.4–5	L a-axis	2.96	25	b-axis	2.183	242	127	82

of the materials that have been studied for AO applications. The acoustic attenuation constant in these tables is defined as

$$\Gamma = \frac{a}{f^2} \tag{11.93}$$

which supposes that the attenuation increases quadratically with frequency. This will be the case for good-quality single crystals, but not for polycrystalline, highly impure, or amorphous materials. For the latter, the constant given in the tables is a rough estimate, based on measurements at the higher frequencies. The light polarization direction is designated as parallel or perpendicular according to whether the light polarization is parallel or perpendicular to the acoustic beam direction. Table 11.2 lists some of the more

TABLE 11.4

Acousto-Optic Properties of Infrared Crystals

Material	Transmission range (μm)	Acoustic mode & prop. div.	ν (cm/s × 10⁵)	Γ (dB/ cm GHz²)	Opt. pol. dir.	λ (μm)	n	M_1 (cm² s/g × 10⁻⁷)	M_2 (s³/g × 10⁻¹⁸)	M_3 (cm s²/g × 10⁻¹²)
Ge	2–020	L[111]	5.50	30	∥	10.6	4.00	10,200	840	1850
		S[100]	3.51	9	∥ or ⊥	10.6	4.00	1430	290	400
Tl_3AsS_4	0.6–12	L[001]	2.5	29	∥	1.15	2.63	620	510	290
GaAs	1–11	L[110]	5.15	30	∥	1.15	3.37	925	104	179
		S[100]	3.32		∥ or ⊥	1.15	3.37	155	46	49
Ag_3AsS_3	0.6–13.5	L[001]	2.65	800	∥	.633	2.98	816	390	308
Tl_3AsSe_3	1.25–18	L[100]	2.15	314	⊥	3.39	3.15	654	445	303
Tl_3PSe_4	0.85–9	L[100]	2.0	150	∥	1.15	2.9	2866	2069	1288
$TlGaSe_2$	0.6–20	L[001]	2.67	240	∥	.633	2.9	430	393	161
CdS	0.5–11	L[100]	4.17	90	∥	.633	2.44	52	12	12
ZnTe	0.55–20	L[110]	3.37	130	∥	1.15	2.77	75	18	19
GaP	0.6–10	L[110]	6.32	6.0		.633	3.31	75	30	71
ZnS	0.4–12	L[001]	5.82	27	∥	.633	2.35	27	3.4	4.7
		S[001]	2.63	130		.633	2.35	14	8.4	5.2
Te	5–20	L[100]	2.2	60	∥	10.6	4.8	10,200	4400	4640

important amorphous materials, which may be useful if large sizes are desired or very low cost is required, but none of which can be used at frequencies much above 30 MHz. Table 11.3 lists the most important class of materials, crystals that are transparent throughout the visible with very low acoustic losses. Table 11.4 lists high-efficiency crystal materials that are transparent in the IR and have reasonably low acoustic losses.

An overall summary of a few outstanding (in one or another respect) selected AO materials presented in these tables is shown in Figure 11.19. Using figure of merit and acoustic attenuation as criteria of quality, it is clear that a trade-off between these two parameters exists, and that the selection of the optimum material will be determined by the system requirements.

11.6 ACOUSTIC TRANSDUCER DESIGN

11.6.1 Transducer Characteristics

The second key component of the AO scanner after the optical medium, is the transducer structure, which includes the piezoelectric layer, bonding films, backing layers, and matching network. Recent advances in this area have made available a number of new piezoelectric materials of very high electromechanical conversion efficiency, and bonding techniques that permit this high conversion efficiency to be maintained over a large bandwidth. Furthermore, the design of high-performance transducer structures utilizing this new technology has been facilitated by new analytical tools[37,38] that lend themselves to computer programs for optimizing this performance.

The most elementary configuration of a thickness-driven transducer structure is shown in Figure 11.20. It consists of the piezoelectric layer, thin film or plate, excited by metallic electrodes on both faces, and a bonding layer to acoustically couple the piezoelectric to the
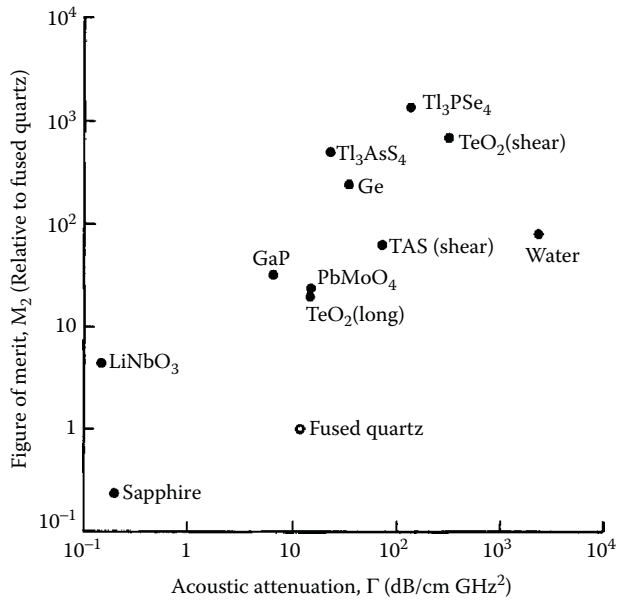
FIGURE 11.19
Figure of merit versus acoustic attenuation.
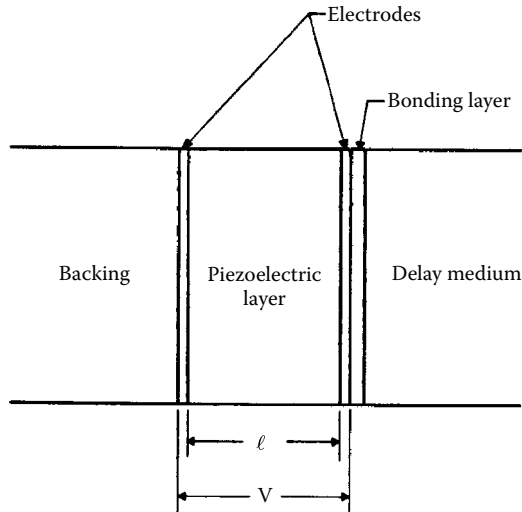
FIGURE 11.20
Transducer structure.

delay medium, or optical crystal. The backing is applied to mechanically load the transducer for bandwidth adjustment but may simply be left as air. The thickness of the transducer is about half an acoustic wavelength at the resonant frequency, and the thickness of the bonding layer is chosen to allow high, broadband acoustic transmission. The most efficient operation of the transducer is obtained when the mechanical impedances of all

the layers are equal. The mechanical impedance is

$$Z = ru \tag{11.94}$$

and in general there is not sufficient choice of available materials to satisfy this condition. When the impedances are unequal, reflection occurs at the interfaces, reducing the efficiency of energy transfer. The reflection and transmission coefficients at the boundary between two media of impedances Z_1 and Z_2 are

$$R = \frac{(Z_1 - Z_2)^2}{(Z_1 + Z_2)^2} \tag{11.95}$$

$$T = \frac{4Z_1 Z_2}{(Z_1 + Z_2)^2} \tag{11.96}$$

The electromechanical analysis is generally carried out in terms of an equivalent circuit model, first proposed by Mason.[39] Several variations of the equivalent circuit have since been developed, but the one due to Mason is shown in Figure 11.21. The fundamental constants of the transducer are permittivity s, acoustic velocity v, and electromechanical coupling factor k. The other parameters are transducer thickness l and area S. With these parameters, the circuit components shown in Figure 11.21 are

$$C_0 = \in \frac{S}{l} \tag{11.97}$$

$$f = k \left(\frac{1}{p} w_0 c_0 Z_0 \right) \frac{1}{2} \tag{11.98}$$

$$Z_A = j Z_0 \tan \frac{g}{2} \tag{11.99}$$

$$Z_A = j Z_0 \tan \frac{\gamma}{2} \qquad \gamma = \pi \frac{f}{f_0}$$

$$Z_B = -j \frac{Z_0}{\sin \gamma} \qquad \theta = \kappa \left(\frac{1}{\pi} \omega_0 C_0 Z_0 \right)^{1/2}$$
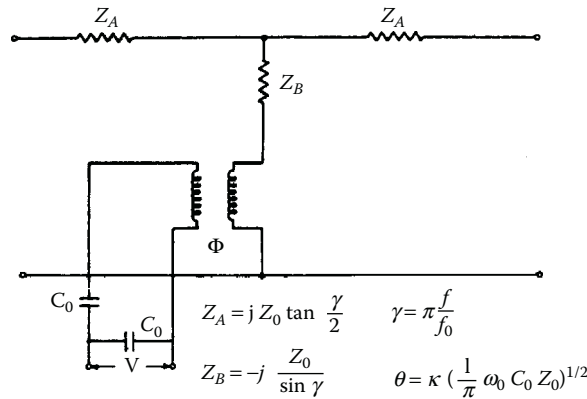
FIGURE 11.21
Equivalent circuit of Mason.

$$Z_B = -j\frac{Z_0}{\sin g} \tag{11.100}$$

where

$$w_0 = \frac{pu}{l} \tag{11.101}$$

$$g = p\frac{w}{w_0} \tag{11.102}$$

$$Z_0 = Sru \tag{11.103}$$

This equivalent circuit was used by Sittig[37] and Meitzler and Sittig[38] to analyze the propagation characteristics of acoustic energy between a piezoelectric and a delay medium. This was done in terms of a two-port electromechanical network, described by the chain matrix

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix} = \Pi_m \begin{pmatrix} A_m & B_m \\ C_m & D_m \end{pmatrix} \tag{11.104}$$

If the equivalent circuit of Figure 11.21 is terminated at the input with a voltage source V_s and impedance Z_s, and at the output with a transmission medium of mechanical impedance Z_t, output voltage V_l, and load impedance Z_l, as shown in Figure 11.22, then the insertion loss is

$$L = 20\log\frac{V_s}{V_l} + 20\log\left|\frac{Z_s + Z_t}{Z_l}\right|\text{dB} \tag{11.105}$$

The impedances Z_s and Z_t are assumed to be purely resistive and

$$\frac{V_l}{V_s} = \frac{2Z_l Z_t}{\left\{AZ_t + B + Z_s(CZ_t + D)\right\}\left\{AZ_t + B + Z_t(CZ_t + D)\right\}} \tag{11.106}$$
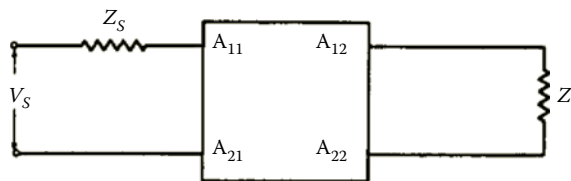
FIGURE 11.22
Terminated two-port transducer.

The two-port transfer matrix was obtained by Sittig,[40] with the result

$$
A = \frac{1}{fH} \begin{vmatrix} A' & B' \\ C' & D' \end{vmatrix} \begin{vmatrix} \cos g + jz_b \sin g & Z_0(Z_b \cos g + z \sin g) \\ \dfrac{j \sin g}{z_0} & 2(\cos g - 1) + jz_b \sin g \end{vmatrix}
\tag{11.107}
$$

where

$$
z_b = \frac{z_b}{z_0}, \quad H = \cos g - 1 + jZ_b \sin g
\tag{11.108}
$$

and

$$
A' = 1, \quad B' = j\frac{f^2}{wC_0}, \quad C' = jwC_0, \quad D' = 0
\tag{11.109}
$$

The impedance Z_b represents the mechanical impedance of layers placed on the back surface of the transducer for loading, $Z_b = S\rho_b v_b$. In case the transducer is simply air-backed, Z_b 0. Electrical matching may be done at the input network by adding inductors either in parallel or in series in order to be electrically resonant with the transducer capacity C_0 at midband, $\omega = \omega_0$. If no inductances are added, the minimum loss condition is achieved for

$$
R_s = \frac{1}{w_0 C_0}
\tag{11.110}
$$

where R_s is the source resistance. The inductance, if added, is chosen so that

$$
L = \frac{1}{w_0^2 C_0}
\tag{11.111}
$$

A result of the matrix analysis shows that when piezoelectric materials with large values of the coupling constant κ are used, it is possible to achieve large fractional bandwidths without the necessity for electrical matching networks. As an example of the results obtained with this formalism, several plots of the frequency dependence of transducer loss for different values of the coupling constant are shown in Figure 11.23.

11.6.2 Transducer Materials

The piezoelectric material itself is perhaps the single most important factor governing the efficiency with which electrical energy can be converted to acoustic energy, this through the electromechanical coupling factor κ. The coupling efficiency is equal to κ. Prior to the discovery of lithium niobate, quartz was the most commonly used high-frequency transducer material, although its coupling factor, even for the most efficient crystal orientations, is rather small. The very-high-efficiency transducers were introduced with the discovery of various new ferroelectrics, such as lithium niobate, lithium tantalate, and the ceramic
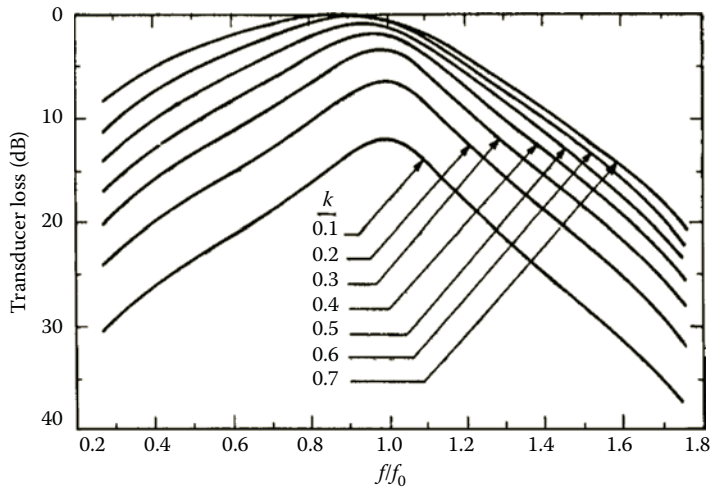
FIGURE 11.23
Transducer loss for various values of κ; $z_{0t} = 0.4$ and $R_s = (\omega_0 C_0)^{-1}$.

PZT materials, lead-titanate-zirconate. While the PZT transducers have among the highest values of κ, up to 0.7, they are not suitable for high-frequency applications since they cannot be polished to very thin plates. The most suitable piezoelectric transducer materials for high-frequency applications and their important properties are listed in Table 11.5, which is based on a compilation of Meitzler.[41] In order to produce transducers in the high-frequency range, say larger than 100 MHz, the piezoelectric crystal must be very thin (<20–30 μm). There are three well-established techniques for fabricating such thin transducers. In the first method, the piezoelectric plate is lapped to the desired thickness by the usual optical shop methods and then bonded to the delay medium. This method becomes impossibly difficult for transducers of even small area as their frequency increases, because such thin plates cannot be manipulated. A much more convenient technique is to bond the piezoelectric plates with a convenient thickness, say several tenths of a millimeter, to the delay medium and then lap the plate to the final thickness. In both methods, one electrode is first deposited on the delay medium, and in the case of thinning the piezoelectric after bonding, the second electrode and back layers are deposited as the final step. Care is required in lapping the bonded transducer so that the base electrode is not damaged by the polishing compound. If a chemically active compound, such as Cyton, is used, the delay medium as well as the electrode may be attacked and must be protected by some appropriate coating, such as photoresist. The final electrical connection to the top electrode must be made in some fashion that does not mass-load the transducer and distort its bandpass characteristics, or be so small as to cause hot spots from high current densities. The usual method is to bond thin gold wire or ribbon onto electrode tabs, as is done for electronic circuit chips. The most successful method for fabricating very-high-frequency transducers for longitudinal wave generation is by deposition of thin films of piezoelectric materials by methods that yield a desired crystallographic orientation.[42,43] The materials used are CdS and ZnO, whose properties are shown in Table 11.5.

Such piezoelectric thin films generally cannot be grown with values of κ as high as that of the bulk material, but in the best circumstances κ may approach 90%. Thin-film transducers with band center frequencies up to 5 GHz can be prepared by these techniques.

TABLE 11.5

Properties of Transducer Materials

Material	Density	Mode	Orientation	K	ε_{rel}	υ (cm/s)	Z (g/s cm²)
$LiNbO_3$	4.64	L	36° Y	0.49	38.6	7.4×10^5	34.3×10^5
		S	163° Y	0.62	42.9	4.56×10^5	21.2×10^5
		S	X	0.68	44.3	4.8×10^5	22.3×10^5
$LiTaO_3$	7.45	L	47° Y	0.29	42.7	7.4×10^5	55.2×10^5
		S	X	0.44	42.6	4.2×10^5	31.4×10^5
$LiIO_3$	4.5	L	Z	0.51	6	2.5×10^5	11.3×10^5
		S	Y	0.6	8	2.5×10^5	11.3×10^5
$Ba_2NaNb_5Oi_5$	5.41	L	Z	0.57	32	6.2×10^5	33.3×10^5
		S	Y	0.25	227	3.7×10^5	19.8×10^5
$LiGeO_2$	4.19	L	Z	0.30	8.5	6.3×10^5	26.2×10^5
$LiGeO_3$	3.50	L	Z	0.31	12.1	6.5×10^5	22.8×10^5
αSiO_2	2.65	L	X	0.098	4.58	5.7×10^5	15.2×10^5
		S	Y	0.137	4.58	3.8×10^5	10.2×10^5
ZnO	5.68	L	Z	0.27	8.8	6.4×10^5	36.2×10^5
		S	39° Y	0.35	8.6	3.2×10^5	18.4×10^5
		S	Y	0.31	8.3	2.9×10^5	16.4×10^5
CdS	4.82	L	Z	0.15	9.5	4.5×10^5	21.7×10^5
		S	40° Y	0.21	9.3	2.1×10^5	10.1×10^5
$Bi_{12}GeO_{20}$	9.22	L	(111)	0.19	38.6	3.3×10^5	30.4×10^5
		S	(110)	0.32	38.6	1.8×10^5	16.2×10^5
AIN	3.26	L	Z	0.20	8.5	10.4×10^5	34.0×10^5

A problem that arises with large-area transducers, or even with small-area transducers at very high frequencies, is that of matching the electrical impedance to the source impedance. It is especially true for the ferroelectric, piezoelectric materials of very high dielectric constant that the impedance of the transducer may be so low that it becomes difficult to efficiently couple electrical power from the source to the transducer. This problem can be largely overcome by dividing the transducer into a series connected mosaic, as reported by Weinert and de Klerk.[44] A schematic representation of such a mosaic transducer is shown in Figure 11.24. If a transducer of given area is divided into N elements, which are connected in series, the capacity of the transducer will be reduced by a factor of N^2. As an example, a 1-GHz lithium niobate transducer of 0.25 cm² (0.4 in²) area would represent a capacitive impedance of only 0.038 Ω; if this area were divided into a 16-element mosaic, the impedance would be increased to 10 Ω. A 40-element thin-film transducer is shown in Figure 11.25. The same considerations will apply at lower frequencies for transducers with large areas, about 1 cm² or more. Because most ferroelectric transducer materials, such as the PZTs or lithium niobate, have high dielectric constants, the large areas will lead to very large capacitance values for frequencies far below 100 MHz. Thus, large-area transducers are usually divided into multiple elements, which are then wired in series to obtain the desired 50-Ω impedance to match to the RF driver. A large-area transducer that has been so wired is shown in Figure 11.26.

11.6.3 Array Transducers

One of the serious limitations of normal (i.e., isotropic) Bragg AOD is that imposed by the bandwidth as limited by the Bragg interaction. The most straightforward method of
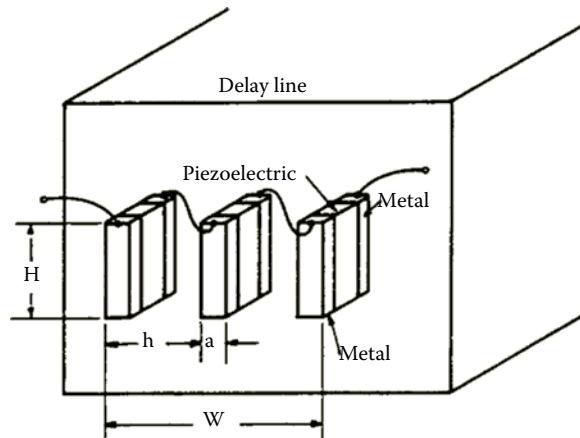
FIGURE 11.24
Schematic of mosaic transducer.
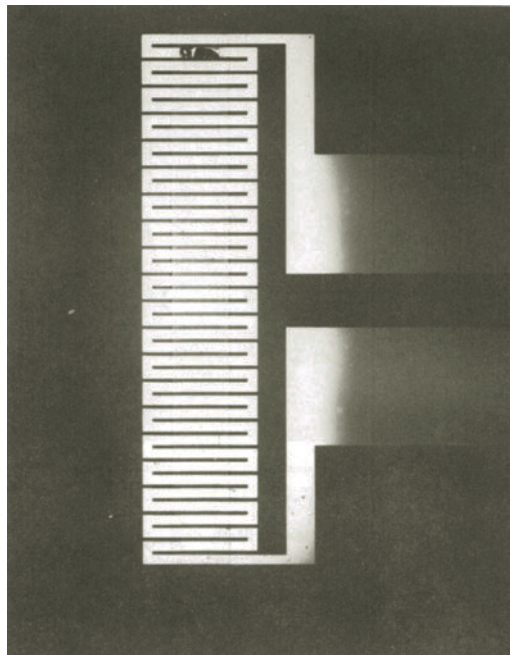
FIGURE 11.25
Forty-element thin-film mosaic transducer array.

enlarging the interaction bandwidth is simply to shorten the interaction length in order to increase the acoustic beam diffraction spread. This is generally not a very desirable method to increase bandwidth for systems in which the light to the Bragg cells is collimated because it wastes acoustic power; only those momentum components of the acoustic beam that can be phase matched to incident and diffracted light momentum components are useful. Furthermore, as the interaction length shrinks, the transducer becomes increasingly narrow, with a corresponding increase in power density. This increase in power
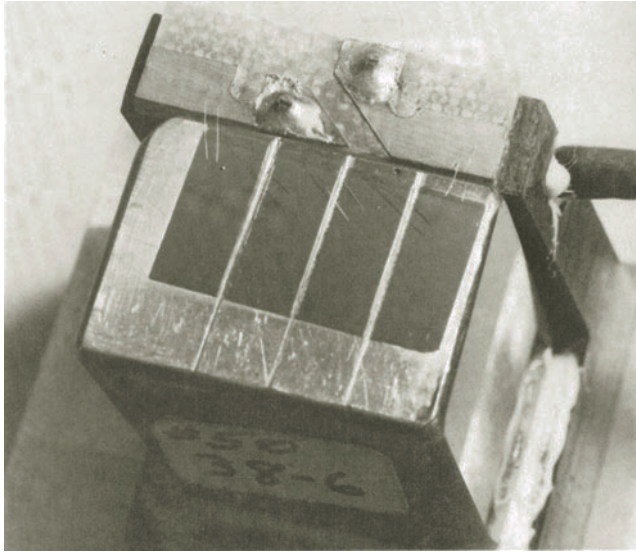
FIGURE 11.26
Four-element, series-connected lithium niobate transducer metal-bonded to Bragg cell.

density may produce heating at the transducer, which can cause thermal distortion in the deflector due to gradients in the acoustic velocity and refractive index.

An ideal solution to this difficulty would be one in which the acoustic beam changes in direction as the frequency is changed, so that for every frequency the Bragg angle is perfectly matched. The first approximation to such acoustic beam steering was carried out by Korpel[45] for a television display system. This transducer consisted of a stepped array, as shown in Figure 11.27. The height of each step is one-half an acoustic wavelength at the band center $\Lambda_0/2$, and the spacing s between elements is chosen so as to optimize the tracking of the Bragg angle. Each element is driven π rad out of phase with respect to the adjacent elements, and the net effect of such a transducer is to generate an acoustic wave with corrugated wavefronts, which are tilted at an angle with respect to the transducer surfaces when the frequency differs from the band center frequency f_0. For this transducer configuration, the acoustic beam steers with frequency but matches the Bragg angle only imperfectly.

To understand the steering properties of such an acoustic array, which was analyzed in detail by Coquin et al.,[45] consider the somewhat simpler arrangement shown in Figure 11.28, in which each transducer element is driven Ψ rad out of phase with respect to the next one, and Ψ may be electrically varied. This causes the effective wavefront to be tilted by an angle θ_e with respect to the piecewise wavefronts radiating from the individual elements. If θ_e is small, it can be approximated by

$$q_e \approx \tan q_e = \frac{\Psi}{2p}\frac{\Lambda}{s} = \frac{\Psi}{Ks} \tag{11.112}$$

If the incident light beam makes an angle θ_0 with the plane of the transducer and if the Bragg angle is $\theta_B = K/2k$, then the angular error from perfect matching is

$$\Delta q = \left(q_0 - q_e\right) - q_B = \left(q_0 - \frac{\Psi}{Ks}\right) - \frac{K}{2k} \tag{11.113}$$
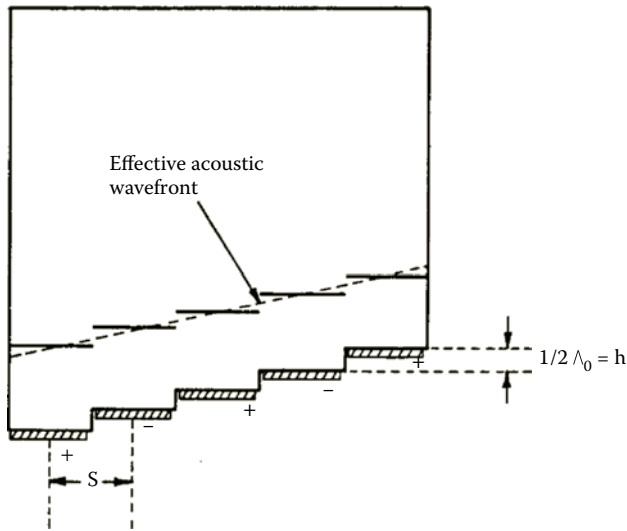
FIGURE 11.27
Stepped transducer array.
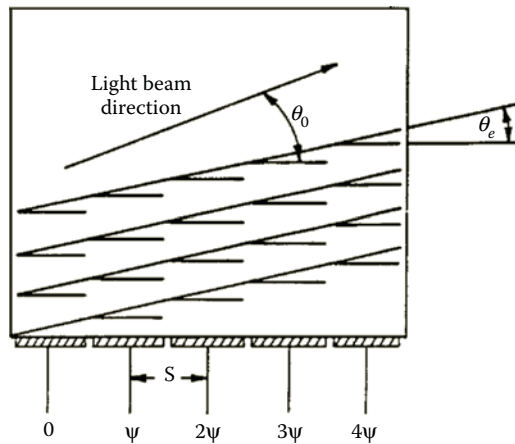
FIGURE 11.28
Steering of an acoustic beam by a phased array transducer.

The condition for perfect beam steering is that $\Delta\theta = 0$ for all values of K; setting $\Delta\theta = 0$, the required phase for perfect beam steering is

$$\Psi_P = q_0 Ks - \frac{K^2}{2k}s \tag{11.114}$$

from which it can be seen that the phase must be a quadratic function of the acoustic frequency.

Most of the work done on acoustic beam steering has involved making various approximations to this condition. One such approximation is obtained by making Ψ a linear function of frequency, with $\Psi = 0$ at f_0, the midband frequency. This was accomplished in the

step transducer method of Figure 11.27, as described in Ref. 46. For this case, the angle that the effective wavefronts make with respect to the transducer plane is

$$q_e \approx \frac{p}{Ks} - \frac{h}{s} \tag{11.115}$$

where h is the step height, and there is 180° phase shift between adjacent elements. The resulting beam steering error is

$$\Delta q = \left(q_0 - \frac{K}{2k} \right) + \left(\frac{h}{s} - \frac{p}{Ks} \right) \tag{11.116}$$

which can be made zero at the midband frequency f_0 by choosing

$$h = \frac{1}{2} \Lambda_0$$
$$s = \frac{\Lambda_0^2}{l} \tag{11.117}$$

and

$$q_0 = \frac{1}{2} \frac{l}{\Lambda_0}$$

A further improvement can be achieved by noting from Equation 11.115 that θ_e varies as $1/f$, whereas perfect beam steering should lead to a linear variation of θ_e with f. Therefore, the constants h, s, and θ_e may be chosen to agree with the perfect beam steering case at two frequencies, rather than only one, as shown in Figure 11.29. This first-order beam steering can yield substantial improvements in performance for systems requiring less than one-octave bandwidth,[47] but bandwidths larger than this require a better approximation to the quadratic dependence of the phase on the acoustic frequency. The next higher approximation to perfect beam steering was carried out by Coquin et al.[45] for a ten-element array, as shown in Figure 11.30. If the phase applied to each transducer corresponds to that for perfect steering, $\Psi_1 = l\Psi_p$, and the element spacing is $s = \Lambda_0^2/l$, the bandwidth extends from 0 to about 1.6f_0, the high-frequency drop-off being determined by the finite element spacing. Coquin pointed out that the deflector performance is very tolerant of errors in the individual phases; for example, if the phase applied to each transducer is within 45° of the perfect beam steering phase, there is a loss of only 0.8 dB in diffracted light intensity. If the phase error is increased to 90°, the loss increases to 3 dB. Thus, for deflectors, in which this degree of ripple is permissible, the transducer array may be driven by logic circuitry that sets digital phase shifters. This requires prior knowledge of the input frequency, or analog phase shifters, which accomplish the same function without the need for logic circuits.

An entirely different approach to broadband Bragg AO interaction matching is the use of the tilted transducer array, first reported by Eschler.[48] A tilted transducer array consists of two or more transducers electrically connected in parallel and tilted in angle with respect to each other, as illustrated in Figure 11.31. Each transducer element in the array is designed to cover some fraction of the entire bandwidth, and its angle with respect to the incident light direction is chosen to match the Bragg angle at the center of its subband.
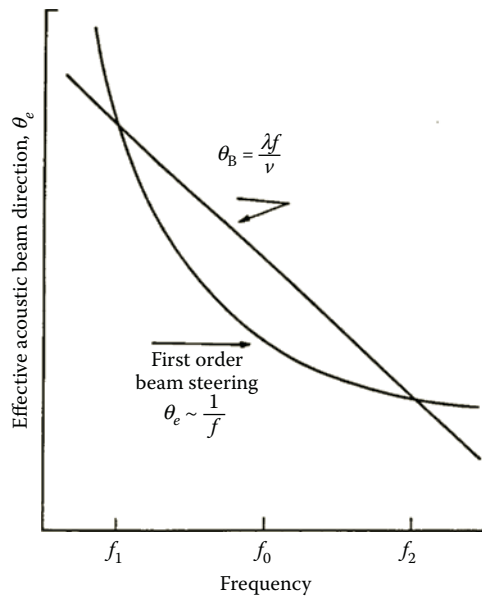
FIGURE 11.29
First-order beam steering with exact match at two frequencies.
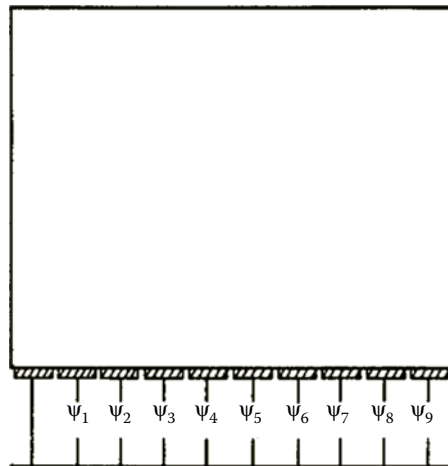
FIGURE 11.30
Ten-element phased array transducer, in which $\Psi = 0, 90, 180$, or $270°$, leading to diffracted intensity less than 0.8 dB lower than for perfect beam steering.

For frequencies near the midband of any of the transducer elements, the incident light will interact strongly only with the sound wave emanating from that element; interaction will be weak from the other elements both because the angle of incidence will be mismatched and the frequency will be far from the resonance frequency of those elements.

On the other hand, for frequencies that are midway between the resonance frequencies of adjacent elements, that is, $(f_{01} + f_{02})/2$ or $(f_{02} + f_{03})/2$, the contributions to the acoustic fields from both elements are about equal and the effective wavefront direction lies
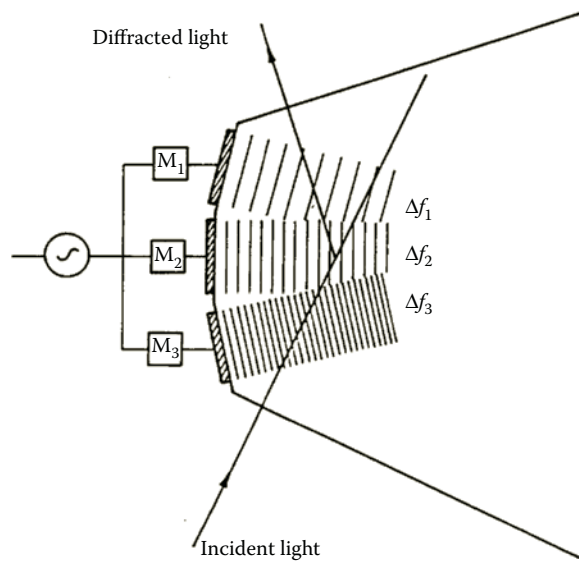
FIGURE 11.31
Tilted transducer array, in which each element is optimized for part of the entire frequency band.
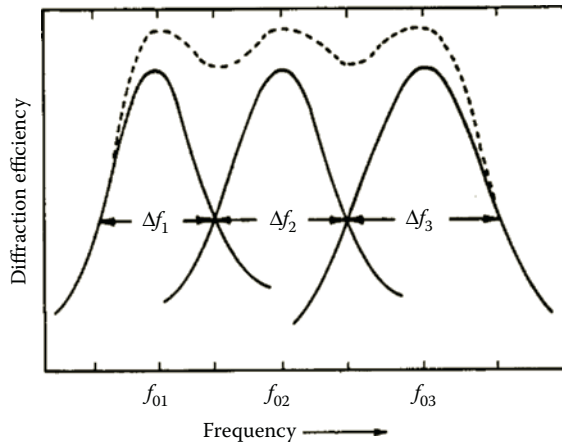
FIGURE 11.32
Diffraction efficiency of a three-element tilted transducer array. Solid curves represent efficiency of individual elements, and dotted curve represents efficiency of entire array.

midway between those of the components. Thus, the array behaves very much as if the acoustic wave were steering with frequency, although this is not true in a strict sense. The diffraction efficiency of the tilted transducer array is shown in Figure 11.32, in which the solid curves represent the efficiency of the individual elements and the dotted curve represents the overall efficiency. There will typically be about 1 dB of ripple across the full band, which is acceptable for most applications.

There are two additional advantages to the tilted array transducer. First, it is obviously relatively simpler to design a larger overall acoustic bandwidth, since each element of the tilted array need be only about one-third of the total bandwidth. Second, tilted-array

transducers can generally be operated more deeply in the Bragg mode, as the combined acoustic wavefronts from adjacent elements are twice the length of that from a single element. If the second-order diffracted light is sufficiently low, then operation over a frequency range larger than one octave is possible. The elements of the array can be connected in parallel since they will tend to behave as bandpass filters, the power being directed to the element with the closest frequency range. In practice, it is generally necessary to provide impedance-matching networks, as indicated in Figure 11.31, because of the low reactance obtained with a parallel network.

11.7 ACOUSTO-OPTIC DEVICE FABRICATION

11.7.1 Cell Fabrication

The AO material is in most cases a crystal or an optical-grade glass. For crystalline materials, specific acoustic and optic propagation axes are typically required and the crystallographic orientation of the material must therefore be identified and marked. A typical cell configuration is shown in Figure 11.33. The optical window surfaces are prepared on the faces perpendicular to the optical axis.

These surfaces are typically polished to a high-quality window surface with a flatness of $\lambda/20$ or better. Because the cells may be relatively thick (up to several centimeters) the homogeneity of the optical material may be a significant factor in the overall wavefront distortion. Therefore, the wavefront distortion may need to be specified for the cell in transmission.

Optical scatter is often a critical parameter for AO devices used in scanning systems. A portion of scatter from the zero-order beam will fall within the aperture of the deflected beam. This scattered light is unmodulated and may limit the extinction ratio in AOM applications. Scatter may occur at the surfaces due to contamination or imperfections or within the material due to defects or inhomogeneities.

Unlike most lenses and mirrors where the clear aperture is in the middle of the optic, the desired clear aperture of an AO device usually starts at the edge of the transducer face. For AOMs, the beams may be centered less than 0.5 mm from the edge of the cell. Therefore,
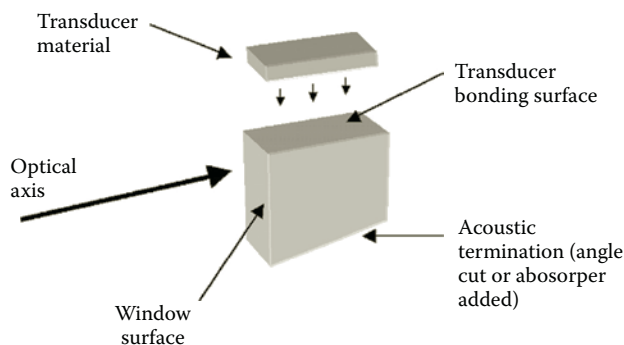
FIGURE 11.33
Bonding of acoustic transducer and acousto-optic medium.

special care must be taken to achieve the surface flatness, polish, and antireflection coating required across the entire clear aperture.

In particular with AOMs, the laser beam is focused to a small spot at the AOM to achieve the desired rise time and can lead to high optical intensity. Therefore, antireflection coatings often must be specified with high damage thresholds.

The acoustic transducer bonding surface is nominally parallel to the optic axis. This surface is prepared for bonding typically by polishing to an optical-quality surface. In most applications, the face of the cell opposite of the acoustic transducer surface is cut or ground at an off angle so that acoustic waves will not reflect directly back to the optical aperture or acoustic transducer. This is to avoid modulation or intermodulation from acoustic echo. Other techniques employed to reduce back reflection are to bond an acoustic absorbing material to back surface or grinding a rough finish on the back surface to diffuse the back reflection.

11.7.2 Transducer Bonding

For transducers in the frequency range for which crystal plates are bonded to the delay medium, the bonding procedure is probably the most critical and most difficult step in fabricating the structure. The bonding layer can drastically modify the transmission of acoustic energy between the piezoelectric and the delay media; this is because the bond layer must provide molecular contact between the two surfaces, which will otherwise result in incomplete transfer, and because the mechanical impedance of the bond layer may produce a large acoustic mismatch with low transmission. In addition to these considerations, if the bond material is acoustically lossy, further decrease in transmission will result.

Because of the special properties required, there is only a very limited number of known bonding materials available. For temporary attachments, a commonly used agent is "salol," phenyl salicylate. It is easily applied as a liquid, which is crystallized by addition of a small seed. It is reliquified by gentle heating, and therefore is useful for various test measurements, but does not yield wide bandwidth or efficient coupling. A more satisfactory bond is made with epoxy resin, mixed to a very low viscosity, which may be compressed to a layer less than 1 µm thick before setting. Such thin layers require a high degree of cleanliness to avoid inclusion of any dust particles. Because of the low impedance of epoxy compared with such transducer materials as lithium niobate, thicker bonding layers would cause serious impedance mismatch problems around 100 MHz, where this technique has been successfully used.

Good results can also be obtained with a low-viscosity, ultraviolet, light-cured cement. For frequencies higher than about 100 MHz, other techniques, capable of yielding still thinner bond layers, which must be kept to a small fraction of an acoustic wavelength, must be used. Vacuum-deposited metallic layers are well suited for this purpose, because their thickness can be very accurately controlled down to the smallest dimensions, and impedances much closer to those of commonly used piezoelectric materials are available. Very good results were first obtained with indium bonds,[49] which are deposited to a thickness of several thousand angstroms on both surfaces, and without removal from the vacuum systems are mated under a pressure of about 100 psi. This technique yields a cold-welded bond, which has excellent mechanical properties with large acoustic bandwidth, if properly designed, and low insertion loss at frequencies of hundreds of megahertz.

The greatest fabrication difficulty is due to the necessity of maintaining the deposited films under vacuum to prevent oxidation. This requires a vacuum system with rather elaborate fixtures to bring the two surfaces together after film deposition and to apply the

hydraulic pressure. The inside of a vacuum system in which this procedure is carried out is shown in Figure 11.34. The substrates are held on either side of the evaporation filament sources during film deposition and are quickly brought into contact before contamination can occur. Using this technique, compression bonds of indium, tin, aluminum, gold, and silver are routinely made. It is essential for these procedures to be carried out in a dust-free atmosphere in order to avoid contaminating the interface with particulates. Even the smallest particles will prevent good acoustic contact between the transducer and the cell, so that fabrication must be carried out in a clean-room facility. A typical AO device clean room is shown in Figure 11.35.

In a modification of the indium compression bond,[50] which allows the freshly deposited indium surfaces to be removed from the vacuum system for handling, the work is then placed in an oven under a pressure of several hundred psi, raised in temperature to slightly below the melting point of indium (156 °C), and slowly cooled. This procedure forms a molecular bond in spite of the oxidation that may occur, and gives results similar to the vacuum bond. The principal drawback is that upon cooling, differential thermal expansion coefficients between the delay line material and the transducer material may set up unacceptable strains in the optical path. For some systems, this may not be a problem; for example, quartz or even lithium niobate transducers on fused or crystal quartz delay lines can be routinely made by this method. On the other hand, such crystals as tellurium dioxide require a great deal of care in handling, since they are extremely sensitive to thermal shock and strains. Differential contraction between the crystal and transducer for a bond made in this fashion may easily be severe enough to fracture the crystal. Therefore, its applicability will depend upon the materials and sizes involved and upon the degree of freedom from residual strain required.
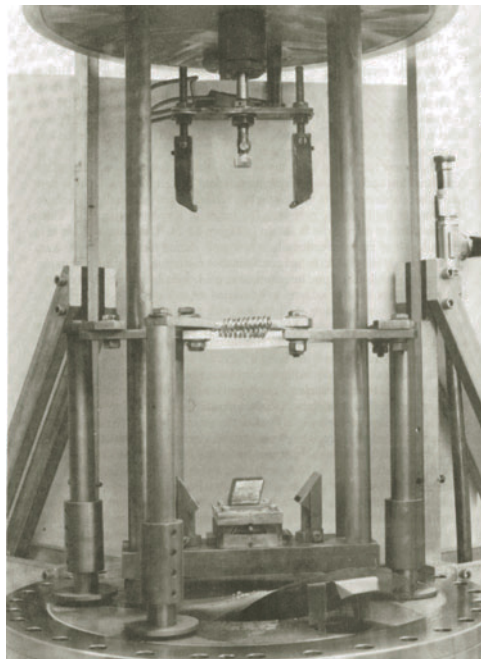
FIGURE 11.34
Vacuum compression bonding system. Metal films are deposited on transducer and delay line surfaces, which are then brought into contact.

FIGURE 11.35
A clean room for acousto-optic device fabrication.

For frequencies approaching 1 GHz, the attenuation of indium layers may become excessive, and better results can be achieved with metals with lower acoustic loss constants. Among such metals are gold, silver, and aluminum.

Although these are made by the vacuum compression method, they generally require higher pressure. Still another method that has been used with these, as well as indium, is ultrasonic welding.[51] The chief advantage to be gained is that the procedure is carried out in normal atmosphere, since the ultrasonic energy breaks up the oxidation layer that forms on the surface. Some heating occurs as a result, but the temperature remains well below that required in the indium thermocompression method, with much lower residual strains. The technique requires the simultaneous application of pressures up to 3000 psi; this may be excessive for easily fractured or deformed materials or where odd-shaped samples are involved. A summary of the important properties of a few bonding materials, also used for electrodes and intermediate impedance-matching layers, is given in Table 11.6.

At lower frequencies, the effects of thin electrode and bonding layers on the performance of the transducer may be entirely negligible, but near 100 MHz, they become increasingly large, and even for layers less than 1 μm thick the effect may not be negligible if the impedance mismatch to the rest of the structure is large. The effects of the electrode layer can be determined by setting $Z_b = 0$ in Equation 11.107, and the entire effect of the back layers will be due to the impedance of the electrode z_{bl} of thickness t_{bl}, so the normalized impedance

$$z_b = jz_{bl} \tan(t_{bl} g) = j \tan d \qquad (11.118)$$

and the matrix of Equation 11.107 becomes more complex.

The effect of the bond layer and front electrode is even more complex, but an interesting illustrative example of varying the bond layer thickness is shown in Figure 11.36. For this example, the normalized impedance of the bond layer is taken to be rather low, $z = 0.1$, and it can be seen that even for a fairly small thickness, the effect on the transducer loss is

TABLE 11.6

Acoustic Properties of Bond Layer Materials

Material	Longitudinal waves			Shear waves		
	Velocity (cm/s)	Impedance (g/s cm²)	Attenuation (dB/µm @ 1 GHz)	Velocity (cm/s)	Impedance (g/s cm²)	Attenuation (dB/µm @ 1 GHz)
Epoxy	2.6×10^5	2.86×10^5	Very large	1.22×10^5	1.34×10^5	Very large
Indium	2.25×10^5	16.4×10^5	8	0.19×10^5	6.4×10^5	16
Gold	3.24×10^5	62.5×10^5	0.02	1.2×10^5	23.2	0.1
Silver	5.65×10^5	38×10^5	0.025	1.61×10^5	16.7×10^5	
Aluminum	6.42×10^5	17.3×10^5	0.02	3.04×10^5	8.2×10^5	
Copper	5.01×10^5	40.6×10^5		2.11×10^5	18.3×10^5	
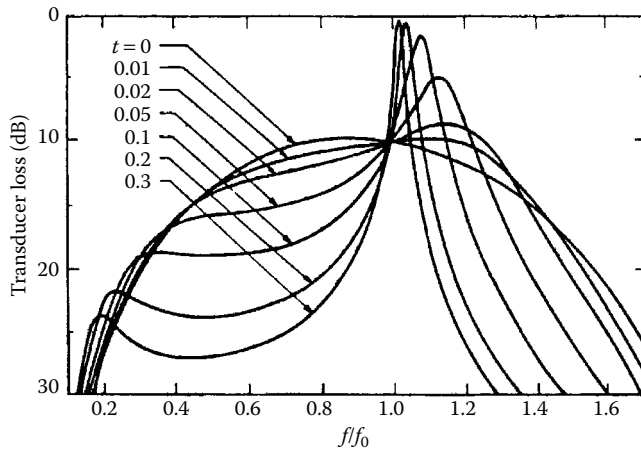
FIGURE 11.36

Transducer loss for various values of normalized transducer thickness t and intermediate layer normalized thickness 0.1. $R_s = (\omega_0 C_0)^{-1}$, $z_{0t} = 1$, $k = 0.2$.

quite marked. Such a low value of impedance would correspond to the nonmetallic bond materials, but for the metallic bond materials the impedance mismatch would not be as severe, and the curve of transducer loss would be correspondingly less influenced. This influence of intermediate layers on the shape of the transducer loss curve can be used to determine the bandpass characteristics of the transducer structure. Such impedance transformers can be used, for example, to make the response symmetric about the band center f_0 by making the intermediate layer thickness one-quarter wavelength at f_0. By choosing other values for the thickness, the bandwidth can be enlarged, ripples smoothed, or various distortions introduced. In general, however, any such objectives are achieved at the expense of increased transducer loss.

11.7.3 Packaging

AO device packaging must take into account needs for optical mounting and electrical connectivity and in many cases, thermal path management as well. Often it is desirable to attach the optical cell to a metal mount suitable for mechanical attachment in a larger

optical system. This is achieved by bonding one or more of the surfaces to the mount with adhesive. Low shear strength adhesives are needed in many cases to alleviate temperature-induced strain between the metal and the optical material. This is because most metals have a thermal expansion coefficient significantly greater than that of glass or other optical crystals. For example, the linear expansion coefficient of aluminum is approximately 23 parts per million per °C compared to only 0.5 parts per million per °C for fused silica. If the cell is bonded with a thin layer of high strength epoxy, difference in expansion over temperature may cause significant strain birefringence in the optical cell.

For devices that use a watt or more of RF power, thermal management becomes important. Most of the power input to the device will end up as heat either from electrical resistive losses or acoustic attenuation loss. This heat must be transferred from the modulator efficiently enough to keep the operating temperature within desired limits. A significant fraction of the input power may be lost at the transducer surface of the AOM, including resistive losses from bond wires and electrodes, and acoustic losses in the transducer and bond layers. Localized heating from the transducer will form temperature gradients in the optical cell that can cause optical distortion due to the change in index with temperature. The effect may be reduced by attaching a heat sink to the back side of the transducer or by choosing an optical aperture further from the transducer. The remaining power is converted into acoustic power that dissipates inside the cell. Because most optic materials are poor thermal conductors, the cell may need to be intimately contacted with a good heat sink path on as much surface area as possible to keep temperature rise to a minimum.

For modulators operating with frequencies of hundreds of MHz or more, the transducer electrode is typically deposited as a thin film (less than 1 µm thickness) by vacuum deposition. Electrical connection to the electrodes is made by bond wiring. Direct soldering to the electrodes may cause damage to the electrode and cause excess loading on the acoustic transducer. A circuit card may be used to bring the input RF from an external connector up to the acoustic transducer. The feed circuit typically includes a passive matching circuit to optimize the impedance matching, nominally 50 Ω.

11.8 APPLICATIONS OF ACOUSTO-OPTIC SCANNERS

11.8.1 Multichannel Acousto-Optic Modulator for Polygonal Scanner

AOM or AOD can be fabricated with many independent channels on a single monolithic device. This approach allows multiple parallel beams to be modulated with a single device. This approach is used in the Etec Systems ALTA 3000 mask writing machine, which writes semiconductor photomasks using a rasterized simultaneous 32-beam scan.

The scanner architecture of the ALTA 3000 system is illustrated in Figure 11.37. An argon-ion laser generates a single Gaussian beam at a wavelength of approximately 364 nm. A beam-splitter subassembly creates 32 separate beams, referred to as *the brush,* that pass through the AOM, which can independently turn on and off each of the beams. The modulation of the beams is controlled by the data path subsystem by varying the RF power to each channel of the AOM. The scan is created by a rotating polygonal mirror, and the scanned angle is converted to a spatial displacement by an $f - \theta$ lens. The final image of the brush at the photomask is obtained after transmitting it through a 20×, 0.6 *NA* reduction lens. At the image plane, the FWHM diameter of the spot size is approximately 360 nm.
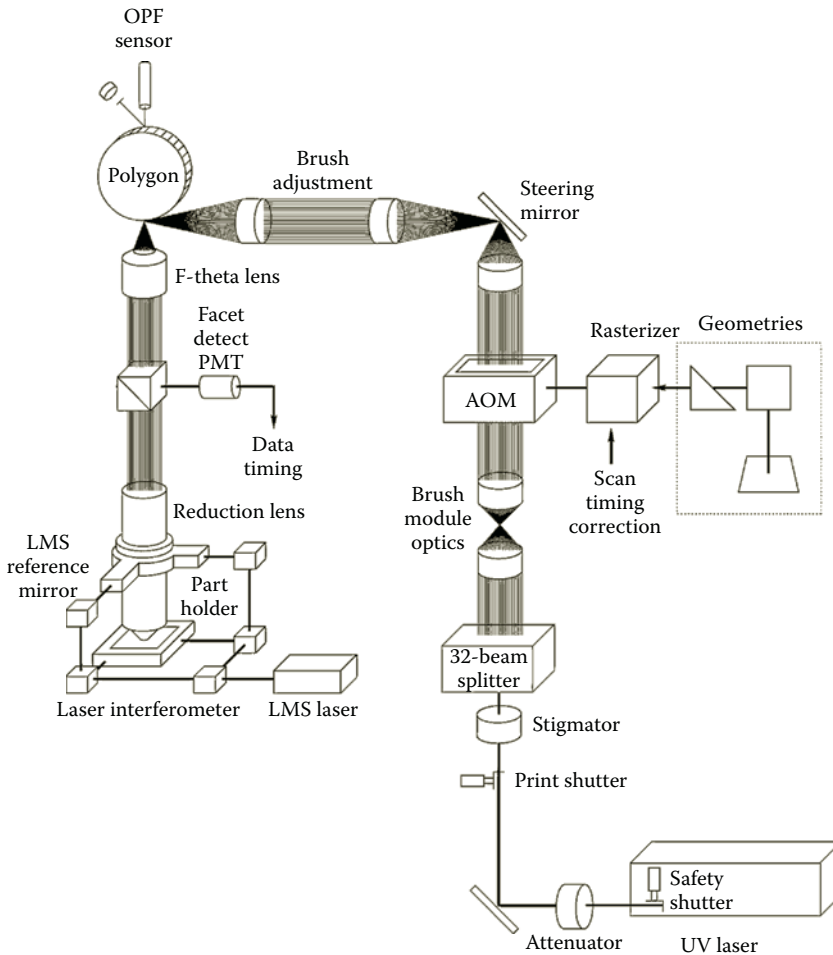
FIGURE 11.37
Use of a multichannel modulator in a precision semiconductor mask writing machine.

During the scan, a translation stage moves the photomask perpendicular to the direction of scan.

The AOM and other optics are made from UV-grade fused silica, which has excellent transmittance and resistance to radiation darkening at UV wavelengths. Unfortunately, the M_2 value for fused silica is relatively low, making power requirements for the modulator an important design parameter. A nominal drive level of 500 mW per channel at 200 MHz is required to achieve a diffraction efficiency of 50%.

The brush used to print at the photomask is an image of the one first created at the AOM. Therefore, the arrangement of beams at the AOM is dictated by the requirements for the final print. In this case, this arrangement is two sets of 16 beams with each beam separated by approximately 3 $1/e^2$ beam diameters. The size of the beams at the AOM is set at 144 μm $1/e^2$ diameter based on the modulation bandwidth requirement of 50 megapixels per second, and the beam-to-beam spacing is 412 μm.

The design of the transducer electrode size and shape must consider the impacts on power efficiency, beam distortion, and channel-to-channel cross-talk. The limited aperture

width and angular acceptance window of this design distort the output beam to a net far-field ellipticity of approximately 1.3:1, which is acceptable for this application. The modulator also introduces astigmatism, which would be detrimental to the system performance. This problem is corrected by precompensating optics referred to as the stigmator on Figure 11.37.

The tight channel spacing requires careful attention to the prevention of electrical cross talk in the feed network to the transducer electrodes. Radio frequency signals are fed to the modulator from the system electronics by an array of coaxial cables. The coaxial cables are connected to a printed circuit card with an individual trace for each of the 32 channels. Each trace ends with a land registering to one of the transducer electrodes. Bond wires are applied to make the final connection between the feed circuit and transducer electrodes.

When all 32 channels are on, this equates to 16 W of RF power. This creates localized heating in the AO cell near the transducer, also corresponding to the useful optical aperture of the AOM. Because the index of the fused silica is temperature dependent, these localized temperature gradients correspond to index gradients in the glass that may cause several waves of distortion across the optical aperture. A temperature gradient across the aperture will have the effect of an optical wedge, causing the angular orientation of the beams to shift. To minimize these effects, a heat sink is applied to the back side of the transducer to keep the temperature rise at the transducer as small as possible. The AOM mount is also water cooled to carry the bulk heat away from the device and the surrounding optical system.

11.8.2 Infrared Laser Scanning

AO beam scanners for use with IR lasers have been under consideration in recent years by the aerospace industry in connection with laser radar and optical communications systems. Where system requirements place excessive demands on mechanical scanning methods, various electronic approaches become attractive. In general, the carbon dioxide laser, with wavelengths from 9 to 11 μm, is the most common one for long-wavelength operation. There are a number of electronic approaches to IR beam scanning besides AO, and all of them are quite difficult to implement, usually for reasons related to the long interaction length needed to achieve large optical phase excursion. For AO diffraction, we have seen that the RF power needed for a given diffraction efficiency increases quadratically with wavelength. At 10.6 μm therefore, 280 times the power for 0.633 μm is necessary. Clearly, there will be severe constraints on the available materials for such devices, and on their performance. Referring to Table 11.4, we can see that only a few materials that transmit to 11 μm and have large AO figure of merit have been identified. The most common of these is germanium, which can be purchased in very large single crystals of excellent optical quality; germanium AO scanners have been commercially available for a number of years. They operate best in the isotropic mode, typically near 100 MHz RF. Another favorable IR material for use in the 9 to 11 μm range is thallium arsenic selenide, which has very recently become available on a commercial basis. This crystal is best used in an anisotropic mode, as described in a previous section. The AO figure of merit is very high due to the low value of shear acoustic wave velocity. The low velocity produces another result that may simplify the design of scanner optics. The scan angles at IR wavelengths are quite large; the angular dispersion for 10.6-μm carbon dioxide wavelength for this Bragg cell is shown in Figure 11.38. For an RF bandwidth of 30% around 110 MHz center frequency, a scan-angle range of 16° is reached. For many applications, no magnification of the scan angle will be needed, as may be the case for AO scanners in the visible.
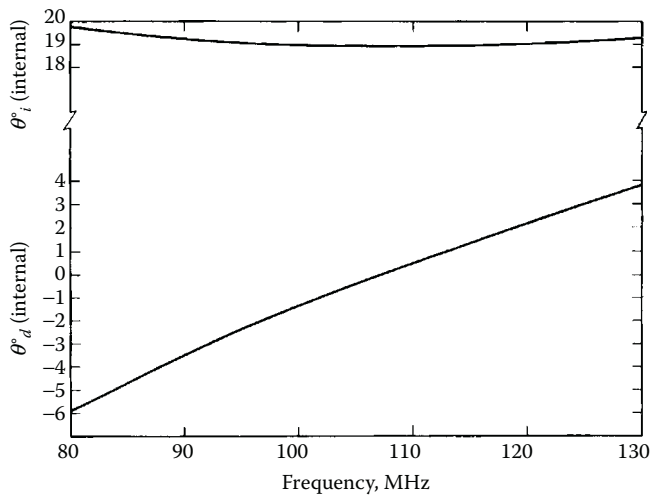
FIGURE 11.38
Anisotropic Bragg diffraction in TAS at $\lambda = 10.6$ μm.

One of the major problems associated with carbon dioxide laser beam scanners is heating, due both to absorption of optical energy and RF power heating. If very high laser beam powers are used, then even a small absorption coefficient in the scanner may cause unacceptable heating. For germanium and thallium arsenic selenide, the absorption coefficients at 10.6 μm are 0.032 and 0.015 cm^{-1}, respectively. The thermal conductivity of germanium is much higher than that of thallium arsenic selenide, but design considerations may favor one or the other, depending upon the detailed effects of thermal gradients. High RF power operation is limited by heating at the transducer, which will eventually damage the transducer bond. Such thermal effects can be reduced by water or air cooling the RF mount and by heat sinking the transducer; a photograph of a high-power transducer with a matching sapphire heat sink, which also serves to make electrical contact, is shown in Figure 11.39. The resolution of such IR AO scanners will, in general, be limited by RF heating if high diffraction efficiency is to be obtained. This comes about because a large interaction bandwidth requires a small interaction length, so that high efficiency can only be achieved by high power. The relationship between resolution and acoustic power for the IR scanner materials is shown in Figure 11.40. If CW operation is required, it can be seen that no more than a few hundred spots can be obtained for an aperture of 1 or 2 cm.

11.8.3 Two-Stage Acousto-Optic Scanner

AO deflection has been successfully applied to semiconductor photomask inspection. The KLA-Tencor 3000 series mask inspection machine checks photomasks for defects by scanning a focused spot across the mask and detecting the transmitted light (brightfield inspection). In this application, the feature sizes are very small (0.3 μm) compared to the size of the mask to be inspected (typically 6 in diameter). To achieve desired throughput, a net scan rate of 50 megapixels per second is required. The number of spots is too great to cover the photomask with a single scan, therefore the approach is to use a high-speed optical scanner to produce a lineal subscan, and use *x*–*y* mechanical translation to pass the photomask beneath the scan location. Significant control is required to provide the precise
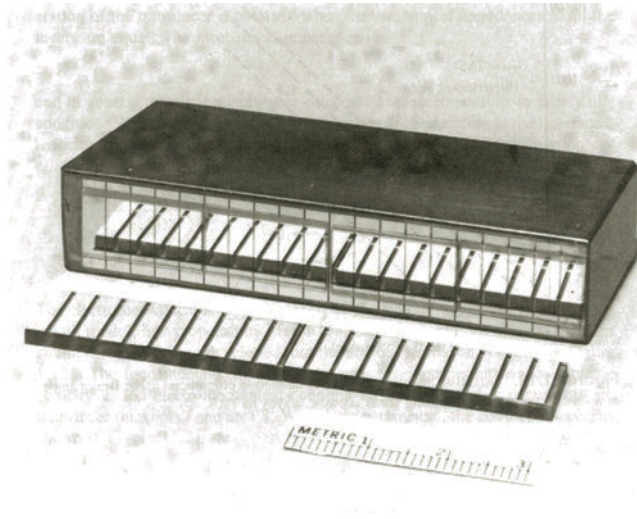
FIGURE 11.39
Transducer structure on TAS acousto-optic device, with matching heat sink for high RF power operation.
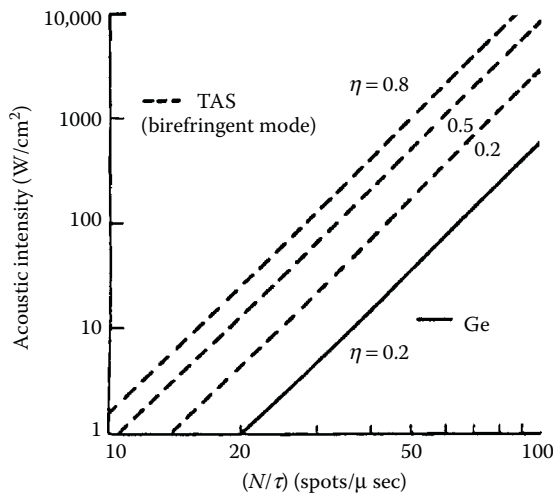
FIGURE 11.40
Acoustic intensity for various deflection efficiencies at the Bragg angle as a function of bandwidth using shear acoustic waves in TAS.

translation required and register all the subscan data together such that the data for the entire photomask can be seamlessly reconstructed.

11.8.3.1 Scanner Optics

The lineal scan is produced by an AO scan pair and has no moving parts. A block diagram of the scan optical portion of the system is shown in Figure 11.41. The source for the scan beam is an argon-ion laser operating at 488-nm wavelength. The beam is passed through a spatial filter and then focused to a 400-μm beam diameter at the first AO device, the
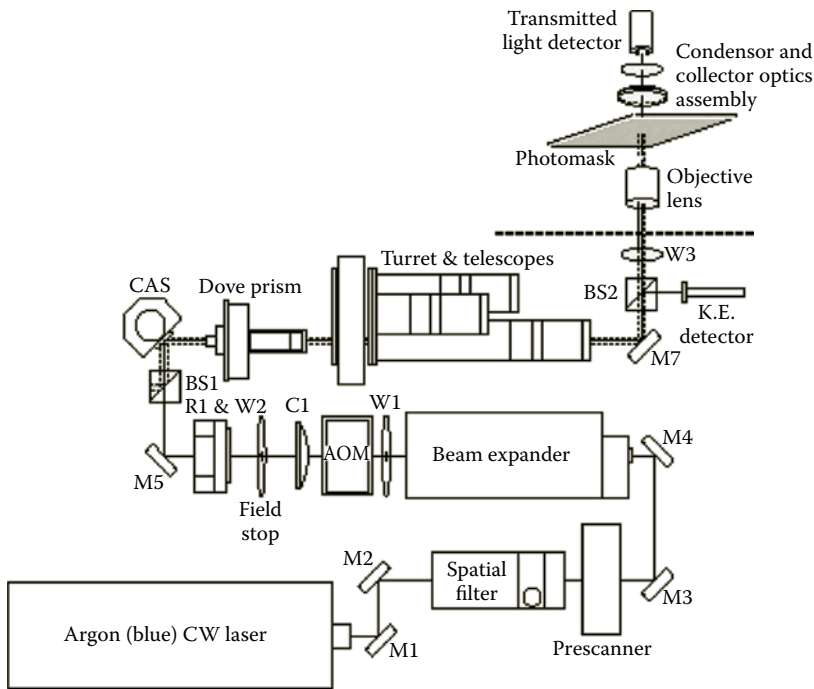
FIGURE 11.41
Use of an acousto-optic deflector (prescanner) and traveling chirp lens (AOM) in a semiconductor mask inspection machine.

predeflector. This device is made from SF-6 glass and operates in the longitudinal acoustic mode. The center frequency for the device is 90 MHz, and the operating bandwidth is 14.4 MHz. This corresponds to a deflection angle in air of 12.5 mrads ± 1 mrads (Figure 11.42).

Following the predeflector, prescan optics transform the deflector output into a telecentric scan with magnification appropriate for the traveling chirp lens. The spot size at the input of the traveling chirp device is 12mm $1/e^2$, and is designed to overfill the traveling chirp lens.

The chirp lens is produced from TeO_2, and uses the slow shear acoustic mode. Unlike the predeflector, the diffraction performance of a slow shear TeO_2 is very polarization dependent, and requires the input to be right circularly polarized. A wave plate is used to convert the linear polarization from the laser to circular polarization ahead of the TeO_2 device. The lens is formed from a linear chirp from 75 to 125 MHz over 7.5 µs. This makes the lens aperture in the scan direction 4.6 mm (7.5 × 0.616).

The aperture size in the cross-scan direction is controlled by the acoustic transducer height and is set to also be approximately 4.6 mm. The length of the scan is 14 µs or 8.6 mm. Therefore, the traveling chirp cell must have a clear aperture of at least 12.2 mm in the scan direction.

A cylinder lens is placed immediately after the chirp device to focus the spot on the cross-scan axis. A scan plane occurs at one focal length from the chirp lens. This is the object plane that is relayed and demagnified to produce the final scan at the photomask. Using the approximation $N = \tau \Delta f$, the number of resolvable spots from the predeflector is 1.6. While this performance would be of little use in direct scanning applications, its purpose in this application is to track the traveling chirp lens and maintain optimal illumination.

FIGURE 11.42
A typical 100-MHz-bandwidth acousto-optic modulator (courtesy MVM Electronics, Inc.).

The spot gain from the traveling chirp, also estimated from the time–bandwidth product, is 375. Therefore, the scan resolution is dominated by performance of the traveling chirp lens. The approximate scan size based on these approximations is 600 spots. However, due to limited aperture at the traveling chirp lens, the scanned spot is better approximated by an Airy disk function than a Gaussian, and the scan size based on the null-to-null spacing approximately 1000 spots.

11.8.3.2 Driver

Both AO devices are driven by analog electronics that consist of a voltage-controlled oscillator followed by amplifier stages. Linear voltage ramps are generated by the system electronics and supplied to the drivers to produce the linear frequency chirps required by the AO devices. Both drive inputs are derived from the same clock to ensure synchronization is maintained between the two devices. Note that chirp linearity is not critical for the predeflector as it would be with an f–θ configuration. In this system, the scan linearity is controlled by the propagation of the traveling lens and the dominant concern is variation of acoustic velocity in the traveling lens cell due to changes in temperature. Chirp linearity is critical for the traveling lens, as the nonlinearities in the chirp signal will appear as aberrations in the lens. Therefore, precompensation is included in the voltage ramp fed to the traveling chirp voltage-controlled oscillator to correct for inherent nonlinearities.

11.8.4 Applications of Acousto-Optic Devices and Acousto-Optic Tunable Filters

AO devices are used for a variety of applications and, depending on the area of specialization, comprise the following: AOM, AOD, frequency shifters (AOFS), tunable filters (AOTF), wavelength selectors (AOWS), and polychromatic modulators (PCAOM). Combining one of these AO devices with a mechanical scanner can produce scanning systems for laser printing, laser machining and engraving, wavelength-programmable scanning, lasik, medical and cosmetic laser systems, and many more. Here we will describe the above mentioned devices to provide insight into their limitations and typical product configurations.

The AO effect is used in a variety of ways for controlling light. From the widely utilized modulation of laser beams, solid-state light scanning, and frequency shifting, to tunable filters and polychromatic light modulation. The basic scheme for all these devices involves the selection of proper AO materials and the optimization of the piezoelectric transducer to produce acoustic waves with efficient conversion of electrical power to ultrasonic power for the desired frequency range. Except in colinear tunable filters, all schemes utilize the propagation of acoustic and optical waves approximately at 90° to each other. The theory of individual devices from the list above appears earlier in this chapter, this section serves to discuss the major aspects required to produce the desirable product in each category.

11.8.4.1 Acousto-Optic Modulators

The basic requirement of an AOM is the highest-speed or wide-bandwidth modulation of light with the highest optical throughput and list amount of electrical power. Limiting factors are the acoustic transit time and the limited fractional bandwidth of the piezoelectric transducer and the inefficient interaction of acoustic waves with optical waves. Optical transmission, damage range, and the acoustic attenuation of the material may become hard to overcome as we push the envelope of the design. The highest bandwidth achieved for an AOM is about 2 GHz or ~0.5 ns rise/fall time, with a top efficiency of less than 20%. These parameters are considerably degraded as the optical wavelength increases toward IR wavelengths such as 10.6 μm. As a variation of a single AOM a linear array of transducers has produced an AOM device with up to 128 channels (Figure 11.43 through 11.46).

11.8.4.2 Acousto-Optic Deflectors

The function of an AOD is to move or scan optical beams electronically. The dependence of the optical diffraction angle on acoustic wavelength makes this possible. The desired parameters are the maximum number of resolvable spots, the highest optical throughput
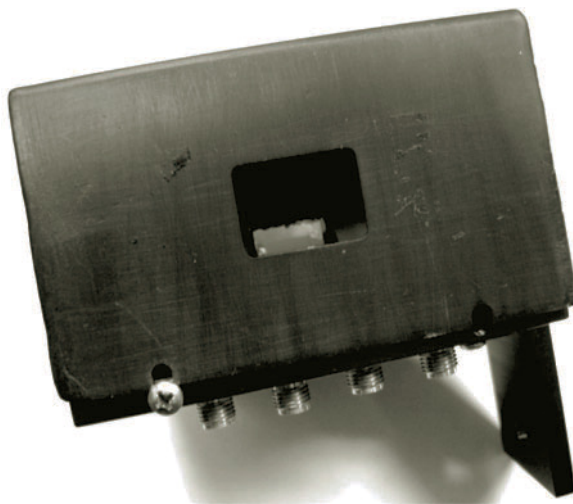
FIGURE 11.43
A typical 16-channel acousto-optic modulator (courtesy MVM Electronics, Inc.).
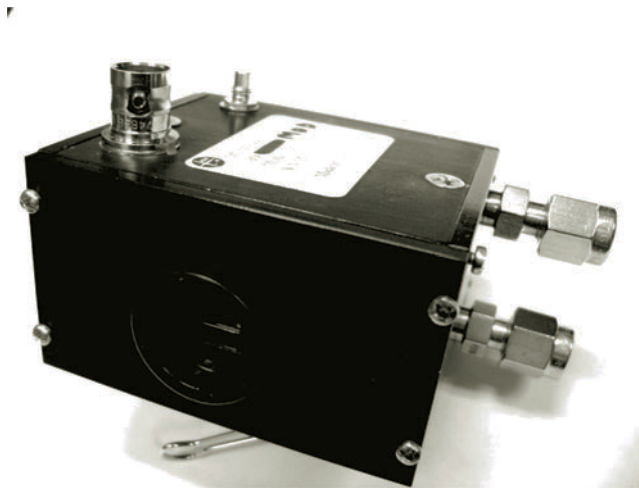
FIGURE 11.44
A typical Ge acousto-optic modulator for infrared modulation (courtesy MVM Electronics, Inc.).

FIGURE 11.45
A typical quartz AOM for laser Q-switching application (courtesy MVM Electronics, Inc.).

and the list electrical power. Here, the transducer bandwidth and acoustic attenuation of the material place severe limits on the achievable results. Speed of scanning and uniformity of optical throughput further reduce the performance. The maximum resolvable spots achieved are about 2000, with a scan time of about 10 μs or larger. Most AO materials have identical AO properties in orthogonal directions. This allows two-dimensional AO deflection with the same block of material.

AOD also serve another function in signal processing. The number of resolvable spots of an AOD scanner translates to the number of resolvable spectral components of the RF in AO spectrum analyzer applications and the number of corelations in AO corelator applications (Figure 11.47 and 11.48).
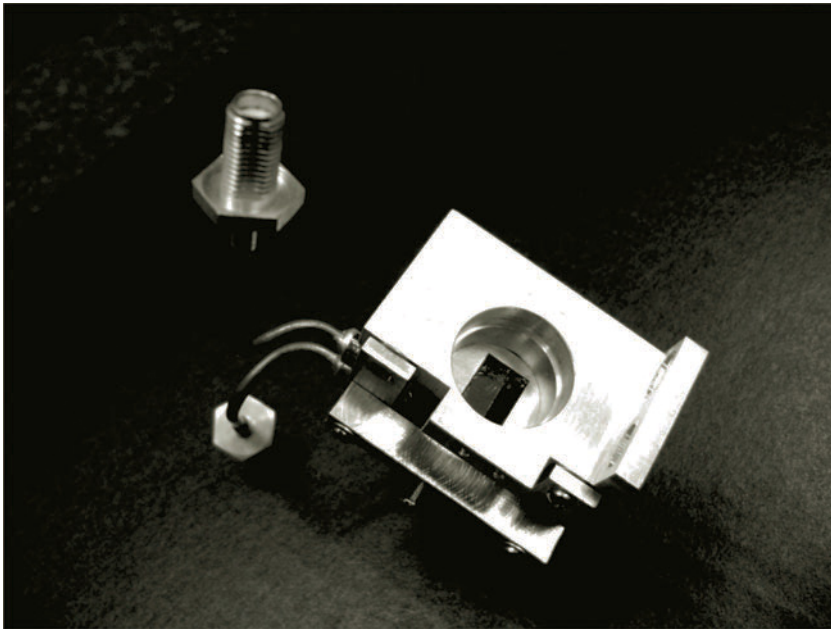
FIGURE 11.46
A typical two-channel GaP acousto-optic modulator for 1-GHz bandwidth modulation. (Courtesy MVM Electronics, Inc.)
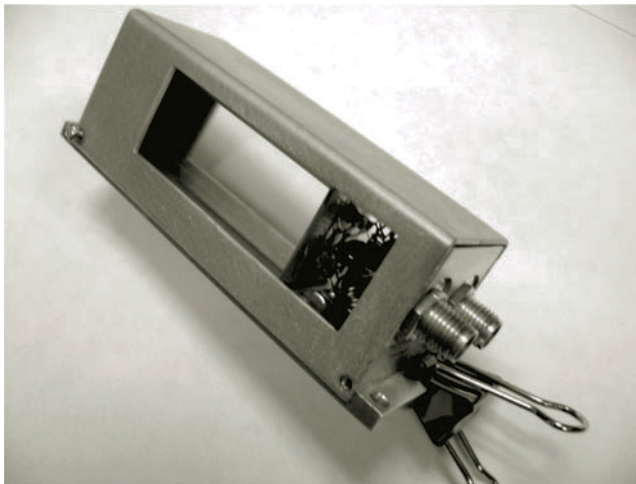
FIGURE 11.47
A typical 1000 resolvable spot acousto-optic deflector. (Courtesy MVM Electronics, Inc.)

11.8.4.3 *Acousto-Optic Frequency Shifters*

The traveling-wave interaction between light and acoustic waves imparts frequency shift to the diffracted light. Though AOFS generally require a narrow frequency band to function, some applications may desire a wide bandwidth. When the bandwidth is large, AOFS and AOM lose their distinction. If the frequency range is small, then the bandwidth of the piezoelectric transducer and the AO interaction can be traded off to achieve higher optical
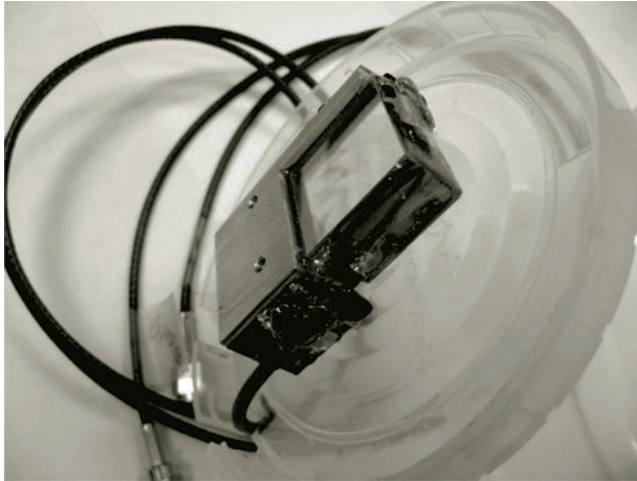
FIGURE 11.48
A typical two-dimensional acousto-optic deflector. (Courtesy MVM Electronics, Inc.)

FIGURE 11.49
6.85-GHz acousto-optic frequency shifter. (Courtesy MVM Electronics, Inc.)

throughput at a much higher center frequency. AO frequency shift to 10 GHz has been achieved, at least for low-percentage optical throughput (Figure 11.49).

11.8.4.4 *Acousto-Optic Tunable Filters*

The AO scheme is the only one available that allows electronically selectable imaging of narrow-band, multispectral optical components from a wide-spectrum light. In colinear propagating acoustic and optical wave devices, diffraction of one optical polarization to another can occur with Bragg matching k-vector supplied by the acoustic waves. Here, the interaction length can be very large resulting in narrow optical bandwidth diffraction. Using simultaneous multitone RF, many narrow optical bands can be diffracted. Moreover,

FIGURE 11.50
Tellurium dioxide acousto-optic tunable filter in the visible and infrared range. (Courtesy MVM Electronics, Inc.)
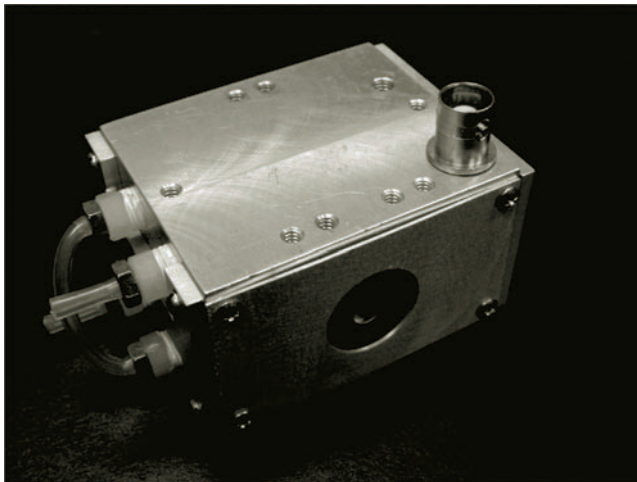
FIGURE 11.51
Crystalline quartz acousto-optic tunable filter in the ultraviolet range. (Courtesy MVM Electronics, Inc.)

the angular sensitivity of the AO interaction is small, resulting in good image preservation. One drawback of the colinear interaction is the difficulty of separating the diffracted and undiffracted components.

In certain materials such as tellurium dioxide and crystalline quartz, a so-called "parallel tangents Bragg matching condition" is possible. This condition allows orthogonal AO interaction and removes the difficulty of separating diffracted and undiffracted optical components (Figure 11.50 and 11.51).

The limitations of AOTF are set by the availability of large-size material, acoustic attenuation, transmission range of the selected material, and RF drive power requirements.
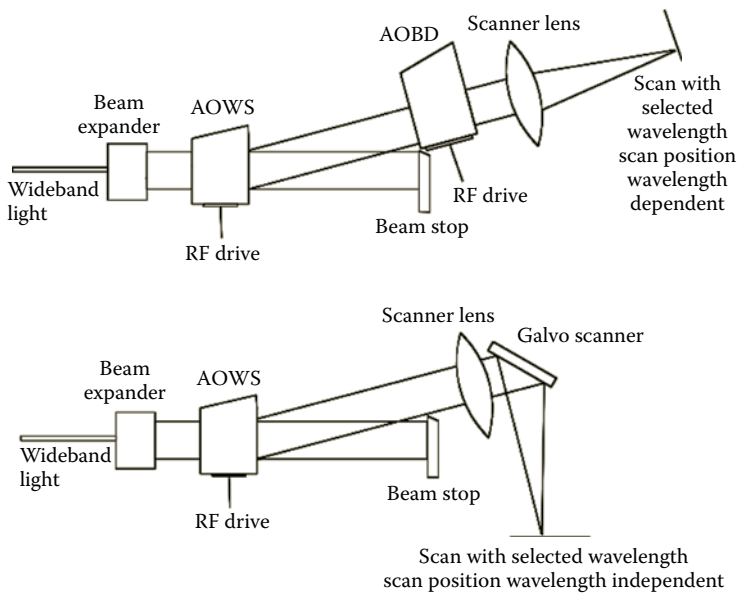
FIGURE 11.52
A schematic of a variable-wavelength optical scanner.
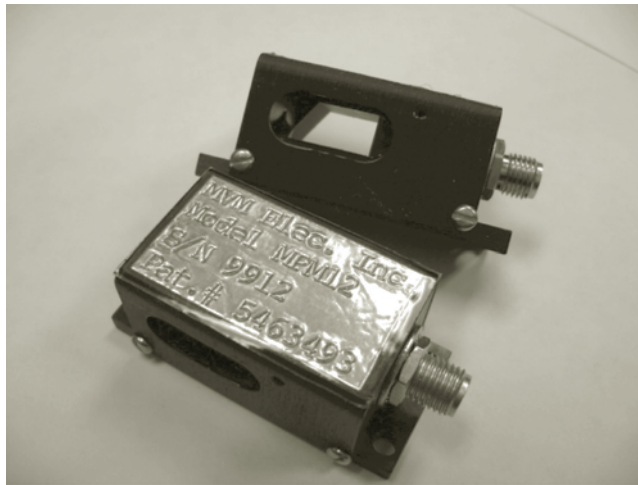
FIGURE 11.53
Device for full color argon and krypton laser projector systems used in entertainment and planetarium applications. (Courtesy MVM Electronics, Inc.)

11.8.4.5 Acousto-Optic Wavelength Selectors

In certain applications, the imaging function available in AOTF is not required. This relaxes several AOTF parameters that can be traded to reduce requirements for RF drive power and material size. In that case AOTF can function as AOWS. If a galvanometric or other mechanical scanner is placed after the AOWS, we could obtain a variable-wavelength optical scanner that may have wide-ranging applications (Figure 11.52).

11.8.4.6 *Polychromatic Acousto-Optic Modulators*

If the diffraction efficiency of an AOWS can be improved further such that almost 100% diffraction can occur with relatively low RF drive power in the range of 20–100 mW, then many RF tones approaching 16 can be used for the AOWS; this device is known as polychromatic AOM. An argonion laser contains six to ten wavelengths in the visible spectrum. Polychromatic AO modulation has been widely used for laser projection and laser shows to produce vivid, full-color images for entertainment and planetarium applications (Figure 11.53).

Material degradation at high laser power and linear polarization restriction is of concern in some applications.

11.9 CONCLUSIONS

AO devices have been utilized in a variety of scanning applications from direct spatial or temporal modulation to predeflection or postscan lensing. For most applications of these devices there are significant design trades between modulation bandwidth, efficiency, and other performance parameters. AOD are often advantageous in applications where high precision is required over a relatively small angular scan. AOM are effective for pixelation of raster scans or video modulation when it is impractical to directly modulate the optical source. The ability to modulate light without moving parts should continue to make acousto-optics an attractive technology in the future.

ACKNOWLEDGMENTS

REFERENCES

1. Debye, P.; Sears, F.W. *Proc. Natl. Acad. Sci.* 1932, *18*, 409.
2. Lucas, R.; Biquard, P. *J. Phys. Rad.* 1932, *3*(7), 464.
3. Raman, C.F.; Nath, N.S.N. *Proc. Indian Acad. Sci.* I 1935, *2*, 406.
4. Gordon, E.I. *Proc. IEEE* 1966, *54*, 1391.
5. Dixon, R.W. *IEEE J. Quantum Electronics* 1967, *QE-3*, 85.
6. Harris, S.E.; Nieh, S.T.R.; Winslow, D.K. *Appl. Phys. Lett.* 1969, *15*, 325.
7. Mertens, R. Meded. K. *Vlaam. Acad. Wet. Lett. Schone Künsten Relg., Kl. Wet.* 1950, *12*, 1.
8. Exterman, R.; Wannier, G. *Helv. Phys. Acta* 1936, *9*, 520.
9. Klein, W.R.; Hiedemann, E.A. *Physica* 1963, *29*, 981.

10. Nomoto, O. *Jpn. J. Appl. Phys.* 1971, *10*, 611.
11. Klein, W.R.; Cook, B.D. *IEEE Trans. Sonics Ultrason.* 1967, *SU-14*, 723.
12. Korpel, A. *J. Opt. Soc. Am.* 1979, *69*, 678.
13. Korpel, A.; Poon, T. *J. Opt. Soc. Am.* 1980, *70*, 817.
14. Chang, I.C. *IEEE Trans. Sonics Ultrason.* 1976, *SU-23*, 2.
15. Uchida, N.; Ohmachi, Y. *J. Appl. Phys.* 1969, *40*, 4692.
16. Warner, A.W.; White, D.L.; Bonner, W.A. *J. Appl. Phys.* 1972, *43*, 4489.
17. Cohen, M.; Gordon, E.I. *Bell Syst. Tech. J.* 1965, *44*, 693.
18. Dixon, R.W. *J. Appl. Phys.* 1962, *38*, 5149.
19. Young, E.H.; Yao, S.K. *Proc. IEEE* 1981, *69*, 54.
20. Foster, L.C.; Crumly, C.B.; Cohoon, R.L. A high-resolution linear optical scanner using a traveling-wave acoustic lens. *Appl. Opt.* 1970, *9*, 2154–2160.
21. Pinnow, D.A. *IEEE J. Quantum Electronics* 1970, *QE-6*, 223.
22. Gladstone, J.H.; Dale, T.P. *Phil. Trans. Roy. Soc. London* 1964, *153*, 37.
23. Wemple, S.H.; DiDomenico, M. *J. Appl. Phys.* 1969, *40*, 735.
24. Uchida, N.; Niizeki, N. *Proc. IEEE* 1973, *61*, 1073.
25. Woodruff, T.O.; Ehrenreich, H. *Phys. Rev.* 1961, *123*, 1553.
26. Mueller, H. *Phys. Rev.* 1935, *47*, 947.
27. Abrams, R.L.; Pinnow, D.A. *J. Appl. Phys.* 1970, *41*, 2765.
28. Dixon, R.W. *J. Appl. Phys.* 1967, *38*, 5149.
29. Pinnow, D.A.; Van Uitert, L.G.; Warner, A.W.; Bonner, W.A. *Appl. Phys. Lett.* 1969, *15*, 83.
30. Coquin, G.A.; Pinnow, D.A.; Warner, A.W. *J. Appl. Phys.* 1971, *42*, 2162.
31. Ohmachi, Y.; Uchida, N. *J. Appl. Phys.* 1969, *40*, 4692.
32. Gottlieb, M.; Isaacs, T.J.; Feichtner, J.D.; Roland, G.W. *J. Appl. Phys.* 1969, *40*, 4692.
33. Roland, G.W.; Gottlieb, M.; Feichtner, J.D. *Appl. Phys. Lett.* 1972, *21*, 52.
34. Isaacs, T.J.; Gottlieb, M.; Feichtner, J.D. *Appl. Phys. Lett.* 1974, *24*, 107.
35. Feichtner, J.D.; Roland, G.W. *Appl. Optics* 1972, *11*, 993.
36. Harris, S.E.; Wallace, R.W. *J. Opt. Soc. Am.* 1969, *59*, 744.
37. Sittig, E.K. *IEEE Trans. Sonics and Ultrasonics* 1969, *SU-16*, 2.
38. Meitzler, A.H.; Sittig, E.K. *J. Appl. Phys.* 1969, *40*, 4341.
39. Mason, W.P. *Electromechanical Transducers and Wave Filters*; Van Nostrand Reinhold: Princeton, NJ, 1948.
40. Sittig, E.K. *IEEE Trans. Sonics and Ultrasonics* 1969, *16*, 2.
41. Meitzler, A.H. *Ultrasonic Transducer Materials*; Mattiat, O.E., Ed.; Plenum: New York, 1971.
42. deKlerk, J. *Physical Acoustics*; Mason, W.P., Ed.; Academic Press: New York, 1970; Vol. IV, Chap. 5.
43. deKlerk, J. *IEEE Trans, on Sonics and Ultrasonics* 1966, *SU-13*, 100.
44. Weinert, R.W.; deKlerk, J. *IEEE Trans, on Sonics and Ultrasonics* 1972, *SU-19*, 354.
45. Coquin, G.; Griffin, J.; Anderson, L. *IEEE Trans, on Sonics and Ultrasonics* 1971, *SU-7*, 34.
46. Korpel, A. et al. *Proa IEEE* 1966, *54*, 1429.
47. Pinnow, D.A. *IEEE Trans, on Sonics and Ultrasonics* 1971, *SU-18*, 209.
48. Eschler, H. *Optics Communications* 1972, *6*, 230.
49. Sittig, E.K.; Cook, H.D. *Proc. IEEE* 1968, *56*, 1375.
50. Konog, W.F.; Lambert, L.B.; Schilling, D.L. *IRE Int. Conv. Rec.* 1961, *9* (6), 285.
51. Larson, J.D.; Winslow, D.K. *IEEE Trans. Sonics and Ultrasonics* 1971, *SU-18*, 142.

12

Electro-Optical Scanners

Timothy K. Deis

Consultant
Pittsburgh, Pennsylvania, USA

Daniel D. Stancil

Carnegie Mellon University
Pittsburgh, Pennsylvania, USA

Carl E. Conti

Consultant
Hammondsport, New York, USA

CONTENTS

12.1 INTRODUCTION

Electro-optic deflection systems are most typically considered in applications where high deflection speed is the paramount selection criteria. They can also have other attractive features, such as high optical efficiency and physical robustness, but these are usually secondary concerns after deflection speeds that can exceed 10^9 rad/s.

Requirements for great speeds are driven by advancements in two technical areas: lasers and computing. Progress in laser technology has resulted in reliable, low-cost compact sources with high power. Many applications, such as material marking, require a certain amount of energy to be delivered and are somewhat independent of the time period that the energy is delivered over. More powerful lasers can deliver the required energy dose in a short period of time, putting pressure on deflection system designers for higher operating speeds.

Advances in computing and communications have resulted in ever-higher data rates delivered to the laser deflection system. Data rates in applications such as displays now far surpass the control bandwidth of most mechanical systems such as galvanometers, piezoelectric deflectors, and microelectromechanical systems (MEMS) mirrors, which are limited by mechanical inertia effects. Relatively low bandwidth, usually less than 10 kHz, makes these systems a poor choice if a laser or other light source can deliver the required energy dose within a time period that matches the desired data rate, which often exceeds MHz speeds. In the realm of electrical circuits, electro-optic (EO) deflectors are essentially capacitors, which can be charged very quickly with appropriate driver circuitry and can operate at MHz speeds with ease.

The most common technical competition to EO deflection is acousto-optic (AO) deflection, covered elsewhere in this volume. Electro-optic systems are often faster and more

optically efficient than AO systems. They often can handle higher beam powers than AO systems because there is no need to achieve a tight beam focus in the device for maximum speeds. The AO systems can be designed to have much higher resolutions, offering hundreds or thousands of resolvable spots compared to 5 to 100 that are more typical of EO systems. Both EO scanners and many AO scanners have polarization-dependent characteristics, which can limit their application. In general, the crystal and glass materials for AO devices are more readily available and have uniform high quality, unlike the materials for EO deflectors, which can be hard to procure or may have uncertain property characteristics. Electronic drivers for AO deflection systems usually require higher input power than those for EO systems.

Examples of current applications where the combination of high-power compact laser sources and high-speed data delivery are driving the evaluation or adoption of EO deflection systems include: encoding image data onto printing plates in plate setting machines, controlling optical paths in communications switching systems, creating jitter-free laser projection displays, and maintaining alignment in free space communications systems over long distances.

Electro-optic deflectors or scanners can be configured to meet a variety of requirements. Choices of the optical material, typically a crystal, the drive electronics, and optical path can be made to address all of the following functional requirements:

1. Error correction: high-speed (10^6 rad/s) and low deflection (order of milli-rad) characteristics can be used to provide facet-to-facet error correction of polygon scanners. The same characteristics can also be used to "debounce" a galvanometer-based display system.

2. Switching: high speed (<100 ns steps) and low deflection (order of 10 resolution spots) can be used to switch collimated light in fiber optic switches, optical backplanes, and optical computers.

3. Modulation: in cases where it is not possible to directly modulate a laser, as with some diode-pumped sources, a deflector can be used to modulate the beam by deflecting it toward or past a beam stop. Applications of this sort include displays, printing plate production, and marking.

There is a class of EO device that is sometimes confused with deflectors or scanners. This class uses the Kerr effect effectively to rotate the polarization of light traversing the device. Such devices can be used as modulators for polarized light sources by rotating the plane of polarization to be parallel or perpendicular to a polarizer downstream in the system. Such systems are commercially available, and have been used in some printing applications.

If a polarization-dependent mirror or beam splitter is used in place of the polarizer in the above system, the beam may be switched to one of two output positions by electro-optically controlling the polarization of the beam. These devices may be cascaded to produce many output positions. These and related systems are not the focus of the current text and will not be covered further.

The present effort builds on the previous edition, authored by Clive L.M. Ireland and John Martin Ley.[1] Much material has been carried over, especially that related to materials properties and basic physics. New material has been added to cover domain-inverted scanner designs, which have recently been applied in commercial products. A new section was added to address some of the options and characteristics of electronic drivers suitable

for commercial applications. Experience has shown that this part of the system can be as difficult to realize as the optical elements. As with the material relating to optical materials and designs, the electronic driver material is intended to serve as a general reference for guiding development efforts, not as a definitive treatise on the subject.

12.2 THEORY OF THE ELECTRO-OPTIC EFFECT

12.2.1 The Electro-Optic Effect

The EO deflectors and scanners discussed below all rely on the EO effect evidenced to some degree by all materials. Only a few materials exhibit property changes with an applied electric field large enough to be exploited for use in deflecting and switching applications. The index of refraction of these materials, typically crystals, changes when an electric field is applied to such a degree that it can be used to usefully deflect a beam according to the rules of normal refractive optics.

In crystals, the direction of the polarization induced by an applied electric field may differ in direction from that of the applied field. Mathematically, this means that the relative permittivity must be represented by a second-rank tensor:

$$D_i = \varepsilon_0 \kappa_{ij} E_j = \varepsilon_0 E_i + P_i \tag{12.1}$$

where ε_0 is the permittivity of free space, κ_{ij} is the relative dielectric tensor, and E_i and P_t are the ith components of the electric field and the induced polarization, respectively, and summation over repeated indices is assumed. We will restrict our discussion to crystals that are essentially neither magnetic nor optically active, and that exhibit negligible absorption. In this case, κ_{ij} is a real, symmetric tensor.

A convenient geometric representation of any symmetric second-rank tensor S_{ij} is an ellipsoidal or hyperboloidal surface defined by

$$S_{ij} x_i x_j = 1 \tag{12.2}$$

Thus we can construct such a surface for κ_{ij}, or its inverse $(1/\kappa)_{ij}$. In contrast, the refractive index—given by the square root of the relative permittivity in isotropic materials—does not transform as a second-rank tensor. Since $\kappa = n^2$ in an isotropic material, it is conventional to adopt the following notation:

$$\left(\frac{1}{n^2}\right)_{ij} x_i x_j = 1 \tag{12.3}$$

This ellipsoidal surface is called the index ellipsoid, or indicatrix. In the coordinate system in which $(1/n^2)_{ij}$ is diagonal, Equation 12.3 reduces to

$$\frac{x^2}{n_x^2} + \frac{y^2}{n_y^2} + \frac{z^2}{n_z^2} = 1 \tag{12.4}$$

This surface has a simple geometric interpretation. The principal axes of the ellipsoid correspond to directions in the crystal for which D is parallel to E, and the refractive indices for waves polarized along these directions are n_x, n_y, and n_z.

12.2.2 The Linear Electro-Optic Effect

The crystal materials commonly used for EO devices do not possess inversion symmetry (see Stancil[2] for a complete explication of the tensor properties of crystals), meaning that the application of an electric field induces a small change in the refractive index that is proportional to the field, and reverses in sign when the field reverses. This is known as Pockel's Effect, or the Linear Electro-Optic Effect.

In the presence of a uniform electric field, the changes in the indices of refraction of such materials can be shown to be

$$\Delta\left(\frac{1}{n^2}\right)_{ij} = r_{ij,k}E_k \tag{12.5}$$

where $r_{ij,k}$ is the linear EO tensor, whose values are readily available from data published in the literature and from crystal suppliers.

As an example, consider the EO effect in KH_2PO_4, also known as KDP. In this crystal (and all crystals with symmetry $\overline{4}2m$) the only nonzero components of the EO tensor are $r_{41} = r_{52}$, and r_{63}. For simplicity, we will consider the case of a static electric field applied along the optic axis, E_3. For further simplicity, for small Δn, which is normally the case, $\Delta(1/n^2) = -2\Delta n/n^3$. Thus, the refractive index seen by an optical wave polarized along the $\langle 110 \rangle$ direction is approximately given by

$$n_1 = n_o - \frac{1}{2}n_o^3 r_{63}E_3 \tag{12.6}$$

where use has been made of the fact that $r_{63}E_3 1/n_o^2$. Similarly, the index along the $\langle 1\overline{1}0 \rangle$ direction is

$$n_2 = n_o + \frac{1}{2}n_o^3 r_{63}E_3 \tag{12.7}$$

The index along the optical axis $\langle 001 \rangle$ is unchanged ($n_3 = n_e$).

Note that the index change observed depends on the polarization of the light transiting the region with the applied field. This effect makes the performance of many EO scanners "polarization dependent," where the deflection achieved depends on the polarization of the beam. This can result in splitting randomly polarized beams and is the reason that most EO scanners are used with polarized beams only.

12.2.3 The Quadratic Electro-Optic Effect

Refractive index changes proportional to the square of the applied field are permitted by symmetry in all materials. Besides crystals such as KDP and $LiNbO_3$, liquids that are strongly polar are of particular EO interest since they can exhibit a high anisotropic, optic polarizability. By applying a strong external field, the molecules of these substances partially align with the field, causing the bulk material to become biréfringent.

The component of a beam polarized parallel to the main polarizability of the molecules, usually nearly parallel to the dipole moment of the molecule, sees an increase in refractive index relative to that of the orthogonal polarization. This effect, which was observed by Kerr in glass and other materials, is generally described by the following equation:

$$n_p - n_s = B\lambda E^2 \tag{12.8}$$

Here λ is the vacuum wavelength of the beam, B is the Kerr constant for the material, and n_p and n_s are the parallel and orthogonal refractive index components, respectively, and E is the applied electric field.

A variety of Kerr materials and devices have been studied in the past, see Lee and Hauser[3] and Kruger et al.[4] At the time of this writing, late 2002, there do not appear to be any commercial devices or systems based on the quadratic EO effect in production.

12.3 PRINCIPAL TYPES OF ELECTRO-OPTIC DEFLECTORS

12.3.1 Basic Topologies

The EO effect can be utilized in a variety of basic ways, with details seemingly limited only by the imagination of very clever practitioners. The design problem is one of selecting a gross geometry that the beam can traverse, and then selecting the geometry and magnitude of desired index change. To provide some order, the following nomenclature is used:

1. Shaping of gross geometry: at its simplest, an EO scanner can consist of a prismatic crystal element with electrodes covering both ends. As the potential across the element is changed, it acts like an electrically controlled prism.

2. Electrically shaped fields: When a bulk prism would be too small for easy handling, or if the field strength desired would result in electrical breakdowns across the side surfaces, electrical fields may be shaped. This is normally achieved by using shaped electrodes of constant potential. A less common approach is to use electrodes with finite or graded conductivity so that a varying voltage profile is obtained when a current is applied.

3. Poled structures: some EO materials, such as $LiNbO_3$ and $LiTaO_3$, may be "poled," a process that can result in geometrically precise crystalline domains within the bulk material. A uniform electric field applied to poled structures will result in equal and opposite changes in index within each domain according to its orientation.

12.3.2 Terminology for Describing Electro-Optic Scanners

Electro-optic scanners can be discussed using much of the same terminology as other types of scanners. The difficult trade-offs between size, voltage, and material properties create some nuances that are important to consider.

12.3.2.1 Beam Displacement and Deflection Angle

A schematic EO scanner is shown in Figure 12.1. In the presence of a linear index variation (constant gradient) across the width of the scanner, the trajectory of the center of the beam within the scanner is described by the parabolic relation[5]

$$X(z) = \frac{1}{n}\frac{dn}{dx}\frac{z^2}{2} \approx \frac{1}{2}\frac{\Delta n}{n}\frac{z^2}{W} \tag{12.9}$$
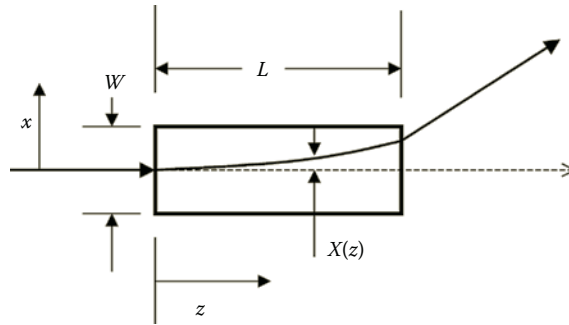
FIGURE 12.1
Schematic electro-optical scanner of width W and length L. An index of refraction profile of the form $n(x) = n_0 + kx$ is assumed. The optical beam to be deflected enters from the left.

where $X(z)$ is the displacement of the beam center from the optical axis at position z, Δn is the total change in index across the scanner, n is the nominal index of refraction (in the absence of an EO shift), and W is the width of the scanner.

The deflection angle at a particular position z is given by the slope of the trajectory at that point, or the derivative of Equation 12.9:[5,6]

$$q_{in}(z) = \frac{1}{n}\frac{dn}{dx}z \approx \frac{\Delta n}{n}\frac{z}{W} \tag{12.10}$$

When the beam exits the material, the angle is increased by the factor n, owing to the small-angle form of Snell's Law. The external deflection angle for the scanner is therefore obtained by evaluating Equation 12.10 at $z = L$ and multiplying by n:

$$q_{def} = \frac{dn}{dx}L \approx \Delta n \frac{L}{W} \tag{12.11}$$

The displacement of the beam at the output facet of the scanner is finally given by

$$d = \frac{1}{2}\frac{\Delta n}{n}\frac{L^2}{W} \tag{12.12}$$

12.3.2.2 Pivot Point

Comparison of Equations 12.12 and 12.10 shows that the output displacement can also be expressed as

$$d = \frac{1}{2}q_{in}L \tag{12.13}$$

This suggests that although the actual trajectory is parabolic, the output angle and displacement are correctly given by assuming that the beam has an abrupt deflection of θ_{in} a distance of $L/2$ from the output plane.[7] We call this the *pivot point*, and define it more generally as

$$L_{P,in} = \frac{X(L)}{q_{in}(L)} \tag{12.14}$$

A scanner has a well-defined pivot point when $L_{P,in}$ does not depend on the magnitude of the index variation Δn.

When the beam deflection is viewed outside the scanner, the deflection also appears to have a pivot point, although displaced, just as an object on the bottom of a swimming pool appears to be displaced when viewed from outside the pool. From outside the scanner, the pivot point appears to be a distance L_P from the output plane:

$$L_P = \frac{X(L)}{q_{def}(L)} = \frac{L_{P,in}}{n} \tag{12.15}$$

The existence of a pivot point is significant for the design of optical systems containing EO scanners. From the optical system point of view, the scanner can be represented simply by a mirror at a distance L_P from the output plane that introduces the deflection θ_{def}.

12.3.2.3 Resolvable Spots

Perhaps the best way to compare various scanner technologies is to use the number of "resolvable spots." The deflection angle can be magnified or reduced with other optical elements, but the number of resolvable spots will remain constant. The number of resolvable spots is given by the number of beam diameters corresponding to the lateral displacement at a certain distance, usually the far field. Clearly this number depends on the way that the beam diameter is defined. For our purposes, we will assume the beam is well described by a fundamental Gaussian beam whose $1/e^2$ intensity radius is given by the Gaussian beam waist $w(z)$:

$$w(z) = w_0 \sqrt{1 + \left(\frac{\lambda(z - z_0)}{p w_0^2} \right)^2} \tag{12.16}$$

where w_0 is the minimum radius at the beam waist, λ is the optical wavelength, and z_0 is the location of the minimum radius, or waist. The number of resolvable spots is given by

$$N_U(z) = \frac{S(z)}{2w(z)} + 1 \tag{12.17}$$

where N_U is the number of spots assuming unipolar deflection (deflection only to one side of the optical axis) and S is the displacement of the beam at the observation plane (see Figure 12.2). Often scanners are used with bipolar drive voltages, resulting in a total beam displacement of $2S$. Thus the number of bipolar resolvable spots is

$$N_B(z) = \frac{S(z)}{w(z)} + 1 \tag{12.18}$$

The displacement S at a distance z from the pivot point is given by

$$S = q_{def} z \tag{12.19}$$
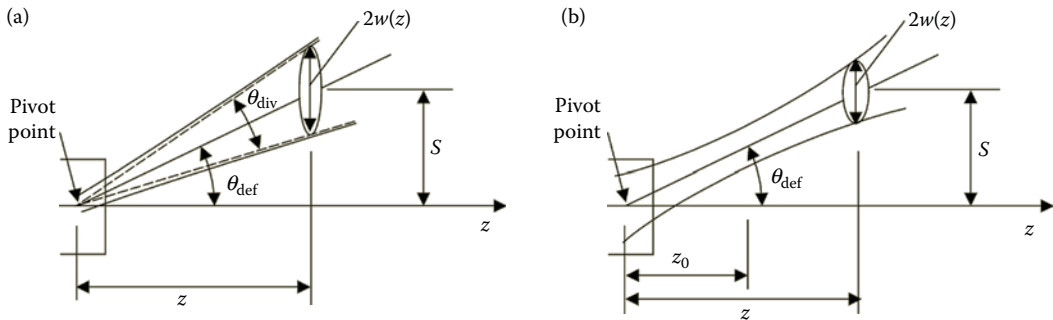
FIGURE 12.2
Geometry illustrating the concept of resolvable spots: (a) beam focus/waist near the pivot point of the scanner, and (b) beam focus/waist beyond the output.

It is instructive to consider the case with the waist collocated with the pivot point ($z_0 = 0$) and the observation point arbitrarily far away. This limit results in the number of spots in the far field:

$$N_{U,FF} \approx \frac{q_{def} p w_0}{2 \lambda} + 1 = \frac{q_{def}}{q_{div}} + 1 \qquad (12.20)$$

$$N_{B,FF} \approx \frac{2 q_{def}}{q_{div}} + 1 \qquad (12.21)$$

where

$$q_{div} = \frac{2 \lambda}{p w_0} \qquad (12.22)$$

is the far-field divergence (full angle) of the Gaussian beam. The displacement clearly gets larger the further away the observation plane is, but this does not result in an increase in spots since the beam divergence causes the spot diameter to increase at the same rate in the far field.

It is important to note that the maximum number of resolvable spots is achieved only in the far field. Many practical EO systems, such as deflection-based modulators, do not require operation in the far field. It is critically important to perform accurate simulations and ray tracings as part of the design process.

Other definitions of beam diameter are sometimes used, depending on the application. For instance, flat-top power profiles are useful for some laser machining operations. The effect of the diameter definition, and the correspondingly correct description of beam divergence, should be considered when discussing the number of resolvable spots that a particular system may exhibit.

12.3.3 Single Elements and Assemblies of Single Elements

One of the most basic optical elements is a prism. Accordingly, one of the most basic EO elements is an electrically controlled prism (Figure 12.3).
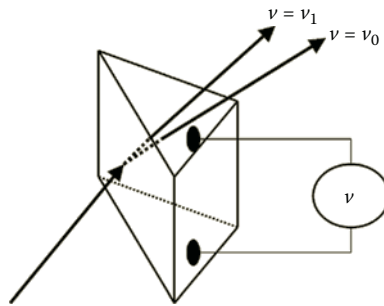
FIGURE 12.3
The simplest electro-optical scanner: a prism fabricated from an electro-optic material having electrodes at each end. As the voltage is varied, the angle of the exiting beam is changed.
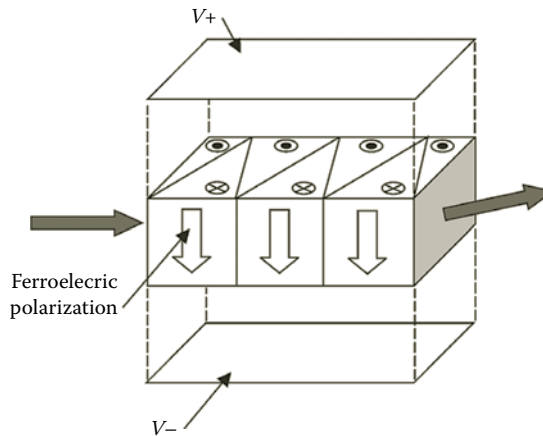
FIGURE 12.4
Electro-optic prism scanner made from alternating discrete prisms. From Lotspeich, J.F. Electrooptic light-beam deflection. *IEEE Spectrum* 1968, 5, February, 45–52. With permission.

In practice, the actual deflection about the mean is very small. For example, an equilateral triangle prism fabricated from lithium niobate and operated to ±1 kV/mm will undergo an index change of approximately ±0.0002, or 0.01%. This change is actually smaller than the accuracy bounds typically quoted by commercial crystal manufacturers for property values.

As discussed previously, the sign of the refractive index change reverses with the direction of polarization (usually coincident with the optic axis of the crystal). Thus a direct extension of the single prism implementation is to assemble several discrete prisms with alternating polarization as shown in Figure 12.4. With proper choice of optical polarization, crystal, and orientation, a voltage applied between conducting electrodes on opposing transverse sides of the assembly increases and decreases the index in alternating prisms generating the Δn required for scanner operation. The major disadvantage of this type of construction is the labor-intensive cutting, polishing, and assembly of the prisms.

12.3.4 Shaped Fields

An analysis of the energized multi-element assembly shows that it can be accurately represented by a band of material with linearly varying index of refraction.[5] This leads to

two interesting design realizations: graded electrical fields with uniform crystal structure, covered in this section, and uniform electrical field and graded crystal structures created through domain inversion, covered in the following section.

12.3.4.1 Graded Index with Uniform Applied Voltage

One way to create a linearly varying spatial index profile is to apply a linearly varying electric field to an EO crystal. Most electrical circuits, especially those designed for low-power operation, are designed to provide a single voltage across an electrode. However, if the geometric spacing between electrodes is not fixed to a single value, the electrical field between the electrodes will exhibit a spatial variation. If this variation is linear where the beam transits the crystal, a good scanner can be produced. A quadrupole field as shown in Figure 12.5 has the desired behavior near the origin.[8,9] If the electrodes are shaped to follow the hyperbolae

$$xy = \pm \frac{R_0^2}{2} \tag{12.23}$$

then the potential in the region between the electrodes is given by

$$V = \frac{V_0}{R_0^2} xy \tag{12.24}$$

The electric field is obtained by taking the gradient of the potential:

$$E_x = -\frac{\partial V}{\partial x} = -\frac{V_0}{R_0^2} y$$

$$E_y = -\frac{\partial V}{\partial y} = -\frac{V_0}{R_0^2} x \tag{12.25}$$
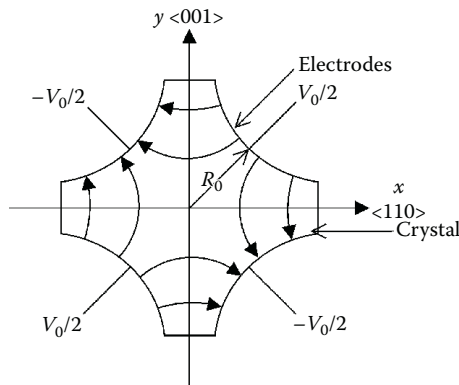
FIGURE 12.5
Geometry for generating a linear electric field profile using quadrupole electrodes in KDP-type materials. Crystallographic directions for proper deflector operation are also shown. Optical beam propagation is perpendicular to the page, and the optical electric field polarization is parallel to the *x*-axis (<110> direction). (From Fowler, V.J.; Buhrer, C.F.; Bloom, L.R. Electro-optic light beam deflector. *Proc. IEEE* 1964, 52(2), 193–194. With permission.)

Thus we see that both components of the electric field vary linearly with position. However, if a crystal with symmetry $\bar{4}2m$ is oriented as indicated in Figure 12.5, then there is no EO effect for E_x, to the first order. Using the index expressions in Table 12.1 for KDP, the index gradient for an optical wave polarized along the x-direction ($\langle 110 \rangle$) becomes

$$\frac{\mathrm{d}n}{\mathrm{d}x} = \frac{n^3 r_{63} V_0}{2 R_0^2} \tag{12.26}$$

The beam displacement and deflection angle at the scanner output can be obtained by substituting Equation 12.26 into Equations 12.1 and 12.3. The results are

$$X(L) = \frac{n^2 r_{63} V_0}{2 R_0^2} \frac{L^2}{2} \tag{12.27}$$

$$q_{\mathrm{def}} = \left(\frac{L n^3 r_{63}}{2 R_0^2} \right) V_0 \tag{12.28}$$

The deflection sensitivity and pivot point are readily found to be

$$\frac{q_{\mathrm{def}}}{V} = \frac{L n^3 r_{63}}{2 R_0^2} \tag{12.29}$$

$$L_{\mathrm{P,in}} = \frac{L}{2} \tag{12.30}$$

It is a difficult task to shape a crystal to accommodate hyperbolic electrodes as shown in Figure 12.5. Instead, Figure 12.6a shows a more practical geometry.[10] The field inside the crystal has been computed using the Finite Element Method[11] for KDP. Although the electrodes are not shaped precisely like hyperbolae, the field near the center of the crystal is still approximately linear, as shown in Figure 12.6b. Another approximation

TABLE 12.1

Eigenvalues and Eigenvectors for KDP with an
Electric Field Applied Along the Optic Axis[a]

Eigenvalue	Eigenvector	Index	Δn
$\dfrac{1}{n_o^2}$ $r_{63} E_3$	$\langle 110 \rangle$	$n_o - \dfrac{\Delta n}{2}$	$n_o^3 r_{63} E_3$
$\dfrac{1}{n_o^2} - r_{63} E_3$	$\langle 1\bar{1}0 \rangle$	$n_o + \dfrac{\Delta n}{2}$	$n_o^3 r_{63} E_3$
$\dfrac{1}{n_e^2}$	$\langle 001 \rangle$		0

[a] For consistency with subsequent sections, Δn is the index change between two regions with oppositely oriented field; thus the index change arising from applying a field with a single polarity to a single domain material is $\Delta n/2$.
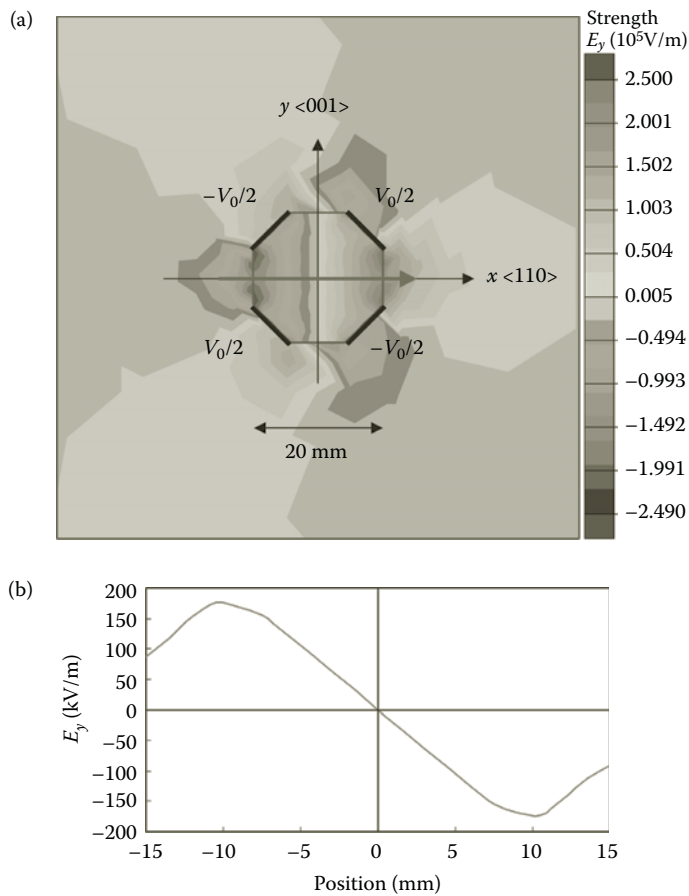
FIGURE 12.6
(a) Finite element analysis[11] of the electric field in a KDP crystal shaped like an octagon to approximate a quadrupolar field. The shading indicates the strength of the y component of the electric field. The vertical stripes near the center indicate that E_y is approximately linear with x and independent of y; (b) Calculation of the electric field E_y along the contour shown in part (a). Note the linearity of the field near the origin.

of hyperbolic electrodes was developed by Ireland and Ley.[12] In this case, cylindrical electrodes were used.

12.3.4.2 Graded Index with Constant Spacing

A second option for producing a linearly graded index of refraction depends on depositing a resistive electrode and a conductive electrode on opposite faces of a slab of EO material, as shown in Figure 12.7. When the voltage is applied at the leads, current flows across the resistive electrode, which is essentially a resistor, resulting in a graded electrical field within the device. The change in optical index, which is proportional to the electric field in the material, is thus graded according to the voltage.

The difficulties of producing this type of electrode have limited their consideration for use to very thin devices, on the order of tens of microns. The optical beams are not typically round in these scanners, since a relatively wide stripe is needed for the voltage gradient
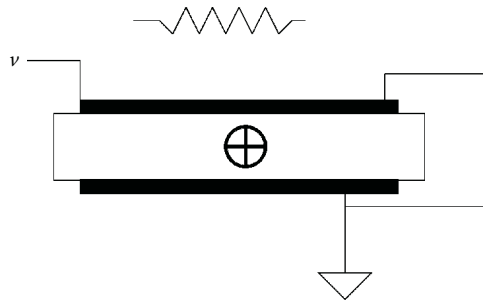
FIGURE 12.7
A resistive electrode may be used to produce graded electrical fields. This is a schematic view, seen end on, of such a device.
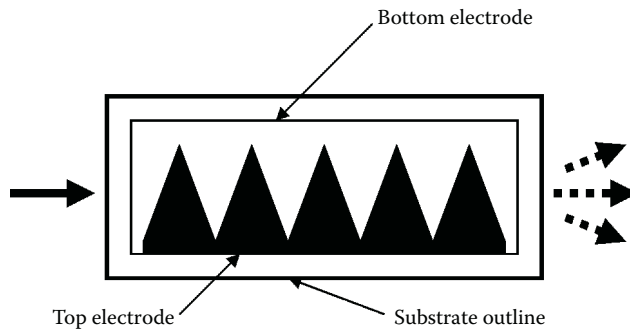
FIGURE 12.8
A serrated electrode, all of which is held at a single voltage, has the effect of producing a graded electrical field over the length of the electro-optical device.

electrodes. Difficulties of coupling optical beams into such wide and thin layers, and their great divergence on the output side add to the difficulties of this approach, although it has been proposed for optical switching for telecommunications use. Electrical heating due to the current flow is another limiting factor.

12.3.4.3 Graded Index with Constant Spacing and Single Voltage

A third approach to producing a gradient in the index of refraction is illustrated in Figure 12.8. The portion of the beam that travels under the root of the conducting top electrode will traverse more material affected by the electrical field than a ray traversing only the tips. This technique suffers from the complex fringing electrical fields around the top electrode. These fields give rise to out-of-plane distortion and other beam quality problems. A wide variety of electrode shapes and spacings are possible,[1] some of which mitigate the effects of fringing fields to a degree.

12.3.5 Poled Structures

There are two ways of achieving an effective linear gradation in index: grading the electric field, or grading the material properties. The techniques covered above, except the

use of multiple discrete inverted prisms, all use device or electrode geometry to effectively grade the electrical field. Grading the material properties is possible using the technique of "poling" or "domain inversion." This process can be performed completely independently of the base material production, making it an effective tool for device fabrication.

One salient advantage of poled devices is that they are not typically affected by electrical fringing fields. They are normally constructed with cover electrodes that are large enough to ensure uniform fields over the active region of the device, which may contain 50 or more interfaces. The result is that the beam will have to pass through only two fringing fields, at the entrance and exit, versus 100 or more in some field-graded devices. The poling fabrication technique is an outgrowth of research directed toward making quasi-phase-matched second harmonic generating gratings.[13–17] It was discovered that by using photolithographic techniques, domain patterns with virtually any shape can be realized in materials such as *z*-cut lithium niobate, lithium tantalate, and potassium titanyl phosphate and its isomorphs such as rubidium titanyl arsenate.

The basic process includes the following steps:

1. Patterning: photolithographic techniques are used to pattern photoresist on one or both of the crystal surfaces with the shape of the final inverted region. Wafers are often used for greatest compatibility with standard semiconductor processing equipment.

2. Apply poling electrodes: either metal or liquid electrodes are placed on the wafer surface.

3. Poling: high electric fields, greater than the coercive field of the medium or on the order of 20 kV/mm for lithium niobate, are applied. Practitioners use voltage pulses, ramps, and quasi-DC waveforms—there does not appear to be an accepted standard practice at this time.

4. Apply operating electrodes: this is often done using standard photolithography and thin film deposition. The operating electrodes are typically much larger than the poling electrodes to ensure uniform electric field in the poled region, but they may be limited in size due to the desire to minimize the device's electrical capacitance and therefore driver power requirements.

5. Annealing: temperatures ranging from about 200 °C to over 1000 °C are used, with the appropriate temperature and cycle profile determined by experimentation.

During the "poling" step, atoms are shifted in the crystal lattice, but only in the volumes defined by the photolithographically applied patterns. This effectively "flips" the ferroelectric domains without having to cut, polish, AR coat, and reassemble the crystal.

Using electric field poling techniques, it is possible to drive the inverted domain regions completely through relatively thick substrates—up to 3 mm have been reported for devices fabricated with profile precision of a few microns in RTA (R. Stolzenberger, personal communication, 2003). For more details on the physics of the domain inversion process, see Gopalan et al.[18] Using such a process, confinement to a waveguide is possible, but not necessary, and scanners that work for any properly polarized beam passing through the substrate can be made (Figure 12.9).[19] For applications not requiring compatibility with waveguide devices, these bulk devices offer easier coupling, lower coupling and propagation losses, and improved beam quality.
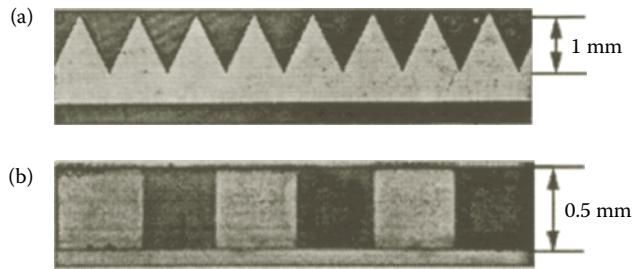
FIGURE 12.9
A bulk electro-optic wafer deflector using patterned domain inversion: (a) patterned Ta electrode used to define the geometry of the deflector; (b) etched Y cross-section of the sample showing domain inversion through the thickness of the wafer. (From Revelli, J.F. High-resolution electrooptic surface prism waveguide deflector: An analysis. *Appl. Optics* 1980, *19,* 389–397. With permission.)
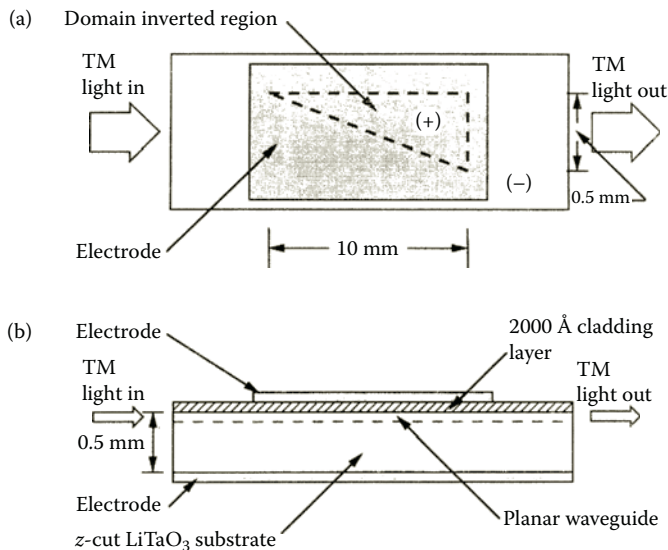
FIGURE 12.10
Geometry of the first electro-optic waveguide deflector using patterned domain inversion: (a) top view of the substrate showing the domain-inverted region; (b) cross section through the prism region. (From Lee, C.L.; Lee, J.F.; Huang, J.Y. Linear phase shift electrodes for the planar electrooptic prism deflector. *Appl. Optics* 1980, 19, 2902–2905. With permission.)

12.3.5.1 Prismatic Poled Structures

The first deflector of this type was a waveguide device fabricated in lithium tantalate, as shown in Figure 12.10.[20] Domain inversion was achieved using patterned proton exchange followed by rapid thermal annealing. This process creates a domain-inverted region extending to a depth of 10–20 μm. The planar waveguide was subsequently formed by proton exchange in 260° pyrophosphoric acid. A 2000 Å thick layer of SiO_2 was deposited as a cladding layer before deposition and patterning of the final cover electrode, to reduce optical loss in the waveguide. Cylindrical lenses were used to edge-couple light into and out of the scanner.

Improved deflection sensitivity can be achieved by using thinner substrates, or by selectively thinning the substrate below the scanner using pulsed laser ablation.[21] Selective

thinning allows the internal field to be increased while maintaining mechanical strength around the border of the substrate.

12.3.5.2 Rectangular Scanners

The simplest scanner geometry is that for which the prisms are enclosed within a rectangular-shaped region. The general case is illustrated in Figure 12.11, where the active region is divided into an arbitrary number of variously shaped prisms. For $\theta_{in} \ll 1$ and $\Delta n \ll n$, the result of applying Snell's law at each interface is a cumulative deflection angle given by

$$q_{in} = \sum_{i=1}^{N} \frac{\Delta n_i}{n} \cot f_i \qquad (12.31)$$

where Δn_i, is the total index change across the ith interface, N is the total number of interfaces in the scanner, and ϕ_i is the angle the ith interface makes with the beam axis. Note that Equation 12.31 is valid regardless of the overall shape of the scanner, and each term in the sum is positive since both Δn_i and ϕ_i change signs from one interface to the next. For rectangular scanners with a fixed $|\Delta n_i| = \Delta n$ at each interface, the scanning angle can be related to the width W and total length L of the device as follows:

$$q_{def} = n q_{in} = \Delta n \sum_{i=1}^{N} |\cot f_i| = \Delta n \sum_{i=1}^{N} \frac{l_i}{W} = \Delta n \frac{L}{W} \qquad (12.32)$$

Note that this is the same result as would be achieved for a scanner with a constant index gradient, from Equation 12.11. We therefore have the somewhat surprising result that the scanning angle does not depend on how many prisms are in the scanner, but only on the ratio L/W! To see why this is, note that as the number of interfaces increases, the angle of incidence becomes closer to normal thereby reducing the refraction at each interface. Consequently, the sum of the effects of all the interfaces is constant. The question of the effects of varying numbers of triangles is considered more closely below from a different point of view.

12.3.5.2.1 Optimum Number of Triangles in Rectangular Scanners

Considering the case of one interface, it is apparent that the scanning properties can, in fact, depend on the number of prisms or triangles. For a sufficiently large value of L/W, total
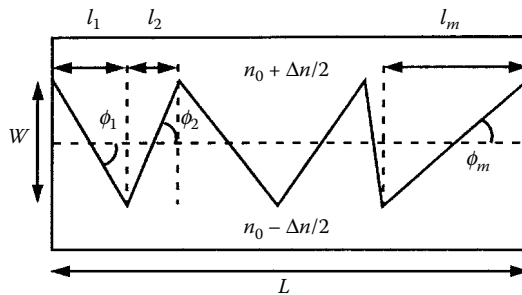
FIGURE 12.11
An arbitrarily divided rectangular prism scanner. (From Sasaki, H.; De La Rue, R.M. Electro-optic multichannel waveguide deflector. *Electronics Letts.* 1977, 13(10), 295–296. With permission.)

internal reflection (TIR) can occur for one voltage polarity, but not the other. Consequently the scanning properties will be strongly asymmetric with respect to drive voltage polarity. On the other hand, if many interfaces are used, the incident angle at each interface will be sufficiently far from grazing that TIR cannot occur for practical values of drive voltage. Chen et al.[20] have considered the scanning asymmetry as a function of the number of interfaces using ray-tracing simulations. As shown in Figure 12.12, the asymmetry in the scanning properties becomes negligible after about 10–15 interfaces.

Another consideration pertaining to the number of interfaces is the Fresnel transmission through the multiple interfaces. For small numbers of interfaces, the reflection at each interface can be high owing to the near grazing incidence (even TIR is possible, as mentioned above). This reflected light is diverted from the optical path, resulting in low transmission through the device. In the opposite limit of very many interfaces, the beam approaches normal incidence at each interface. Thus the reflection at each interface approaches a finite value, while the number of such reflections continues to increase. Thus the total reflection increases for both large and small numbers of interfaces. An optimum number of interfaces can be found that minimizes the total reflection, or equivalently, maximizes the transmission through the device. For $\Delta n/n \ll 1$ and $R \ll 1$, the normalized reflected intensity R is approximately given by

$$R = m \left[1 + \left(\frac{L}{mW} \right)^2 \right]^2 \left(\frac{\Delta n}{2n} \right)^2$$

(12.33)

which has a minimum at the optimum number of interfaces:

$$m_{\mathrm{opt}} = \sqrt{3} \, \frac{L}{W}$$

(12.34)

Interestingly, this condition corresponds to filling the scanner with equilateral triangles. The behavior of the reflected intensity given by Equation 12.33 is illustrated in Figure 12.13.
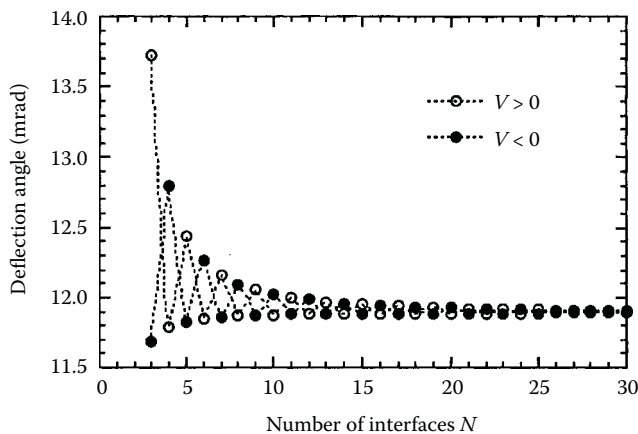
FIGURE 12.12
Symmetry of deflection angle versus the number of interfaces from ray tracing analysis. (From Takizawa, K. Electrooptic Fresnel lens-scanner with an array of channel waveguides. *Appl. Optics* 1983, 22(16), 2468–2473. With permission.)
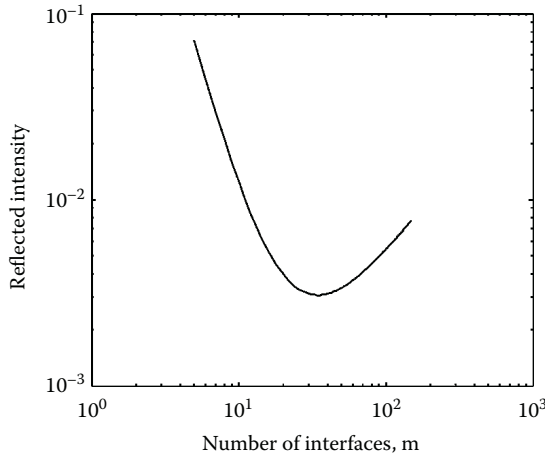
FIGURE 12.13
Normalized reflected intensity as a function of number of interfaces for $\Delta n/n = 10^{-4}$, $L/W = 20$, $m_{\mathrm{opt}} = 35$.

Examination of this figure shows that although an optimum does exist, the reflected intensity is negligible over the practical range of interfaces large enough to satisfy the symmetry condition discussed above. We conclude that as long as the number of interfaces exceeds 10–15, it is satisfactory to assume that the properties of the scanner are independent of the number of interfaces.

12.3.5.2.2 *Deflection Sensitivity for Rectangular Scanners*
To obtain the deflection sensitivity for a rectangular device, we need only to substitute the expression for Δn into Equation 12.32. Using $E_3 = V/h$ in Table 12.2 for crystals of symmetry $3m$, the result is

$$\frac{q_{\mathrm{def}}}{V} = \frac{n_e^3 r_{33}}{h} \frac{L}{W} \tag{12.35}$$

where h is the thickness of the substrate, and the incident beam is polarized along the z-axis of the crystal (extraordinary wave). Bulk scanners will also work if the incident beam is polarized in the plane of a z-cut substrate (ordinary wave), but with reduced deflection. In this case, the deflection sensitivity is given by (Table 12.2):

$$\frac{q_{\mathrm{def}}}{V} = \frac{n_o^3 r_{33}}{h} \frac{L}{W} \tag{12.36}$$

It should be noted that Equations 12.35 and 12.36 are applicable to bulk devices; for waveguide devices, the voltage drop across the cladding layer must also be taken into account. This will result in a slight decrease in deflection sensitivity.

12.3.5.2.3 *Pivot Point Location for Rectangular Scanners*
The internal pivot point for rectangular scanners can be obtained directly from Equations 12.13 and 12.14:

$$L_{\mathrm{P,in}} = \frac{L}{2} \tag{12.37}$$

TABLE 12.2

Eigenvalues and Eigenvectors for $LiNbO_3$ and $LiTaO_3$
with an Electric Field Applied along the Optic Axis[a]

Eigenvalue	Eigenvector	Index	Δn
$\dfrac{1}{n_o^2}$ $r_{13}E_3$	<100>	$n_o - \dfrac{\Delta n}{2}$	$n_o^3 r_{13}E_3$
$\dfrac{1}{n_o^2}$ $r_{23}E_3$	<010>	$n_o - \dfrac{\Delta n}{2}$	$n_o^3 r_{23}E_3$
$\dfrac{1}{n_e^2}$ $r_{33}E_3$	<001>	$n_e - \dfrac{\Delta n}{2}$	$n_e^3 r_{33}E_3$

[a] Δn is defined as in Table 12.1. Note that for these crystals
$r_{13} = r_{23}$.
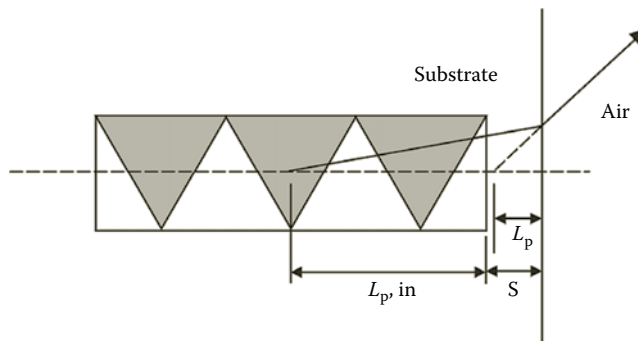
FIGURE 12.14
Shift of the apparent location of the pivot point by substrate refraction, including the effect of a spacing s
between the output of the scanner and the substrate edge.

In practical geometries, there is usually a spacing s between the output plane of the scan-
ner and the edge of the crystal, as shown in Figure 12.14. This spacing provides a lon-
ger electrical creepage path around the end of the scanner, and allows for cutting and
polishing operations to be performed without impinging on the active scanner area. In
this case, the location of the pivot point as viewed from outside of the crystal is given by
(cf. Equation 12.15)

$$L_P = \frac{1}{n}\left[L_{P,in} + s\right] \tag{12.38}$$

12.3.5.3 Trapezoidal Scanners

One difficulty with the design of rectangular scanners results from the beam displacement
within the scanner, as given by Equations 12.12 and 12.13. Clearly the width of the scanner
must be increased to accommodate this displacement, but increasing the width reduces
the deflection sensitivity. However, because of the shape of the trajectory, this increase is
only needed at the output of the scanner. As discussed earlier, one way to address this is
to focus the beam through the scanner so that the reduction in output beam diameter is
comparable to the displacement. However, we saw that this technique reduces the number

of resolvable spots. Another possibility is to increase the width of the output forming a trapezoidal shape (Figure 12.15).

12.3.5.3.1 Deflection Sensitivity of Trapezoidal Scanners

The deflection angle for a trapezoidal scanner is given by[5]

$$q_{\text{in}} = \frac{\Delta n}{n} \frac{L}{W_1 - W_0} \ln\left(\frac{W_1}{W_0}\right) \tag{12.39}$$

Re-expressing this in terms of the external deflection angle and substituting for the change in refractive index gives the deflection sensitivity

$$\frac{q_{\text{def}}}{V} = \frac{n_e^3 r_{33}}{h} \frac{L}{W_1 - W_0} \ln\left(\frac{W_1}{W_0}\right) \tag{12.40}$$

12.3.5.3.2 Pivot Point Location for Trapezoidal Scanners

The displacement of the beam at the output of a trapezoidal scanner is given by[5]

$$X(L) = \frac{\Delta n}{n} \frac{L}{W_1 - W_0} \left[\frac{W_1}{W_1 - W_0} \ln\left(\frac{W_1}{W_0}\right) - 1\right] L \tag{12.41}$$

The internal pivot point is obtained using the definition (14) along with Equations 12.39 and 12.41. The result is

$$L_{\text{P,in}} = \left[\frac{W_1}{W_1 - W_0} - \frac{1}{\ln\left(W_1 / W_0\right)}\right] L \tag{12.42}$$

The pivot point as viewed externally can be computed from Equation 12.38, as before.
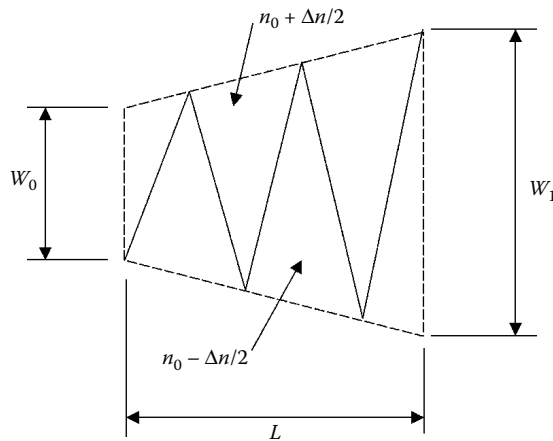
FIGURE 12.15
Geometry of a trapezoidal prism scanner. (From Chiu, Y.; Zou, J.; Stancil, D.D.; Schlesinger, T.E. Shape-optimized electrooptic beam scanners: Analysis, design, and simulation. J. Lightwave Technol. 1999, 17(1), 108–114. With permission.)

12.3.5.3.3 Comparison of Trapezoidal and Rectangular Scanners

The increase in deflection sensitivity compared to rectangular scanners can be illustrated by considering rectangular and trapezoidal scanners with the same lengths, and the width of the rectangular scanner equal to the average of the input and output widths of the trapezoidal device. Thus, if W_R is the width of the rectangular scanner and W_0, W_1 are the input and output widths of the trapezoidal scanner, we require

$$W_R = \frac{W_0 + W_1}{2}$$

(12.43)

Assuming the same maximum index difference Δn Figure 12.16 shows the improvement gained by the trapezoidal geometry. The improvement is modest (<10%) for $W_0/W_R > 0.5$. Since the usual case is that the beam diameter is much larger than the output displacement, W_0/W_1 and W_0/W_R are normally only slightly smaller than unity, yielding a typical improvement of a few percent.

12.3.5.4 Horn-Shaped Scanners

Since the scanning sensitivity is diminished as the width increases, the optimal solution is to increase the width gradually so as to track the beam trajectory. If the change in beam diameter through the scanner is neglected, the shape of the scanner can be obtained by simply adding a constant offset to the beam displacement, as shown in Figure 12.17a. The width of the scanner can therefore be written in the form

$$W(z) = W_0 + 2X_{max}(z)$$

(12.44)

where $X_{max}(z)$ is the displacement at position z with maximum voltage applied. The factor of 2 accommodates bipolar operation of the scanner. The general shape $W(z)$ has been obtained by Chiu et al.,[5] and is plotted in terms of normalized coordinates in Figure 12.17b.
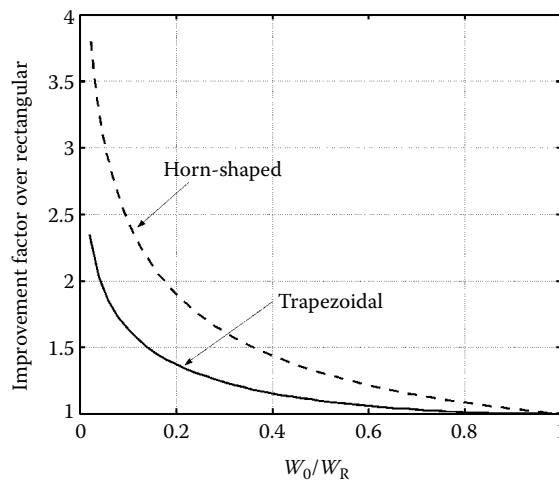
FIGURE 12.16
Comparison of scanning sensitivity of rectangular, trapezoidal, and horn-shaped scanners. (From Chiu, Y.; Zou, J.; Stancil, D.D.; Schlesinger, T.E. Shape-optimized electrooptic beam scanners: Analysis, design, and simulation. *J. Lightwave Technol.* 1999, 17(1), 108–114. With permission.)
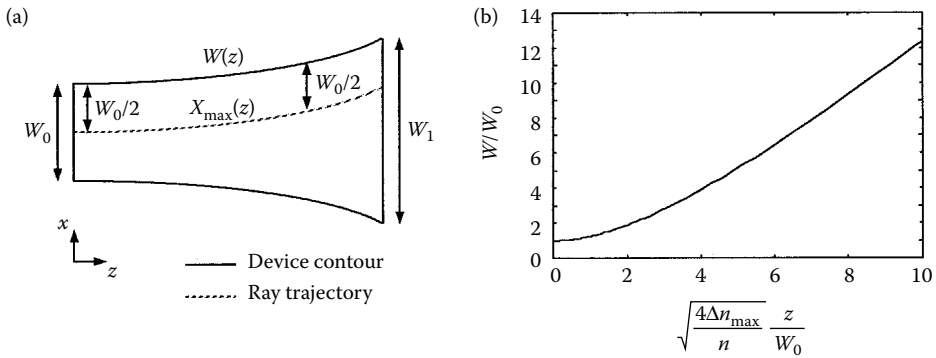
FIGURE 12.17
Normalized shape-optimized scanner design curve: (a) reference geometry and (b) normalized shape contour (from Chiu, Y.; Zou, J.; Stancil, D.D.; Schlesinger, T.E. Shape-optimized electrooptic beam scanners: Analysis, design, and simulation. *J. Lightwave Technol.* 1999, 17(1), 108–114. With permission.) Tabulated values are given in Table 12.3.

TABLE 12.3

Tabulated Values for the Normalized Horn-Shaped Scanner Curve Shown in Figure 12.17[a]

Z^*	W^*	Z^*	W^*	Z^*	W^*	Z^*	W^*
0.00	1.00	2.50	2.49	5.00	5.36	7.50	8.84
0.25	1.05	2.75	2.73	5.25	5.68	7.75	9.21
0.50	1.11	3.00	2.99	5.50	6.01	8.00	9.58
0.75	1.21	3.25	3.26	5.75	6.35	8.25	9.96
1.00	1.34	3.50	3.53	6.00	6.69	8.50	10.34
1.25	1.48	3.75	3.82	6.25	7.04	8.75	10.73
1.50	1.65	4.00	4.11	6.50	7.39	9.00	11.11
1.75	1.83	4.25	4.41	6.75	7.75	9.25	11.50
2.00	2.04	4.50	4.72	7.00	8.11	9.50	11.89
2.25	2.26	4.75	5.03	7.25	8.47	9.75	12.29
						10.00	12.68

$$Z^* = \frac{z}{W_0}\sqrt{\frac{4\Delta n_{\max}}{n}} \quad W^* = W/W_0.$$

[a] These values can be used for designing masks for scanner fabrication.

To facilitate use of this curve in designs, values of the normalized width as a function of z are tabulated in Table 12.3.

The shape of the optimal curve allowing for the change in beam diameter through the scanner depends on the Gaussian beam parameters. Figure 12.18 shows a simulation of an optimal scanner with the beam focused on the output plane.[22] The simulation was performed using the beam propagation method (BPM).[23,24]

12.3.5.4.1 Deflection Sensitivity of Horn-Shaped Scanners

The deflection angle as a function of Δn for an optimum horn-shaped scanner is

$$q_{\text{in}} = \frac{\Delta n}{n}\sqrt{\frac{n}{\Delta n_{\max}}\ln\left(\frac{W_1}{W_0}\right)} \tag{12.45}$$
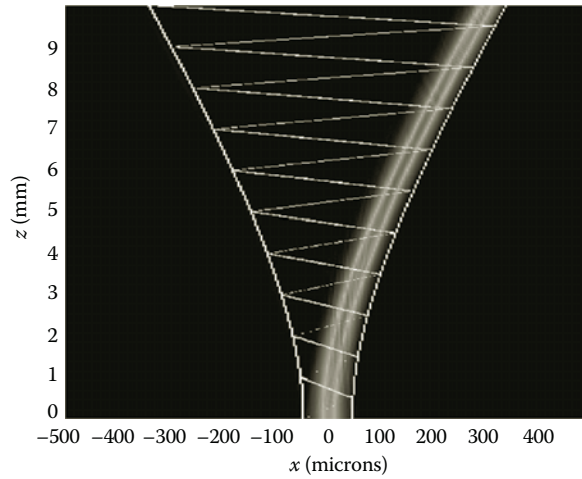
FIGURE 12.18
Simulation of horn-shaped scanner operation using the beam propagation method (BPM). Parameters are $W_0 = 92$ μm, $W_1 = 678$ μm, $L = 10$ mm, $\lambda_0 = 0.6328$ mm, $n_e = 2.1807$ (lithium tantalate), and $\Delta n = 2.1 \times 10^{-3}$. The radius of the beam waist is 30 μm and is focused at the output of the scanner. (From Fang, J.C.; Kawas, M.J.; Zou, J.; Gopalan, V.; Schlesinger, T.E.; Stancil, D.D. Shape-optimized electrooptic beam scanners: experiment. *IEEE Photonics Technol. Lett.* 1999, 11(1), 66–68. With permission.)

with a maximum value of

$$q_{\text{in,max}} = \sqrt{\frac{\Delta n_{\max}}{n} \ln\left(\frac{W_1}{W_0}\right)} \tag{12.46}$$

where $\Delta n_{\max}$ is the index variation across the scanner with the maximum applied voltage to be used. The deflection sensitivity is obtained by multiplying Equation 12.45 by n and substituting for Δn from Table 12.1. The result for crystals with symmetry $3m$ (e.g., lithium niobate) is

$$\frac{q_{\text{def}}}{V} = n_e^2 \sqrt{\frac{r_{33}}{hV_{\max}} \ln\left(\frac{W_1}{W_0}\right)} \tag{12.47}$$

12.3.5.4.2 Pivot Point Location of Horn-Shaped Scanners
The unipolar displacement at the output of the horn scanner is

$$X(L) = \frac{\Delta n}{\Delta n_{\max}} \frac{W_0}{2} \left[\frac{W_1}{W_0} - 1\right] \tag{12.48}$$

Taking the ratio of Equations 12.48 and 12.45 gives the pivot point:

$$L_{\text{P,in}} = \frac{(W_0/2)\left[(W_1/W_0) - 1\right]}{\sqrt{(\Delta n_{\max}/n)\ln(W_1/W_0)}} \tag{12.49}$$

Remarkably, the fact that Δn drops out of this equation means that a well-defined pivot point exists (i.e., $L_{P,in}$ does not depend on voltage), even in such a complex horn-shaped device.

12.3.5.4.3 *Comparison of Horn-Shaped Scanners with Trapezoidal and Rectangular Scanners*

The deflection sensitivity of a horn-shaped scanner with the same input width, output width, and length as a trapezoidal scanner is shown in Figure 12.16, normalized to that of a rectangular scanner with the same average width. The improvement over trapezoidal and rectangular scanners is clear.

A summary of the design equations for rectangular, trapezoidal, and horn-shaped scanners is presented in Table 12.4.

12.3.5.5 *Domain Inverted Total Internal Reflection Deflectors*

Domain inversion in ferroelectric crystals can also be used to create relatively long, straight interfaces within a bulk crystal. Since the domains are antiparallel across the interface, an applied electric field will result in a step change in the index of refraction. If a light beam intersects this interface at very high angles (i.e., near grazing incidence) a state of TIR can result (Figure 12.19).

Eason and coworkers report an analysis of such a device,[25] with an application to telecommunications switching in mind. Such a device would typically be operated in a digital fashion, ON or OFF, in order to steer a collimated beam into one or another optical fiber (with appropriate collection optics on the end of each fiber). Compared to a scanner composed of many triangles, such a TIR device will exhibit greater deflection for the same voltage and device size. This gives some advantage by shortening the package length since the two beams separate faster providing more room for the output fibers and their collimation optics. A further advantage is that TIR device performance may be engineered to exhibit nearly polarization independent performance, while triangular prisms are strongly polarization dependent.

12.3.5.6 *Domain Inverted Grating Structures*

The component based upon patterned domains that has received the most attention in recent years is the quasi-phase-matched grating for second harmonic generation (SHG).[13–17] The frequency shifting effect occurs in a "periodically poled" region in the crystal, which is produced by forming many precisely spaced parallel domains in materials such as lithium niobate or KTP.[17] For instance, some early efforts were focused on producing blue light from IR laser diodes for use in data storage applications.

In SHG applications, the light propagates perpendicular to the poled stripes, and generally there is no applied electric field. By rotating the incident beam to near grazing incidence and allowing for the application of an electric field, an EO Bragg grating is produced. This is analogous to an AO Bragg deflector where the electric field rather than the acoustic intensity controls the distribution of the light into the various orders.

Gnewuch and coworkers have designed and tested such a device operating at 633 nm.[26] Similar to the TIR device above, the incident light can be switched to two different positions. The advantage of this structure, however, is that it operates at approximately 25 V, versus 500 V for a prismatic domain poled device operating with a similarly sized optical beam. This offers the potential for enormous electrical power savings during

TABLE 12.4

Summary of Design Formulas for Rectangular, Trapezoidal, and Horn-Shaped Scanners

Scanner type	Geometry	Deflection θ_{def}	Output beam displacement $X(L)$	Pivot point location $L_{P,in}$
Rectangular		$q_{def} = \Delta n \dfrac{L}{W}$	$X = \dfrac{1\Delta n}{2n}\dfrac{L^2}{W}$	$L_{P,in} = L/2$
Trapezoidal		$q_{def} = \Delta n \dfrac{L}{W_1 - W_0}\ln\left(\dfrac{W_1}{W_0}\right)$	$X(L) = \dfrac{\Delta n}{n}\dfrac{L}{W_1 - W_0}$ $\times\left[\dfrac{W_1}{W_1 - W_0}\ln\left(\dfrac{W_1}{W_0}\right) - 1\right]L$	$L_{P,in} = \left[\dfrac{W_1}{W_1 - W_0} - \dfrac{1}{\ln(W_1/W_0)}\right]L$
Horn-shaped		$q_{def} = \Delta n \sqrt{\dfrac{n}{\Delta n_{max}}}\,\ln\left(\dfrac{W_1}{W_0}\right)$	$X(L) = \dfrac{\Delta n}{\Delta n_{max}}\dfrac{W_0}{2}\left[\dfrac{W_1}{W_0} - 1\right]$	$L_{P,in} = \dfrac{(W_0/2)[(W_1/W_0) - 1]}{\sqrt{(\Delta n_{max}/n)}\,\ln(W_1/W_0)}$
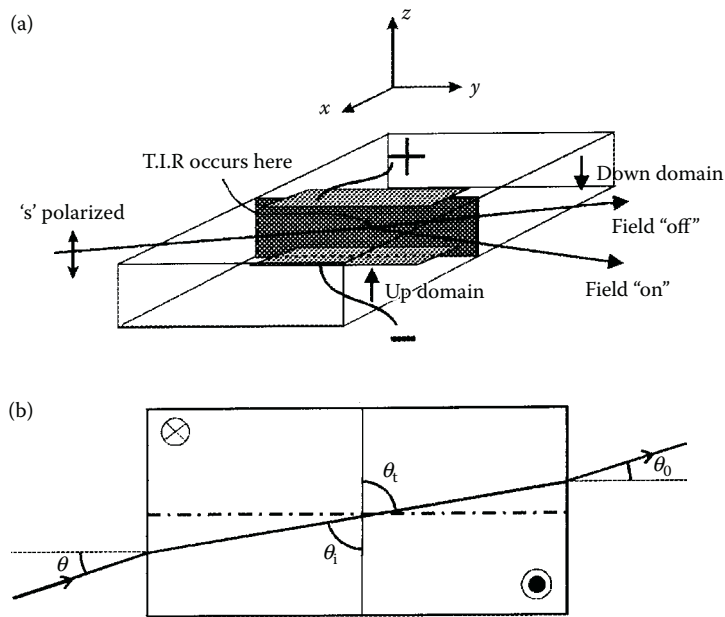
FIGURE 12.19
(a) Schematic of domain engineered total internal reflection (TIR) deflector; (b) plan view of scanner showing grazing angle of the input optical beam to the poled interface. (From Eason, R.; Boyland, A.; Mailis, S.; Smith, P.G.R. Electro-optically controlled beam deflection for grazing incidence geometry on a domain-engineered interface in LiNbO$_3$. *Optics Commun.* 2001, 197, 201–207. With permission.)

high-frequency operation. Without acoustic velocity issues, the EO version of the Bragg deflector can operate at high speeds and handle higher beam powers than an AO version could.

12.3.5.7 Other Poled Structures

The power of patterned ferroelectric domain inversion is that a wide variety of optical components can be fabricated by designing masks with appropriate shapes. For example, although patterning the optimal horn-shape photolithographically is straightforward, the fabrication of a horn-shaped scanner by assembling discrete prisms would be prohibitively difficult. Further, with patterned domain inversion multiple components can be easily integrated on a common substrate.

As another example, EO cylindrical lenses can be formed by patterning oval, semicircular, or circular inverted domains.[27,28] Stacks of such lenses are readily combined with a prism scanner.[29] An optical integrated system of this type could be used, for example, to collimate the light from a fiber before entering the scanner. It should be noted, however, that lenses made in this way are cylindrical lenses, and hence only focus in the plane of the substrate. This is of no consequence in planar waveguide devices, but must be considered if bulk operation is desired.

A periodically poled structure can also be integrated with prism scanners to realize a device capable of generating and steering blue light.[30] The SHG conversion efficiency of this bulk integrated device was relatively low, however, since the optical intensity in
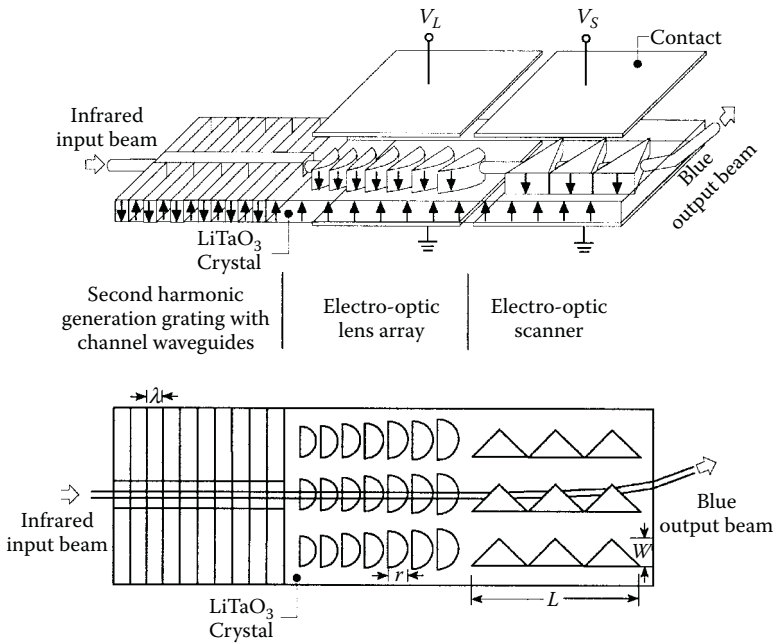
FIGURE 12.20
Schematic of an integrated device containing a channel waveguide with an SHG grating, collimating lens stack, and scanner. (From Chiu, Y.; Gopalan, V.; Kawas, M.J.; Schlesinger, T.E.; Stancil, D.D.; Risk, W.P. Integrated optical device with second-harmonic generator, electrooptic lens, and electrooptic scanner in LiTaO$_3$. *J. Lightwave Technol.* 1999, 17(3), 462–465. With permission.)
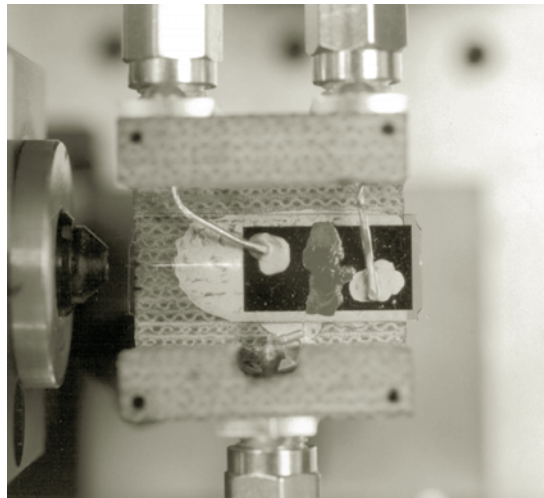
FIGURE 12.21
Photograph of the device described in Figure 12.20. Blue light generated by the SHG waveguide can be seen on the left, followed by the lens and scanner electrodes.

the grating was limited by the focused beam diameter. To achieve the highest conversion efficiency, the light must be confined to a channel waveguide.

The issue of low conversion efficiency encountered with the first SHG scanner device can be addressed by combining the above elements to form an integrated SHG grating, lens stack, and scanner, as shown in Figures 12.20 and 12.21.[31] The channel waveguide in the grating keeps the optical intensity high to maximize the nonlinear frequency doubling efficiency. The output of the channel waveguide then opens into a planar waveguide, and the beam subsequently diverges. A lens stack is next used to collimate the beam before entering the scanner.

12.4 ELECTRONIC DRIVERS FOR ELECTRO-OPTIC DEFLECTORS

12.4.1 Overview

A primary consideration when applying EO deflector devices as part of a system, and an issue often raised by potential users, is the electronic driver. The driver is essentially an amplifier, converting a low-voltage control signal to the higher voltage required to achieve the desired optical index change. It is usually custom-designed for each application, requires specialized circuit design skills and may present safety hazards. Depending on the specific system requirements, the design challenges can vary; however, the high-voltage (HV) nature of these devices can be the single most restricting factor, as component availability can restrict the design space. Also, the HV supplies required to power the driver(s) can be a considerable cost and design challenge.

Other parameters that can influence the design are: switching speeds; repetition rate; switch/component package density requirements; electromagnetic interference (EMI) and radio frequency interference (RFI) considerations; switching power efficiency; and thermal considerations.

Significant strides in recent years in power field effect transistor (FET) and insulated gate dipolar transistor (IGBT) technology have alleviated many practical constraints and enabled power densities and performance that were not possible in earlier years. These devices, developed by Motorola, IXYS, International Rectifier, Toshiba, ST Microelectronics, and others, primarily for the motor drive/control industry, can be well suited for EO scanner drive applications. Also, very fast HV diodes by these same manufacturers and surface mount HV capacitors from Johansson Dielectric and others have driven achievable performance up considerably.

12.4.2 High-Voltage Power Supplies

Most EO scanner drivers rely on a constant high voltage supply as a subsystem. For instance, a modulator system can be realized by switching one electrode of an EO element between a high voltage and ground while the other electrode is held at ground. Benchtop HV power supplies are readily available from several manufacturers including Spellman, Ultravolt, Trek, and others. The market for application-specific HV supplies is not large and, as such, cost can often be quite high and standard packaging options somewhat limited. In many cases it is desirable to design a custom HV supply that suits the application.

There are several design topologies available for HV power supply design. Whether the input power is AC line voltage or a DC supply, a voltage boost converter of some type

needs to be employed. A typical boost converter is discussed and then some higher efficient topologies covered.

12.4.2.1 Conventional Boost Converters

Switching supply topologies, such as boost converters, can be utilized when a high voltage needs to be generated from a much lower one. Typical boost converters consist of a switching transistor (Q1), usually a FET, with an inductor (L1) connected between the drain and the low voltage supply. Current is transferred discontinuously at the switching frequency, and the stored energy amount and pulse duration are proportional to the output voltage feedback signal (Figure 12.22).

The load in a boost converter is usually fed through a rectifying diode. The current in the inductor $I_{L1}(\text{pk}) = \left(V_{dc}{}^*t_{on}\right)/L1$, ramps up linearly during the ON cycle of the FET. The energy stored is $E = 1/(2 * L1 * I_{L1}^2(pk))$. When the FET is turned off that energy is then transferred to the load via the rectifying diode. A portion of V_{out} is fed back through a pulse width modulation (PWM) converter to control the desired pulse width of the FET drive. Boost converters are typically only used in lower power applications of less than 10 W.

12.4.2.2 Flyback Converters

In many HV applications with moderate to high power requirements, the size of the inductor in the boost converter needed to store the proper amount of energy becomes unwieldy, and losses become high. In these cases, a transformer can replace the boost converter inductor. This topology is referred to as a flyback converter. A schematic example is shown in Figure 12.23.

The basic operation of a flyback converter is that when the FET is turned ON, current ramps up at a rate $di/dt = (V_{dc})/L_{pm}$, where L_{pm} is the magnetizing inductance of the primary winding of transformer T1. When the FET is subsequently turned off the current has
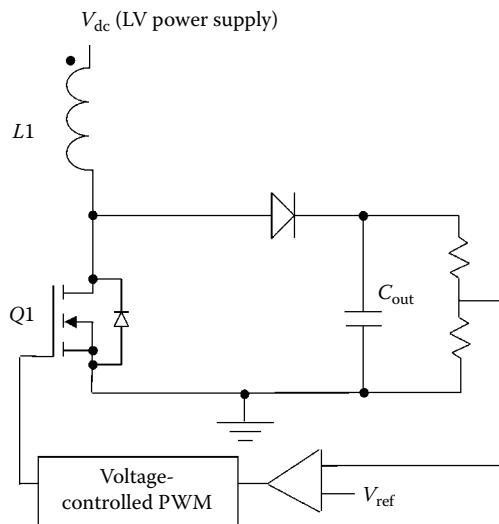
FIGURE 12.22
Schematic of conventional boost converter. The driver circuit for the EO element is connected across the output storage capacitor, C_{out}. This type of high voltage supply is typically used in applications requiring less than 10 W of output power.
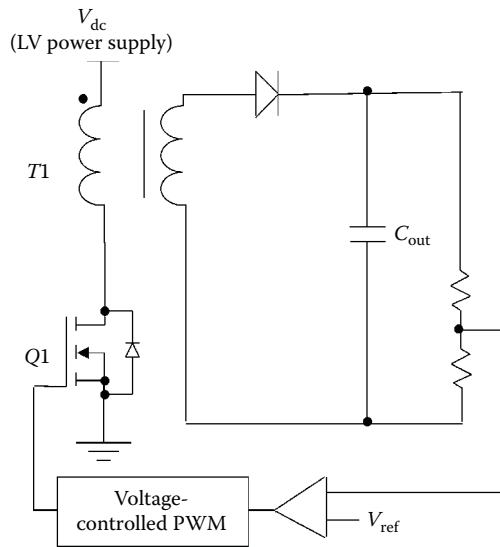
FIGURE 12.23
Schematic of a flyback converter.

ramped to the $I_{pk} = (V_{dc})^* T_{on}/L_{pm}$ thus storing the energy $E = L_{pm}{}^*(I_{pk})^2 2$. With the FET turned OFF, the magnetizing inductance causes an instantaneous reversal in polarities of all windings' voltages and the primary current transfers to the secondary as $I_s = I_{pk}(N_p/N_m)$ where N_p and N_m are the primary and secondary winding count. Using a higher voltage primary V_{dc} can help minimize transformer size and keep the I_{pk} to manageable levels. Also, one can employ multiple winding outputs to increase output voltage as needed or to select one of multiple output levels.

The flyback converter topology can be used for 5–150 W supplies for $V_{out} \leq 5000$ V. A limiting factor can be that the required current in the primary of the flyback transformer is not excessive while keeping within an acceptable, and realistic, transformer size. Efficiencies of the order of 85% and above are achievable.

12.4.3 Digital Drivers

Many EO systems such as telecommunications switches and display system optical beam modulators, can be realized using drivers with only two output states. The term "digital driver" is used here, although the voltages being considered can range to 1000 V and above, requiring far different components than digital logic circuits.

12.4.3.1 Simple Totem Pole Circuits

The simplest and highest speed driver for controlling a capacitive EO device is a "totem pole" or "half bridge" arrangement of FETs where the totem pole is connected across a HV power supply and the device is connected between the FETs to ground (Figure 12.24). Each FET is turned ON or OFF in turn to set the voltage across the output (load) EO device to either the high or low state. Both FETs are never turned ON at the same time—this would lead to a high current directly from the high voltage supply and would likely destroy the FETs unless proper limiting circuitry is used.
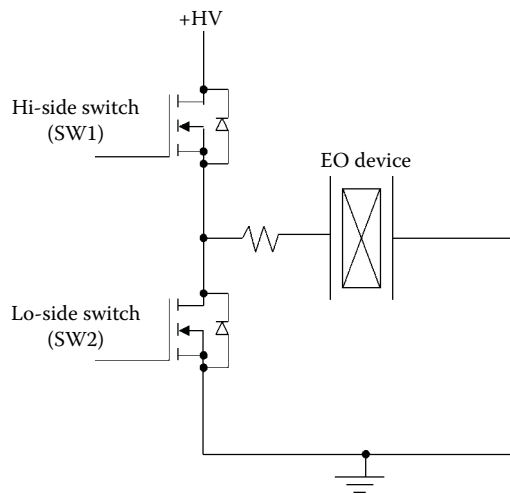
FIGURE 12.24
Schematic of a totem pole driver. The +HV lead of this driver would be attached to the high side of C_{out} of a high-voltage power supply.

In some cases, the FETs would have resistors in series with them to dissipate some of the charge/discharge energy (without the resistors, nearly all of the energy is dissipated in the FETs during the switching transition, which can cause substantial heating). This will, of course, slow down the charging and discharging edges slightly. This straightforward totem pole results in a very high speed, although very lossy, method to drive the output capacitor, in this case, an EO scanner.

The high-side drive can be easily crafted from a p-channel FET where HV < 200 V. However, many practical applications can require a higher voltage, and such FETs are not available. As such, n-channel FETs are recommended for most applications and can be utilized with a floating high-side gate drive. A simple implementation of this would be a transformer-coupled FET gate drive circuit.

It should be noted that this high-side gate driver must be crafted such that in the "dwell ON state," the gate drive must be able to keep the FET ON for the longest system dwell time, which may be an undetermined duration. If continuous (infinite) dwell is required then a refresh circuit should be added to the transformer-coupled FET gate driver circuit.

Gate drivers for the FETs can be crafted from discrete elements—many of those that were designed for the motor drive industry can be utilized. The main FETs must be selected for low junction capacitances rather than minimal R_{ds}(on) as a primary constraint as these capacitances must be charged and discharged during the switching cycle and can be comparable to the load capacitance of the EO device itself and contribute significantly to circuit losses.

Devices such as the highly integrated totem pole driver ICs from International Rectifier (IR2213), ST Microelectronics (L6285), and others can be utilized for applications up to 1000 V. Beyond that, IGBTs can be employed. These devices have T_{on}/T_{off} propagation delays and these parameters can drift significantly with temperature such that timing control circuitry must take it into account. Switching speeds on the order of 100 V/ns are possible.

The maximum voltage rating of most devices in the circuit can be cut in half, thus greatly increasing the component choices and safety factors, by utilizing a full bridge drive where two totem poles are set up to switch $1/2$ HV each to the load. Efficiency benefits can also be realized as the loss from the two charge cycles can be less than one-half that of switching the

entire HV in one step. Further, this topology enables "adiabatic" switching to be utilized such that further efficiency gains can be realized—this concept is covered in Section 12.4.3.2.

12.4.3.2 Adiabatic Drivers

Inherent in traditional switching logic design is the $CV^2/2$ of energy that is dissipated every time a transistor is turned on to charge or discharge a capacitive load. This dissipation is a direct consequence of the fact that, in traditional switching logic configurations, charge for the load is taken from a power or ground rail and that the device to be charged initially sits at a fixed potential very different from that of the rail.

In a simple totem pole application of charging and discharging a load C to a voltage *V*, the energy dissipated to flip the output is $E = \frac{1}{2}CV^2$. This energy does not depend on the needed time to switch, nor the clock rate, but is strictly related to the energy transfer process. In fact, during the rising transition, the power supply delivers all the charge $Q = CV$ at voltage *V*, while during the falling transition that charge is returned at zero voltage. So, actually, the energy $E = CV^2$ is drawn from the HV supply, with half of the energy being stored in the load capacitor, and the other half being dissipated in losses.

Put another way, half of this energy is dissipated by the FET in the pull-up network (rising transition) and the other half by the pull-down network (falling transition), independent of how fast the transitions are. To reduce the dissipated energy, only methods that reduce the load capacitances or the supply voltage can be applied, but in any case they are strictly limited by the load *C* and voltage *V*.

The term "adiabatic process" is most often applied to the thermodynamic cycle by which a gas, such as air, is heated or cooled by expansion or compression without an external source or sink of heat.

The energy transfer processes in the electronic driver can be done adiabatically if, during the rising (or falling) transition, the power supply delivers (or recovers) the charge to (or from) the load at a potential close to that of the source/supply potential. In other words, in order to implement adiabatic processes in a switching circuit, the switching devices should be turned on only when the source-drain voltage is zero, and source-drain voltage should be changed only while the device is off; and, if possible, given the desired performance of the circuit, any voltage change must be done as gradually as possible.

There are difficulties in implementing this solution, however. First, the logic must be designed so that switching transitions can occur only at suitable times (that is, only when there is no potential drop across the switching devices). Choreographing this timing can add considerable complexity as there are switching delays inherent in the FETs and gate driver elements. FET drivers and control circuitry also have temperature-dependent characteristics. Secondly, zero energy dissipation only occurs with arbitrarily slow switching: with realistic switching rates, the energy savings might not be enough to make up for the additional complexity. Lastly, adiabatic design relies on the assumption that one can efficiently provide the moving supply (in fact a clock) to the circuit that it drives. This last characteristic is not achievable with HV applications and is more suited for logic level designs.

In Figure 12.25, note that the two half bridges are configured such that the "lo-side" –HV switch is referenced to the –HV rail. This is fairly easily implemented as will be discussed later.

The full bridge circuit operates adiabatically as follows. Start with the crystal discharged, that is PH and NL are off (or open) while PL and NH are turned on. To charge the crystal, one of the grounded FETs, PL is turned OFF—this is adiabatic, in that it is
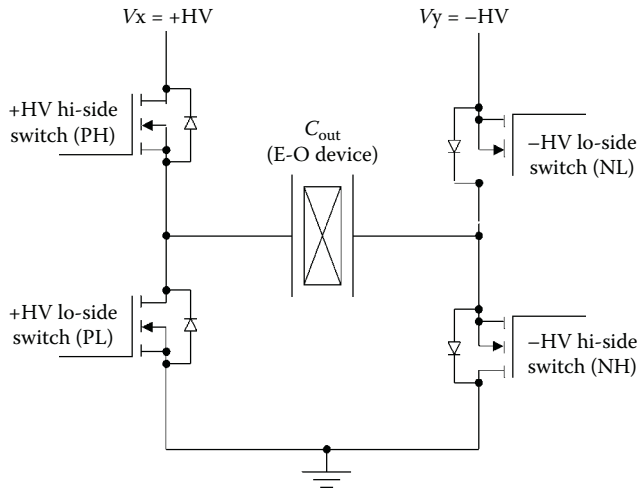
FIGURE 12.25
A dual totem pole driver. By using both a positive and a negative high-voltage supply, the absolute voltage level may be halved compared to a simple totem pole circuit. This can simplify some design tasks. This type of circuit can also be controlled to operate "adiabatically," which offers significant power savings.

done while no current is flowing ($V_{ds} \sim 0$ V). Only a very short period of time is required for it to fully turn off, then immediately turn ON switch PH. Obviously PL must be fully OFF before PH can begin to turn ON to avoid a catastrophic shoot-through condition HV to GND. Note the basic totem pole circuit always switched with V_{ds} at $\sim$ HV, leading to losses in the FET.

This switch transition will result in the following: the +HV device, PH, will need to supply energy to charge the load and the parasitic capacitance of PL, C_{PL}, which includes the capacitance of the protection diode, which is usually a significant component of the total load. Thus the energy required from the +HV supply is $(C_{load} + C_{PL})$ (HV)^2. Since each HV supply is one-half the voltage of what it was in the simple totem pole arrangement, this translates to $(C_{load} + C_{PL})$HV^2/4 of energy drawn so far.

The next step, to fully charge the crystal, is to turn NH OFF (again while $V_{ds} \sim 0$ V) and turn on NL. As before the –HV supply needs to charge up $C_{load} + C_{NH}$. The required energy thus to turn ON the C_{load} is:

$$E = (C_{load})(+HV)^2 + (C_{load})(-HV)^2 \tag{12.50}$$

Since each HV in this case is one-half the voltage compared to the simple totem pole case, the energy is (C_{load})(HV/4) + (C_{load})(HV/4) or (C_{load})(HV^2/2), a saving of approximately half of the energy (minus that to charge the stray and parasitic capacitances of the additional switching device) for each cycle.

For the adiabatic discharge, NL is turned OFF, then NH is turned ON. This will force the recovery of half the charge in the crystal to go back to +HV supply (or C_{load}*HV^2/4 of energy is recovered). Also, the parasitic capacitance of Cp(NL) will be dissipated and Cp(NH) will be charged to Vy, requiring Cp(NH)*Vy*2 energy from Vy supply. This total energy is equivalent to Cp(NH)Vy*2/4. Finally, to completely discharge the crystal, we need to turn OFF PH and turn ON switch PL.
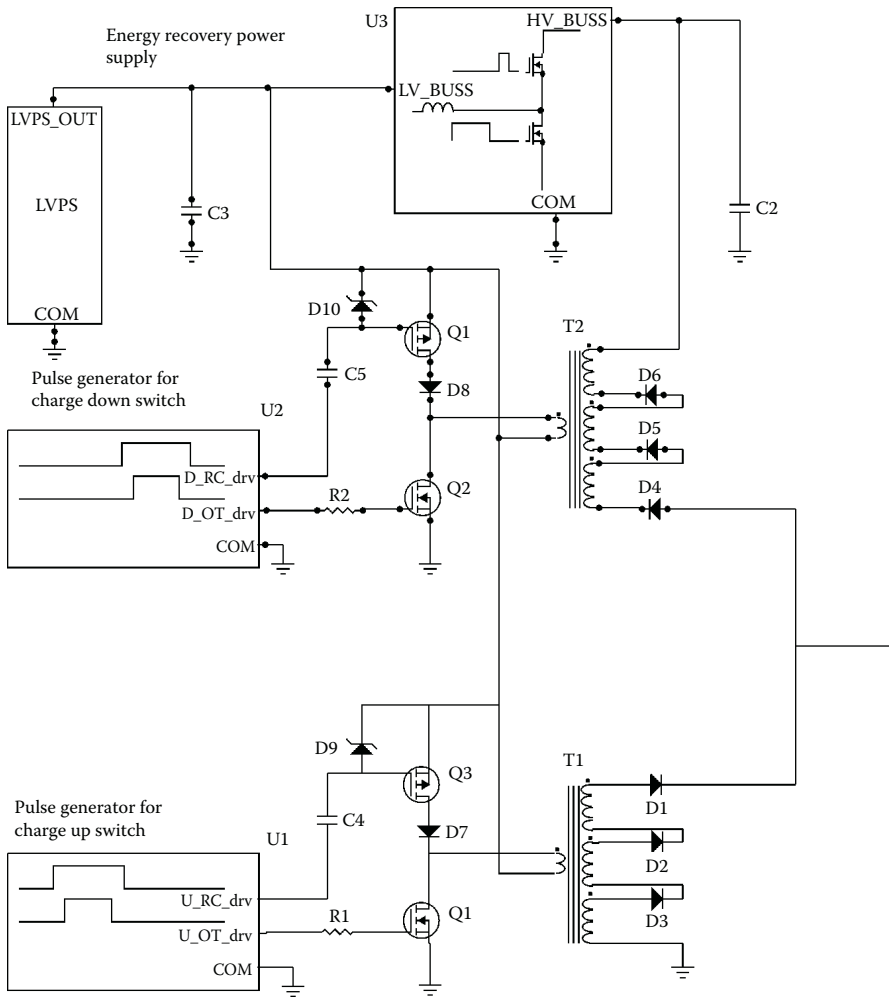
FIGURE 12.26

A schematic of flyback high-voltage supply with an energy recovery circuit, allowing for connection to an adiabatic driver circuit.

Of course, the HV supplies used in an adiabatic driver control scheme must be able to source and sink current—which is not true of most HV supplies. However, additional circuit elements can be added to the flyback converter to accommodate this requirement. Further gains in overall energy efficiency can be realized by making the HV power sink operate as an energy recovery system. One option for this is shown in Figure 12.26, where the energy pushed back to the HV rail is fed into a step-down DC–DC converter with the output connected to a system level low voltage bus.[32]

12.4.4 Analog Drivers

The wide variety of EO applications and elements leads naturally to a wide variety of drivers, especially for "nondigital" applications. Possible designs range from those using densely packaged hybrid circuitry to those using 4000 V, or higher, vacuum tubes for

control elements. A variety of IC op-amps are also available up to 600 V, allowing for essentially arbitrary waveforms within the bandwidth of the amplifier.

Analog drivers are generally specified by their operating voltage and bandwidth. The bandwidth of a driver is always specified in relation to the load; the usable bandwidth of the driver will decrease as the load capacitance increases. For this reason, the cabling or other connection between the driver and the EO element must be considered in addition to the actual EO element load when specifying the driver, since they can sometimes dominate the overall system performance. Efficiency, single-ended or differential input, linearity, and other factors may also be specified.

A sampling of the variety of designs possible is:

1. Single ended: this is the simplest, most straightforward variety of driver. The driver is connected to one electrode on the EO device, with the other usually held at ground. Op-amps are typical single-ended drivers.

2. Differential: some amplifiers work best in the middle of their output voltage range. A differential driver is constructed using two amplifiers, each connected to one side of the EO element and each biased to one-half the peak voltage. As the voltage on one is raised, the other is lowered. Differential topologies can be used where deflection in either direction from nominal is required. They also effectively act as a single-ended design but with twice the voltage—which may be useful if FETs at the desired voltage are not available.

3. Resonant: since EO elements act as capacitors, they can be coupled with an inductor and driven at the resonant frequency of the pair. Voltage amplification is possible, and standard techniques can be used to synchronize the circuit with other elements of the complete optical system.

4. Transformer coupled: voltage amplification is also possible using a transformer. This type of circuit may be operated in a resonant mode, or to produce some other waveform. Transformer coupling is best suited to periodic waveforms since the bandwidth of a transformer can be limited, and special tuning and compensation techniques may be required for complex waveforms. Triangle-wave drive voltages, useful for some display applications, can be done very effectively.

12.5 PROPERTIES AND SELECTION OF ELECTRO-OPTIC MATERIALS

12.5.1 General

The optical performance of an EO deflection system depends on the material chosen, the operating electric field, and the characteristics of the beam to be deflected (wavelength, diameter, divergence, M^2 quality factor, and so on). Unfortunately for system designers, there is no one "best" material that meets most application requirements.

Most practical EO materials are crystals with anisotropic properties, and almost all of these are grown by crystal suppliers because they are not abundant in nature. Properties can vary by manufacturer and grade, making it important to work in concert with the material suppliers when selecting a material; they are also good sources of material prop-

TABLE 12.5

Basic Properties of Some Popular EO Materials[11]

Material	EO coefficient $r_{ij}(10^{-12}$ m/V)	Index of refraction	Dielectric constant
PLZT	$r_{13} = 67$	$n_0 = 2.312$	
	$r_{33} = 1340$	$n_e = 2.299$	
$LiNbO_3$	$r_{13} = 9.6$	$N_0 = 2.286$	$\varepsilon_1 = \varepsilon_2 = 78$
	$r_{22} = 6.8$	$n_e = 2.200$	$\varepsilon_3 = 32$
	$r_{33} = 31$		
$LiTaO_3$	$r_{13} = 8.4$	$n_0 = 2.176$	$\varepsilon_1 = \varepsilon_2 = 51$
	$r_{33} = 30.5$	$n_e = 2.180$	$\varepsilon_3 = 45$
KH_2PO_4 (KDP)	$r_{41} = 8$	$n_0 = 1.507$	$\varepsilon_1 = \varepsilon_2 = 42$
	$r_{63} = 11$	$n_e = 1.467$	$\varepsilon_3 = 21$
KD_2PO_4 (KD*P)	$r_{63} = 24.1$	$n_0 = 1.502$	$E_3 = 50$
		$n_e = 1.462$	
$(NH_4)H_2PO_4$ (ADP)	$r_{41} = 23.41$	$n_0 = 1.522$	$\varepsilon_1 = \varepsilon_2 = 58$
	$r_{63} = 7.83$	$n_e = 1.477$	$\varepsilon_3 = 14$
$Ba_{0.25}Sr_{0.75}Nb_2O_6$	$r_{13} = 67$	$n_0 = 2.3117$	$\varepsilon_3 = 3400$ (15 MHz)
(SBN, Tc = 395 K)	$r_{43} = 1340$	$n_e = 2.2987$	
	$r_{51} = 42$		
$KNbO_3$	$r_{13} = 28 \; r_{23} = 1.3$	$n_1 = 2.280$	
	$r_{42} = 380 \; r_{33} = 64$	$n_2 = 2.329$	
	$r_{51} = 105$	$n_3 = 2.169$	

[a] All properties are measured at 633 nm (for optical properties), and low frequency (for dielectric constant). All properties are approximate, and should be verified with material vendors during procurement. (From Yariv, A. *Optical Electronics in Modern Communications;* Oxford University Press: New York, 1997. With permission.)

erty information. Extensive lists of material properties can also be found in the literature.[33] Electro-optical properties of some common crystals are shown in Table 12.5.

Crystal purity, optical quality, internal strain, physical size, doping, domain structure, electrical conductivity, and other quality measures can vary widely from vendor to vendor, from one boule to the next by a single vendor, and within a single boule as well. These and related factors have led the industry to focus on a few materials that are in relatively large-scale production, such as lithium niobate and ADP, which are covered in Section 12.5.2. Other materials, such as lithium tantalate and KTP, are also covered since they are steadily increasing in quality and availability.

Materials with very high EO coefficients, such as bulk form strontium barium niobate (SBN) and PLZT in the form of deposited films, are starting to appear. Other "new" materials are doped forms of lithium niobate or tantalate, and stoichiometric lithium niobate and tantalate. Many of these materials still exhibit performance variations between suppliers—it is best to discuss your needs with several crystal suppliers and perform multiple sample runs to ensure that the proper selection is made.

12.5.2 ADP, KDP, and Related Isomorphs

Relatively easy commercial growth processes make ADP (NH4H2PO4) and KDP (KH_2PO_4) popular materials for bulk EO devices. High optical quality crystals can be grown to over 10 cm diameter, and can be cut, polished, and mounted without undue difficulty. In both materials, the hydrogen atoms may be replaced with deuterium. In this case they are referred to as AD*P and KD*P. This replacement results in an increase in the linear EO coefficients by a factor of ~2.5.

The resistivity of ADP and related materials is very high, typically $> 10^{10}$ Ω cm when operated near room temperature. As operating temperatures approach the Curie temperature, C_T, the loss tangent and the dielectric constant increase, leading to heating and high power draws on electronic drivers if operated at high speeds. Heating in the crystal also leads to beam distortion via the thermo-optic properties of the material.

One major drawback to using ADP and its isomorphs in EO scanning systems is that the materials are hygroscopic. They are typically housed in a hermetic package that is filled with a dry gas or an index matching liquid that also provides electrical insulation. Some of the liquids are toxic, although the packaging is generally reliable.

12.5.3 Lithium Niobate and Related Materials

A large class of ferroelectric materials have the form $A^{1+}B^{5+}O_3$ or $A^{2+}B^{4+}O_3$, and are related to the mineral perovskite ($CaTiO_3$). Several of these materials are in mass production for devices based on their piezoelectric properties, such as $LiNbO_3$ for surface acoustic wave filters, which are found in cell phones and a host of other signal processing applications. Czochralski growth is the typical practice, with boule diameters approaching 15 cm for $LiNbO_3$, although 7.5 and 10 cm is more common. During the processing of crystal boules, they are typically poled to form a single domain throughout by heating to a point near the Curie temperature and then applying a DC electric field, which is maintained during the cooldown. This ensures that all crystal domains have uniform orientation—a critical consideration for good optical quality of the ensuing device.

The perovskite materials are not water soluble, eliminating some of the packaging problems encountered with ADP and similar materials. Another use of perovskites, in particular lithium tantalate, is as a pyroelectric detection element. If precautions are not taken in handling and processing, extremely high voltages can be generated between the faces of wafers. This charging can lead to electrical flashovers, which can damage electrodes or other coatings, or it can damage attached electrical equipment such as drivers or thermocouples. Wherever possible, controlled slow heating and cooling are recommended; the use of air ionizers in the work space also mitigates these effects.

The piezoelectric properties of the perovskites must be considered whenever the deflector will be operated at high speeds. The electrostrictive strain component of the index change can be as large as the EO component if a mechanical resonance is present. The mechanical performance of the entire device assembly—crystal, electrical leads, mounting adhesive and mounting base—must be considered early on, but careful testing is still a requirement.

Lithium niobate ($LiNbO_3$) and lithium tantalate ($LiTaO_3$) are the most common perovskite materials in use. For reasons of producibility and quality, they are typically grown to be slightly lithium-rich—this is referred to as congruently grown niobate or tantalate. These congruent materials exhibit Curie temperatures of 1470 and 890 K, respectively, giving them stable EO properties at room temperature or slightly elevated temperatures. They are also commercially available with fairly consistent properties across several vendors.

Recently, stoichiometric lithium niobate and lithium tantalate have been produced in commercially relevant sizes.[18,34] These materials exhibit lower coercive fields, which leads to easier fabrication of poled devices, and higher EO coefficients, as well as a broader transmission range. Being relatively new, it is best to contact the crystal growers for detailed information on properties and processing practices for stoichiometric materials.

Various dopants can be introduced when growing lithium niobate or tantalate to alter properties for special applications. Magnesium is a common dopant for lithium niobate, added to mitigate photorefractive damage from short wavelengths. This variant of the material is often used when producing SHG or other nonlinear devices, typically in the visible spectrum.

Significant effort has been applied to the problem of domain inversion processing of lithium niobate and, to a lesser extent, lithium tantalate. These efforts were driven primarily by interests in SHGs and related nonlinear devices, but the practices are transferable to the production of scanners, as mentioned previously.

Barium titanate ($BaTiO_3$) and KTN (a solid solution of $KTaO_3$ and $KNbO_3$) belong in the perovskite group of materials. They have good EO properties, but have critical temperatures near or below room temperature. This makes some properties very temperature dependent, and can result in creating or changing ferroelectric domains simply by handling and processing the crystal. They are not in wide use at this time.

12.5.4 Potassium Titanyl Phosphate (KTP)

In 1976 the Du Pont Company reported[35] on the growth and properties of the crystalline material $K_xRb_{1-x}TiOPO_4$. The material is ferroelectric and has found use in the production of SHGs and other nonlinear devices. The material is relatively difficult to grow (compared to lithium niobate), although the situation is improving due in part to military interest in the material, with slabs over 40 mm square being produced.

Dopants and special processing are used to produce various grades of KTP, one must check with manufacturers for availability and detailed properties. It is difficult to produce domain-inverted devices in KTP since the coercive field and dielectric strength are very near each other. Problems have also been encountered with relatively high electrical conductivity of the crystals, especially in large flux-grown crystals, further limiting its appeal.

12.5.5 Other Materials

12.5.5.1 AB-Type Binary Compounds

The main interest in these materials has been for EO devices in the infrared, particularly at 10.6 µm for use in CO_2 laser systems. GaAs, ZnTe, ZnS, CdS, and CdTe are among the most common materials that are available in large sizes. The EO coefficients of these materials are relatively small, only about 10% of lithium niobate, but their transmission beyond 10 µm may make them useful in some applications, especially for military uses.

12.5.5.2 Kerr Effect in Liquids

Much attention has been paid in the past to Kerr effect liquids, especially nitrobenzene.[1] The attraction was due to the high purity of materials compared to most crystals, and the basically unlimited size of the resulting device.

Advances in crystal growth techniques have largely mitigated the perceived quality and size advantages a liquid material may offer. In addition, the liquids can exhibit heating,

currents, turbulence, and other behavior that affects performance. They are not widely used or considered for application at this time.

12.5.5.3 Electro-Optic Ceramics in the (Pb, La)(Zr, Ti)O₃ System

Lanthanum-modified lead zirconate titanate (PLZT) ceramic materials have been investigated since 1969 for their EO properties, which can be tailored to a degree by controlling the precise chemical makeup of the material.[1]

These materials are not available in large single crystal form, limiting their application to scanning systems. Scanning devices based on thin films of PLZT have been proposed, but the problems of coupling into and out of such films are daunting.

12.5.5.4 Other Materials

Significant EO materials development is still ongoing, driven by the reality that almost all current EO systems could be improved by using materials with higher EO coefficients, higher optical damage limits, lower conductivity, or improvements in other technical parameters. Efforts can be roughly characterized as either creating new materials or modifying existing ones.

An example of a relatively new material is SBN. The crystal is grown by the Czochralski method, with boule diameters typically under 50 mm. There are several, slightly different formulations available, and quality and properties can vary even in modestly sized samples. Very high EO coefficients and relatively low Curie temperature make the material attractive for future applications. It is best to contact the material growers for current specifications prior to developing a design with an SBN element.

An example of modifications to a standard material is the development of magnesium doped lithium niobate, available from a variety of suppliers. The doping raises the optical damage threshold in the short wavelength part of the transmission band—a characteristic important for SHG devices, among others.

Careful selection, specification, inspection, and qualification of materials are a key to successful EO system design. Given the continuous improvement in crystal growing practices, inspection techniques, and materials formulation it is imperative to work closely with the materials suppliers during the design process.

12.5.6 Material Selection

Currently, there are only a handful of materials suitable for use in commercial EO systems. In addition to optical transparency at the desired wavelength, the following factors are common to nearly all applications.

1. High electrical resistivity: greater than $10^{10} \, \Omega$ cm is desired. This requirement stems from the desire for no resistive heating when operating voltages (typically hundreds of volts) are applied to the device. Ion migration can also occur, especially in the presence of DC fields, which can create substantial optical perturbations.

2. High optical homogeneity: refractive index variations of less than 1 in 10^6 are desired. This requirement helps to preserve beam quality, and can also be a stand-in for crystal compositional variations.

3. Large EO effect: absolute index variations of at least 10^{-4} are desired, with reasonable applied voltage. Excessively high voltages create packaging problems, long-term drift due to ion migration and require high driver power when high-speed operation is required.

4. Processability: standard handling and process operations should not impact material quality or device performance. For reasonable cost, the materials must be able to be oriented, cut, polished, AR coated, and mounted with only minor (if any) departure from practices used for other optical materials.

If the device is going to be operated at high frequencies, the dielectric constant and loss tangent become important considerations, with thermal conductivity and temperature dependence of optical properties also needing to be considered.

If the device under consideration will be produced via a domain inversion process, other factors must be added to the list of considerations. Obviously it must be ferroelectric, which implies that piezoelectric and pyroelectric effects must also be taken into consideration. Also, the desire to pole the material will restrict the orientations available. Poled devices should also not be operated near to their coercive field or near the Curie temperature, which can be quite low for some materials, or there is a chance of depoling occurring.

12.6 ELECTRO-OPTIC DEFLECTION SYSTEM DESIGN PROCESS

Selecting a system design for a particular set of operating parameters is an iterative process, likely requiring thorough analysis of multiple trial designs. The complex interplay of material properties, EO element geometry, electronic power consumption, operating speeds, and the realities of current fabrication methods does not lend itself to a closed form solution, nor are there large catalogs of standard alternatives to choose from.

The recommended process is:

1. Verify that the speed, optical efficiency, or ruggedness of an EO deflector is required. If other technologies such as galvanometers or AO deflectors can be used, they are likely going to triumph in a head-to-head comparison of total system cost and complexity.

2. Consider the wavelength of the laser to be used. Few, if any, materials are optically clear across the entire spectrum of wavelengths available today. Also, many material properties vary with wavelength so it is important to look at the properties of interest at the wavelength of interest.

3. Consider whether the system can be built using a single linear polarization. If not, it is likely that beam splitting and recombining after deflection will be required, adding significant complexity and cost to the system.

4. Review the operating environment in light of safety and reliability of high-voltage electrical systems. If moisture and dust are present, the EO scanner will likely need to be placed in a sealed housing, adding length and additional windows to the optical path.

5. Select appropriate design guidelines and safety factors such as electrical creepage distances, dielectric strength, and optical power per unit area on surfaces. These guidelines and safety factors may bound the design options.

6. Generate and analyze trial designs. One aspect that is sometimes overlooked is the mechanical response of the EO material. All attractive EO crystals exhibit piezo-electric responses to some degree. In some cases, high-speed electrical pulses can excite mechanical resonances that create a time-varying strain in the material, which can alter the deflection from that expected or contribute to beam losses.

7. Verify that the selected design can actually be built by consulting with the appropriate material suppliers, electronics designers, optical designers, and engineers familiar with current practice for all manufacturing steps.

12.7 CONCLUSIONS

Electro-optic scanning systems can be very fast and optically very efficient. This performance often comes at the cost of working near the cutting edge of materials, electronics, and processing technologies. The demands for speed are likely to continue increasing, however, as laser and computing technology continue to advance. To address these demands, a true systems approach should be used when designing an EO scanning system. The overlapping implications of decisions in areas as diverse as HV electronic drivers, mechanical isolation, temperature control, and beam size must be weighed carefully.

Large scale application of EO scanners has not yet occurred due to the difficulties and uncertainties discussed in this chapter. Recent advances in telecommunications switching, computer-to-plate printing, and biomédical imaging applications are driving development of EO devices and electronic drivers at a rapid pace. New EO materials, with new combinations of properties, are also being developed. Such progress in each of the key areas of EO scanning system design and construction may eventually lead to their wider application.

ACKNOWLEDGMENTS

REFERENCES

1. Ireland, C; Ley, J. Electrooptical scanners. In *Optical Scannin.* Marcel Dekker: New York, 1987; 687–778.
2. Stancil, D.D. Electro-optical scanners. In *Encyclopedia of Optical Engineering.* Marcel Dekker: New York, 2003; 456–474.

3. Lee, S.M.; Hauser, S.M. Kerr constant evaluation of organic liquids and solutions. *Rev. Sei. Instruments* 1964, *35*, 1679.

4. Kruger, R.; Pepperl, R.; Schmidt, U. Electrooptic materials for digital light beam deflectors. *Proc. IEEE* 1973, *61*, 992.

5. Chiu, Y.; Zou, J.; Stancil, D.D.; Schlesinger, T.E. Shape-optimized electrooptic beam scanners: Analysis, design, and simulation. *J. Lightwave Technol.* 1999, *17*(1), 108–114.

6. Lotspeich, J.F. Electrooptic light-beam deflection. *IEEE Spectrum* 1968, 5, February, 45–52.

7. Lee, T.C.; Zook, J.D. Light beam deflection with electrooptic prisms. *IEEE J. Quantum Electronics* 1968, *QE-4*(7), 442–454.

8. Fowler, V.J.; Buhrer, CF.; Bloom, L.R. Electro-optic light beam deflector. *Proc. IEEE* 1964, *52*(2), 193–194.

9. Fowler, V.J.; Schlafer, J.A. Survey of laser beam deflection techniques. *Appl. Optics* 1966, *5*(10), 1675–1682.

10. Kiyatkin, R.P. Analysis of control field in quadrupole optical-radiation deflectors. *Opt. Spectrosc.* 1975, *38*(2), 209–210.

11. QuickField, for finite element calculations. Retrieved from http://www.quickfield.com March 22, 2004.

12. Ireland, C; Ley, J. Electrooptical scanners. In *Optical Scanning;* Marshall, G., Ed.; Marcel Dekker: New York, 1987; 752–754.

13. Armstrong, J.A.; Bloembergen, N.; Ducuing, J.; Pershan, P.S. Interactions between light waves in a nonlinear dielectric. *Phys. Rev.* 1962, *127*, 1918–1939.

14. Fejer, M.M.; Magel, G.A.; Jundt, D.H.; Byer, R.L. 'Quasi-phase-matched second harmonic generation: tuning and tolerances.' *IEEE J. Quantum Electronics* 1992, *28*(11), 2631–2654.

15. Mizuuchi, K.; Yamamoto, K. Highly efficient quasiphase-matched 2nd harmonic generation using 1st-order periodically domain-inverted $LiTaO_3$ waveguide. *Appl. Phys. Lett.* 1992, *60*(11), 1283–1285.

16. Wang, Y.; Petrov, V.; Ding, Y.J.; Zheng, Y.; Khurgin, J.B.; Risk, W.P. Ultrafast generation of blue light by efficient second-harmonic generation in periodically-poled bulk and waveguide potassium titanyl phosphate. *Appl. Phys. Lett.* 1998, *73*(7), 873–875.

17. Ktaoka, Y.; Narumi, K.; Mizuuchi, K. Waveguide-type SHG blue laser for high-density optical disk system. *Rev. Laser Eng.* 1998, *26*(3), 256–260.

18. Gopalan, V.; Sanford, N.A.; Aust, J.A.; Kitamura, K.; Furukawa, Y. Crystal growth, characterization, and domain studies in lithium niobate and lithium tantalate ferroelectrics. In *Handbook of Advanced Electronic and Photonic Materials and Devices,* Nalwa, H.S., Ed.; Academic Press: New York, 2001; Vol. 4, Ferroelectrics and Dielectrics, 57–114.

19. Li, J.; Cheng, H.C.; Kawas, M.J.; Lambeth, D.N.; Schlesinger, T.E.; Stancil, D.D. Electrooptic wafer beam deflector in $LiTaO_3$. *IEEE Photonics Tech. Letts.* 1996, *8*(11), 1486–1488.

20. Chen, Q.; Chiu, Y.; Lambeth, D.N.; Schlesinger, T.E.; Stancil, D.D. Guided-wave electro-optic beam deflector using domain reversal in $LiTaO_3$. *J. Lightwave Technology* 1994, *12*(4), 1401–1404.

21. Chen, Q.; Chiu, Y.; Devasahayam, A.J.; Seigler, M.A.; Lambeth, D.N.; Schlesinger, T.E.; Stancil, D.D. Waveguide optical scanner with increased deflection sensitivity for optical data storage. In *SPIE Proc. Series,* Vol. 2338, 1994; Topical Meeting on Optical Data Storage, Dana Point, CA; May 16–18, 1994; 262–267.

22. Fang, J.C.; Kawas, M.J.; Zou, J.; Gopalan, V.; Schlesinger, T.E.; Stancil, D.D. Shape-optimized electrooptic beam scanners: experiment. *IEEE Photonics Technol.* Lett. 1999, *11*(1), 66–68.

23. Chiu, Y.; Burton, R.S.; Stancil, D.D.; Schlesinger, T.E. Design and simulation of waveguide electrooptic beam deflectors. *J. Lightwave Technol.* 1995, *13*(10), 2049–2052.

24. Feit, M.D.; Fleck, J.A., Jr. Light propagation in graded-index optical fibers. *Appl. Opt.* 1978, *17*(24), 3990–3998.

25. Eason, R.; Boyland, A.; Mailis, S.; Smith, P.G.R. Electro-optically controlled beam deflection for grazing incidence geometry on a domain-engineered interface in $LiNbO_3$. *Optics Commun.* 2001, *197*, 201–207.

26. Gnewuch, H.; Pannell, C; Ross, G.; Smith, P.G.R.; Geiger, H. Nanosecond response of Bragg deflectors in periodically poled $LiNbO_3$. *IEEE Photonics Technol. Lett.* 1998, *10*(12), 1730–1732.

27. Kawas, M.J. Design and characterization of domain inverted electro-optic lens stacks on $LiTaO_3$. Department of Electrical and Computer Engineering; Carnegie Mellon University, 1996; M.S. Thesis.

28. Kawas, M.J.; Stancil, D.D.; Schlesinger, T.E. Electrooptic lens Stacks on $LiTaO_3$ by domain inversion. *J. Lightwave Technol.* 1997, *15*(9), 1716–1719.

29. Gahagan, K.T.; Gopalan, V.; Robinson, J.M.; Jia, Q.; Mitchell, T.E.; Kawas, M.J.; Schlesinger, T.E.; Stancil, D.D. Integrated electro-optic lens/scanner in a $LiTaO_3$ single crystal. *Appl. Optics* 1999, *38*(4), 1186–1190.

30. Gopalan, V.; Kawas, M.J.; Gupta, M.C.; Schlesinger, T.E.; Stancil, D.D. Integrated quasi-phase-matched second-harmonic generator and electrooptic scanner on $LiTaO_3$ single crystals. *IEEE Photonics Technology Lett.* 1996, *8* (12), 1704–1706.

31. Chiu, Y.; Gopalan, V.; Kawas, M.J.; Schlesinger, T.E.; Stancil, D.D.; Risk, W.P. Integrated optical device with second-harmonic generator, electrooptic lens, and electrooptic scanner in $LiTaO_3$. *J. Lightwave Technol.* 1999, *17*(3), 462–465.

32. Cleland, A.; Gass, H. Energy recirculating driver for capacitive load. Patent Cooperation Treaty application, document #WO 02/14932, August 16, 2001; revised February 21, 2002.

33. Yariv, A. *Optical Electronics in Modern Communications;* Oxford University Press: New York, 1997.

34. Furukawa, Y.; Kitamura, K.; Suzuki, E.; Niwa, K.J. Stoichiometric $LiTaO_3$ single crystal growth by double crucible Czochralski method using automatic powder supply system. *Crystal Growth* 1999, *197,* 889.

35. Zumsteg, F.; Bierlein, J.; Gier, T. $K_xRb_{1-x}TiOPO_4$: A new nonlinear optical material. *J. Appl. Phys.* 1976, *47,* 4980.

36. Revelli, J.F. High-resolution electrooptic surface prism waveguide deflector: An analysis. *Appl. Optics* 1980, *19,* 389–397.

37. Lee, C.L.; Lee, J.F.; Huang, J.Y. Linear phase shift electrodes for the planar electrooptic prism deflector. *Appl. Optics* 1980, *19,* 2902–2905.

38. Sasaki, H.; De La Rue, R.M. Electro-optic multichannel waveguide deflector. *Electronics Letts.* 1977, *13*(10), 295–296.

39. Chiu, Y.; Burton, R.S.; Stancil, D.D.; Schlesinger, T.E. Design and simulation of waveguide electrooptic beam deflectors. *J. Lightwave Technol.* 1995, *13*(10), 2049–2052.

40. Takizawa, K. Electrooptic Fresnel lens-scanner with an array of channel waveguides. *Appl. Optics* 1983, *22*(16), 2468–2473.

13

Piezo Scanning

James Litynski and Andreas Blume

Piezosystem Jena, Inc.
Hopedale, Massachusetts, USA

CONTENTS

13.1 INTRODUCTION

The piezoelectric effect is the production of positive and negative charges on the surface of certain types of crystalline structures in the presence of an externally applied force that changes the shape of the crystal. The converse is also true. When positive and negative charges are applied to the surface of these same crystals, forces are produced within the crystals which change its shape. This inverse piezoelectric effect has recently been used in micropositioning devices to good effect. In low voltage (<200 V) applications, a potential of −20 V to 150 V is applied to bonded stacks of such crystals so that they expand. With current piezoelectric stacks available commercially today, such expansion is on the order of 1 μm for every 1 mm of stacked piezoelectric crystal. Due to this rather small expansion, and the unpredictable trajectory of the crystal faces, flexure hinges are commonly used to

both amplify and guide the motion in a piezo stage. Major advances in stage design within the past 15 years have resulted in a large variety of scanning and fine positioning devices which take advantage of the rather unique properties of this type of drive mechanism.

Piezoelectric flexure drive mechanisms offer a number of advantages over traditional stepping and servo motor bearing guide stages. One of the main advantages of this type of drive is their very high resolution. Since the entire flexure mechanism is driven by the smooth expansion of a crystal, it is a friction and stiction free system with a resolution that is only limited by the electrical noise on the voltage potential being applied to create the inverse piezoelectric effect and the mechanical noise that may affect the flexure structure from the environment. This has been the driving factor behind the development of extremely low-noise amplifiers and highly stiff, robust mechanical flexures with high resonant frequencies to keep outside vibrations that may excite the stage to a minimum. A second advantage of piezoelectric stages is the crystal's ability to exert extremely high forces. A 25-mm^2 area of PZT (lead zirconium titanate) can exert a force in excess of 1000 N. A third advantage of piezoelectric stages is their high resonant frequency. PZT preloaded stacks by themselves may safely operate up to 75 kHz and, when integrated into a flexure mechanism, 4 kHz is not uncommon for a tilting stage, making them very useful in laser scanning applications.

A number of obstacles present themselves in practice when using piezoelectric materials in a laser scanning system. These include a large hysteresis in voltage versus position, drift or creep, unpredictable force vectors during crystal expansion, temperature limitations due to the Curie temperature and reduction of the piezoelectric effect at low temperatures, environmental noise amplification by the flexure design, low tolerance for tensile forces, and electrical failure due to ion migration at the electrical contacts.

The object of this chapter will be to give a practical understanding of the use of these devices allowing the scientist or engineer to make the most of their inherent advantages while addressing the constraints they impose.

13.2 STRUCTURE AND DESIGN

Piezoelectric ceramics exhibit a Perovskite ionic lattice structure (AXO_3). Below their Curie temperature these types of crystals exhibit an inherent polarization. This is caused by a deviation of the titanium ion from its center position within the crystal lattice which creates an electrical dipole.

Under normal conditions such a ceramic exhibits an anisotropic structure and all the dipoles within the material domains of the ceramic are oriented randomly. However, in the presence of a strong electric field these domains become ferroelectric. The dipoles all align in the direction of the electric field. This effect is long term. After removal of this strong electric field, the structure relaxes but remains strongly polarized (Figure 13.1).

Depending on the orientation of the electric field to the crystal structure, forces generation by the piezoceramic can be characterized as longitudinal d_{33}, transverse d_{31} or shear d_{15}. The transverse mode is used in bi-morph (bending) and tube piezoelectric actuators. Stacked-type actuators use the longitudinal mode for expansion. To somewhat limit the scope of this chapter, we will concentrate on theory and devices which take advantage of the longitudinal mode.

Modern piezo stack structures consist of hundreds of thin layers of PZT with a thickness of approximately 100 μm. Electrodes are arranged on opposite sides of the stack structure
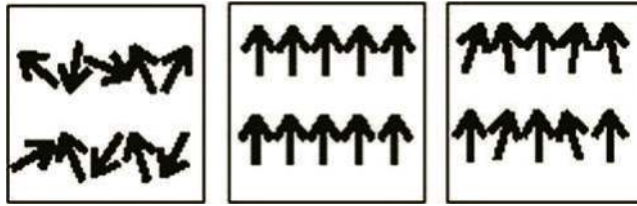
FIGURE 13.1
Domain polarization randomly ordered, saturated, and remanent (from left to right).
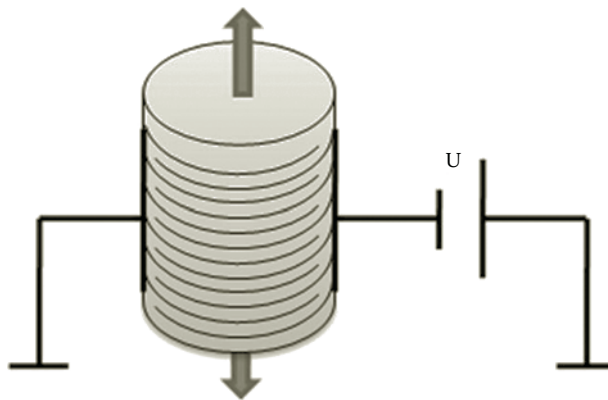
FIGURE 13.2
Construction of a piezo stack.

with connections to deposited metallic film such that alternating charged and grounded plates are sandwiched between the crystal stacks (Figure 13.2). The very thin layers of PZT require relatively low voltages in order for all dipoles to become completely aligned. A reversal in the polarity of these plates and therefore the direction of the electric field can result in some contraction of the crystalline structure, but the majority of the motion can be achieved by expansion.

Imagine the domain dipoles as arrows which orientate along the direction of the electric field. Applying a high field means the arrow stands vertically, head on top (maximum stroke). Without the electrical field the arrow lays horizontally (minimum stroke). A reversal of the polarity of the electric field lets the arrow stand also, but head on bottom (maximum stroke) (Figure 13.3).

This is caused by the essential deviation of the titanium ion. In reality the "arrows" rotate only a few degrees. Saturation may exist even when all "arrows" are not aligned exactly parallel to each other; most of the "arrows " reached the maximum position early and some later, but the magnitude of the whole stroke remains the same for a given time period.

Many devices take advantage of the extended range offered by this contraction and hence most power supplies offer a voltage range which extends down to –20 V. Stack manufacturers recommend upper voltage limits of 100 V to 150 V. This limit is imposed by the breakthrough field strength of the ceramic and is affected by the thicknesses of the layers.
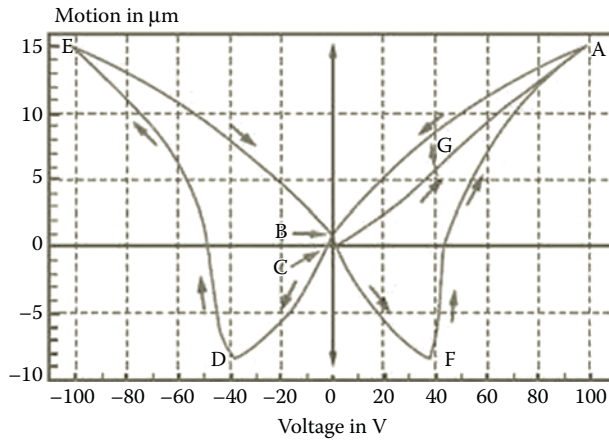
FIGURE 13.3
According to the applied voltage, the motion of a piezo element will follow the points ABCDEF.

The physical structure of these stacks—layers of charged metal plates sandwiched between an insulating material is, of course, the classical structure of a parallel-plate capacitor and, as you might imagine, these stacks have a quite large capacitance.

Solving for the capacitance of a parallel-plate capacitor we have the formula:

$$C = (\varepsilon_o \varepsilon_r A)/d$$

where A is the area of the plate, and d is the distance between them. The material specific relative permittivity ε_r takes into account the charge bonded by the ceramic material (dielectric) and greatly increases the capacitance.

From this formula the capacitance of a single layer of 5 mm × 5 mm PZT can be calculated:

$$C = (8.9\text{E}{-}12)F/m * 50\text{E}6 * (25\text{E}{-}6)m^2/100(\text{E}{-}6)m$$

And therefore a 10-mm stack of 100 such disks would have a capacitance of appr. 1.0 μF at a 50-Hz measuring frequency.

One might think from this calculation that d is, in this case, a variable which depends on electric field strength which is directly proportional to the voltage applied to the metal plates within the stack. As the stack expands or contracts, this value will change accordingly. However, the change in d for an application of 130 V is typically 0.1% (0.1 μm for a 100-μm layer) and for calculations concerning electrical requirements, this value is too small to be significant and other factors such as temperature and strain dependence are much larger than this. It should also be noted that current production methods for PZT layers are such that the thickness can vary quite a bit. Tolerance on stack thickness (and therefore capacitance, and motion/V (sensitivity)) can be as high as 10%.

There are two different methods to prepare multilayer stacks. In the first method, discs are cut from a ceramic bulk, and then sandwiched between electrical contact films and epoxy glue. This is finished by a pressing process. In the second method the ceramic is powdered and applied to the contact film, layer by layer followed by sintering process (without epoxy). The tensile strength of the electrical contact/PZT bonds determines the tensile strength of the stack.

This tensile strength is relatively weak compared to the compressive load that the stacks can endure. Under high dynamical operation internal accelerations inside the stack can generate forces which exceed the tensile strength limit of the bonds resulting in a delaminating of the layers. To prevent this, it is necessary to apply a preload to the stack of sufficient strength to overcome any internal force. A typical preload for a stack is 150 N. Even with this strong preload it is possible to damage the stack if it, or the mechanical structure of which it is a part, is driven at resonant frequency with a large amplitude. It is therefore important to determine the resonant frequency first and try to operate below it if possible. Operating above the resonant frequency is possible for low amplitudes (I would recommend <1% of the total motion), but of course the primary resonance will be excited in this case.

Since piezoelectric stacks are rather stiff and brittle it is important to take care when applying heavy loads or generating shock forces. The main concern here is to be sure that any load or force is applied directly on the same axis as the center of the stack. Off axis, shear loads or shock forces will typically crack the stack. A common strategy to minimize off-axis loading is to utilize a ball tip on the end of the stack or integrate the stack into a preloaded flexure which directs any outside force axially (Figures 13.4 and 13.5).
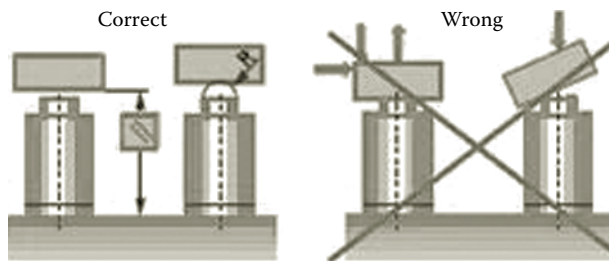
FIGURE 13.4
Applying load to a stack.
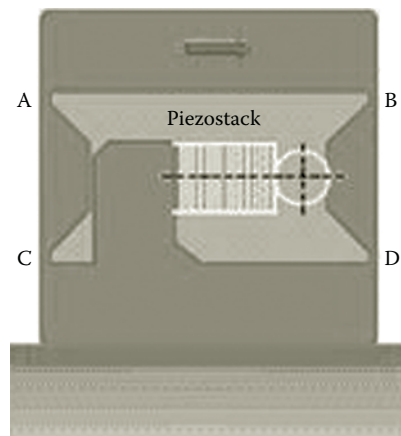
FIGURE 13.5
Integrated stack with flex hinges at ABCD. Lower bar is fixed and arrow indicates direction of motion.

13.3 TEMPERATURE EFFECTS

Temperature variations play an extremely important role in micrometer scale positioning systems. Often a neophyte piezosystem user will complain that his system is "drifting all over the place" when, in fact, he is trying to measure on a scale where material expansions, invisible to the naked eye, are causing relatively huge shifts in position and he is simply witnessing his thermostat kicking the heat or air conditioning on and off.

Consider: The thermal coefficient of expansion for steel from 20°C–100°C is about 16 µm/m °C. A simple piezo flexure stage may incorporate up to 50 mm of steel in its design with specified resolutions for closed loop systems of 10 nm or less. A variation of just half a degree will result in an expansion of the steel of 400 nm—or more than 400× the resolution of the stage!

To demonstrate this effect I measured the position of a parallelogram flexure design with an integrated PZT stack under static voltage conditions over an extended period of time with a michaelson-type interferometer. Temperature measurements of the air were taken simultaneously and the results are shown in Figure 13.6.

An interesting feature of PZT is that it has a *negative* coefficient of expansion at room temperature of −6 µm/m °C. The total displacement caused by thermal effects can be expressed by the following formula:

$$dl_{therm}/dT = L_{piezo} * \alpha_{piezo} + L_{metal} * \alpha_{metal}$$

To see the effect of this on a stack type actuator we can use the example of a 50-µm stack which uses a combination of three 16-mm stacks in series (Table 13.1).

It is also possible to construct a simple temperature compensated block by selecting the proportionally correct amount of steel block and PZT and arranging them in series. So a stack that has a motion of 16 µm (16 mm in length) would require a steel block of 6 mm to make a temperature compensated simple actuator. Materials with a higher temperature extension coefficient will decrease the additional length, for example, brass.
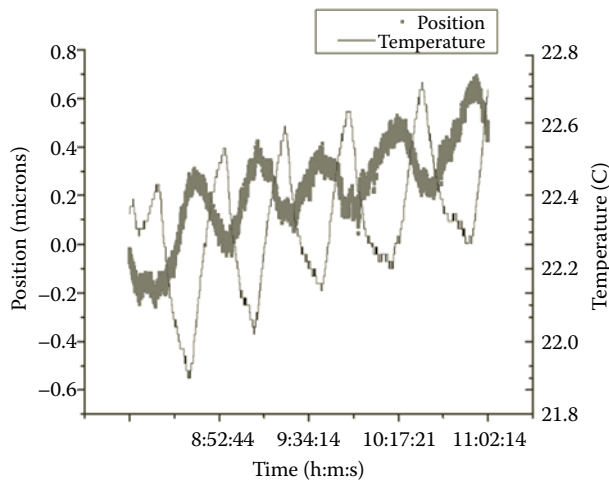
FIGURE 13.6
Temperature-induced motion on a 400-µm flexure stage.

TABLE 13.1

Thermal Effects on a PA 50 Preloaded Stack Actuator

All dimensions: μm/°C

Temp	Displacement	Thermal expansion[a]	Sum	Displ. three stacks	Therm exp. three stacks	Thermal exp. steel	Sum actuator
30	16.1	0.0	16.1	48.3	0.0	0.0	48.3
40	15.7	−1.8	13.9	47.1	−5.4	0.3	42.0
60	15.3	−3.5	11.8	45.9	−10.5	0.9	36.3
80	15.2	−5.5	9.7	45.6	−16.5	1.4	30.5
100	14.6	−7.4	7.2	43.8	−22.2	2.0	23.6
120	14.0	−9.4	4.6	42.0	−28.2	2.6	16.4

[a] Thermal expansion of steel parts is calculated with $16*10^{-6}*K^{-1}$ which causes an expansion of 30.4 nm/K.
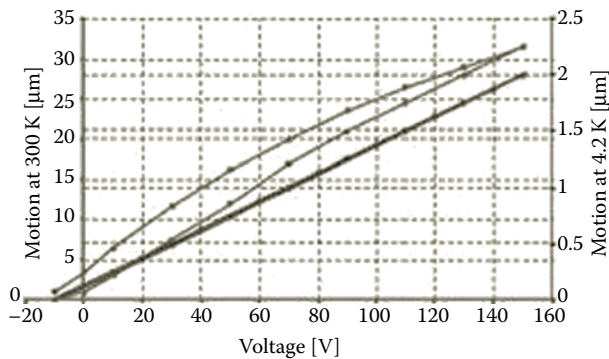
FIGURE 13.7
Piezo motion at room temperature (upper curve) and 4 K (lower curve).

Modern flexure designs utilize a variety of materials and methods for temperature compensation. But not all systems take temperature into consideration.

For most applications (–10°C to 90°C) the piezoeffect is relatively constant. However, outside this temperature range the effect begins to decrease and at extremely low temperatures (i.e., 4 K) the motion generated by an electric field on a stack can be as little as 6% of the motion generated at room temperature. The hysteresis curve of a stack which makes 32 μm of motion is shown in Figure 13.7 for both room temperature and liquid helium operation.

Of course even this reduced motion can be very useful for many cryogenic studies and as you can see, the problem of hysteresis is reduced to the point where a closed loop system is often unnecessary.

13.4 PROPERTIES OF MOTION

Due to the discrete structure of a piezoelectric stack, its expansion is often a bit unpredictable. Stacks have a tendency to exhibit all kinds of bad behavior such as twisting and tilting. Additionally, the expansion is nonlinear and exhibits a great deal of hysteresis.

Taken as a percentage of the full motion of the stack, this hysteresis can be as much as 12%. To further complicate matters, this hysteresis is temperature and load dependent. At higher temperatures and under higher loads, this hysteresis will be even greater. Figure 13.8 shows the typical response of a piezoelectric to a triangle wave function.

Piezoelectric stacks also exhibit some drifting in position or "creep." This is an asymptotic decay toward a final position when given a step function which is a result of some crystal domains within the layers coming into alignment with the electric field more slowly. This creep is dependent on the expansion of the PZT, the external load, and time. It can be calculated as a logarithmic function with the following formula:

$$dL/dt = dL_{0.1} [L + \gamma \lg (t/0.1s)]$$

L = length
t = time
$dL_{0.1}$ = change in length after 0.1 s after the step function
γ = drift constant (varies depending on loading, but typically 0.015)

A typical graph of creep over a period of 10 min is shown in Figure 13.9.
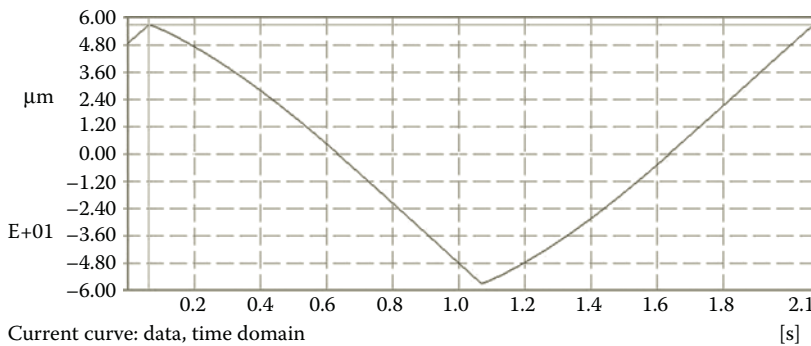
FIGURE 13.8
Piezoelectric stack total motion as a function of time as measured with an interferometer when driven with a triangle function at 0.5 Hz.
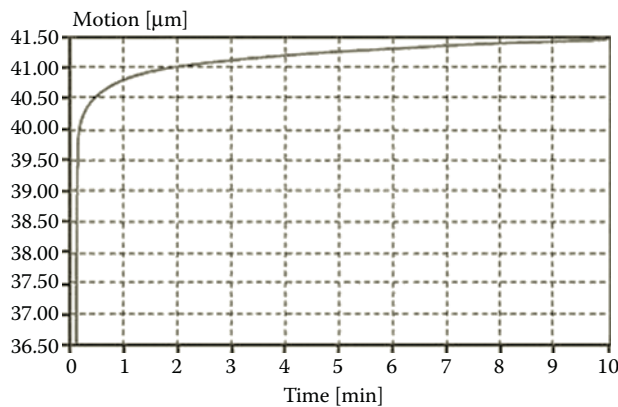
FIGURE 13.9
Creep exhibited by a PZT stack structure after given an electrical step function.

FIGURE 13.10
Demonstration of open loop repeatability of a PZT flexure.

A simple strategy often used for minimizing this creep is to overshoot the desired position by 5% or so and come back to it. This strategy brings the dipoles within the crystal structure more rapidly to their final rest state.

As you can see, the creep effect in piezosystems is long term and often negligible for dynamic operation. In fact, for high speed periodic signals piezosystems are highly repeatable. Although the function will still be subject to hysteresis, motion will follow the same track down to the noise of the system—often sub—nm. To provide an example, I drove a flexure stage capable of 240 μm at a frequency of approximately 1 kHz over a distance of 1 μm with a sine wave function provided by a Tektronix function generator and amplified with a Piezosystem Jena power supply. Results were measured and graphed with a SIOS interferometric vibrometer system (Figure 13.10).

13.5 PROPERTIES OF STACK-FLEXURE STRUCTURES

The integration of piezoelectric stacks into solid state flexures has two major benefits when predicting behavior:

1. The structure can be easily described with spring formulae of classical dynamics.
2. Finite Element Analysis computer modeling can predict behavior under a variety of loading conditions.

For instance, we can estimate a change in natural resonant frequency when a stack-flexure structure is affected by an external load. Let's take the example of a stack-flexure stage from Piezosystem Jena shown in Figure 13.11.

FIGURE 13.11
NanoX 200 flexure stage.

The following specifications for this stage are given as:

Unloaded resonant frequency: 700 Hz
Stiffness: 1.1 N/µm

We can use the following formula to calculate the distributed mass moved or "effective mass" of the unloaded structure:

$$m_{\text{eff}} = c_T/(2\pi f^{\circ}_{\text{res}})^2$$

m_{eff} = effective mass
c_T = stiffness
f°_{res} = unloaded resonant frequency

So for this case we calculate an effective mass of 57 g.

Now we need to move a mirror with a mass of 200 g and determine the resonant frequency of the system. We can use the following formula:

$$f^1_{\text{res}} = f^0_{\text{res}} \frac{\sqrt{m_{\text{eff}}}}{m_{\text{eff}} + \text{M}}$$

So for this case the resonant frequency has dropped from 700 Hz to 330 Hz.

Although a good order-of-magnitude estimate, things are actually not quite so simple since the stiffness c_T is a variable dependent on the amount of expansion of the stack and the load. Also, off-axis loading may cross couple out-of-plane resonances. When very accurate determinations of resonant frequency are necessary, the best method for determining this is by applying load to the stage under the conditions of use and measuring the response of the structure to a mechanical spike impulse and allowing the structure to "ring" at its resonant frequency.

A simple way to generate a spike mechanical impulse to a stack-flexure structure is by giving the piezo an electrical step function. I loaded this stage with 200 g and measured the response of the stage with an interferometer (Figure 13.12). A bandpass filter has been applied to the frequency domain graph (Figure 13.13) to isolate the region of interest.

From this you can see the primary resonance is about 300 Hz. Rise times for these types of systems are dependent on the resonant frequency and can be estimated to be about

$$\frac{1}{3f^1_{res}} \text{ or in this case about 1 ms for the loaded stage.}$$
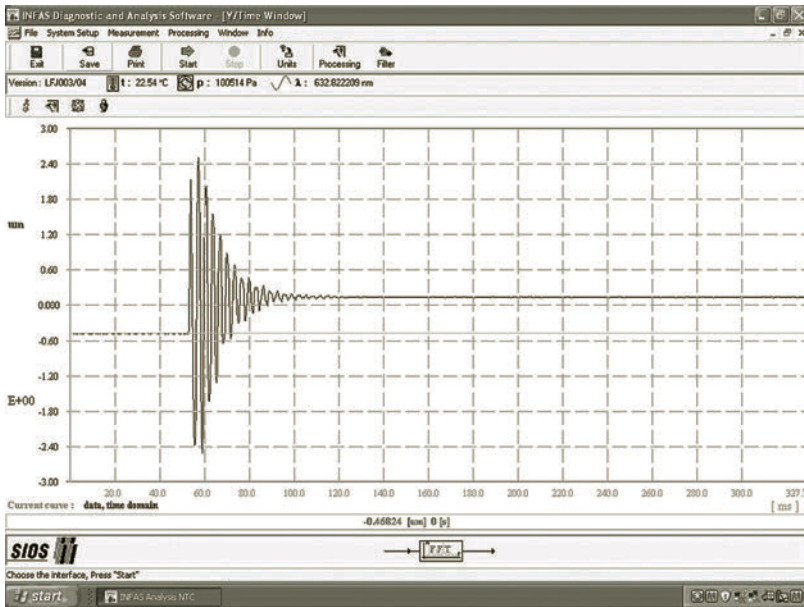
FIGURE 13.12
Step impulse time domain.
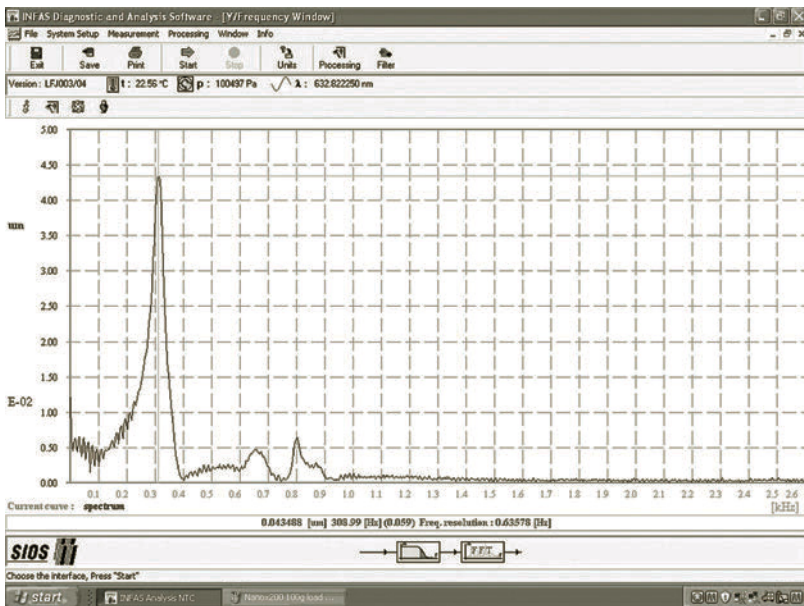
FIGURE 13.13
Step impulse frequency domain.

13.6 ELECTRICAL DRIVES

13.6.1 Noise

The positional noise inherent to a piezoelectrical system is only limited by the voltage noise of the power supply that is used to drive it. Commercially available amplifiers built for piezoelectric stages today typically have voltage noise of <300 μV rms over a large bandwidth. Since piezoelectric stacks typically operate over a voltage range of 150 V this voltage noise can be expressed as a relative noise of 2E–6. So for a piezo stage capable of a total motion of 240 μm as in the previous example, we can calculate the rms noise contribution from the electronics as 0.5 nm. It's pretty close to the resolution of the interferometer (0.3 nm in this case).

13.6.2 Current

The main consideration for current requirements of piezo stages is their large capacitance. Formulae for calculating bandwidth and rise times based on available current are simply those used for charging and discharging a capacitor with a very large resistance. So, you can imagine that if you charge a piezo stage to a potential of 70 V and leave it there it will draw almost no current from the amplifier. In fact, for many static applications a few mA will be sufficient. Let's consider the case of a single axis tilting stage PSH 4/1 as shown in Figure 13.14.

This stage has a capacitance of 200 nF. Let's say we have an amplifier capable of a maximum average output current of 50 mA and a peak current of 300 mA and we need to make small adjustments to the position to within 10 μrad to steer a laser beam. Based on the current we have, how fast can we do this? We can use the following formula for the time needed to charge or discharge a large capacitor:

$$dt = CdV/i_{max}$$

dt = change in time
C = capacitance
dV = voltage change
i_{max} = peak current available from the amplifier
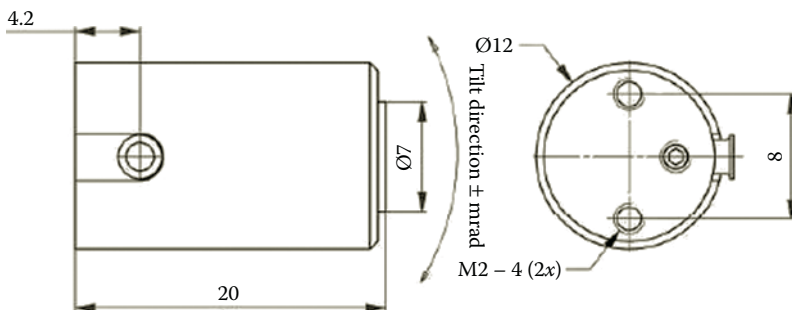
FIGURE 13.14
Small single axes mirror tilt scanning stage.

For this case we estimate the voltage change as 10 µrad/4 mrad × 150 V = 375 mV and solving for *dt*:

$$dt = (200E-9F)\ (375E-3V)/300E-3A$$

$$dt = 0.25\ \mu s$$

The limiting factor when considering whether or not to increase the current available to the stage in this type of application should be the mechanical rise time based on the resonant frequency of the system. For this stage the resonant frequency is 6.5 kHz. Using the rise time estimation of $dt = 1/3f$ we can calculate the maximum rise time limited by the mechanics to 50 µs. Therefore, increasing our current will not increase the speed of the stage for this application.

Now let's consider the case with a large dynamical requirement. Let's say we would like to scan the same laser beam by tilting the mirror over +/− 1.6 mrad at 4 kHz. How much current will we need from our amplifier? For a sinusoidal function we can use the following formula for oscillating a large capacitor:

$$i_{max} = \pi f C Vpp \quad \text{peak current required}$$

$$i_{average} = f C Vpp \quad \text{average current required}$$

For this case we calculate the *Vpp* to be 3.2 mrad/4 mrad × 150 V = 120 V and the limiting factor here will be average current available

$$i_{average} = (4000/s)\ (200E-9F)\ (120V) = 96mA$$

So we need a bigger amplifier.

These calculations are complicated by the fact that the capacitance of the stacks may increase by up to 200% due to changes in amplitude, strain, and temperature. So, for this case an amplifier with 200 mA should be sufficient under a variety of environmental conditions.

Some mention should be made here that driving piezo stacks with very large currents (>1 A) may result in heating of the stack due to the electrical power requirements. When the stack reaches its Curie temperature (around 170 C) it will depolarize and stop working. Once it cools, you can usually repolarize the crystal by cycling it a few times from 0 to 150 V. Generally speaking I wouldn't recommend this type of high-temperature operation for extended periods of time since it affects the lifetime of the PZT stack.

13.7 RELIABILITY

Stack-flexure piezo stages have no parts that move against each other. As friction-free systems, the mechanics exhibit no wear or fatigue as long as the deformation of the metal is kept below its elastic limit. Commercial stages have been operational for more than 15 years of continuous use. The main mechanism of failure for a piezo stage is ion migration of the electrodes into the PZT which eventually causes a short circuit of the stack. This

effect is accelerated by three conditions which should be avoided if possible:

1. Constantly applied high voltage
2. High-humidity conditions
3. High-temperature conditions

I could put a graph in here showing MTBF for these conditions, but in my experience handling and sealing play a big role in the environmental factors and cycling away from 150 V a few times a day will do wonders for piezo health and well-being.

So be forewarned if you want to park your stack with greasy fingers at 150 V for weeks on end in the jungle!

13.8 TILTING STAGE DESIGN

There are a few varieties of tilt and tip-tilt stages commercially available. Usually they are temperature compensated for tilt, but not *z*. Since these are mainly used for laser scanning applications and are usually in direct competition with galvo type scanners, the design emphasis here is on high resonant frequency and stiffness. Often flexure hinges are dispensed with in the interest of speed. Three or four piezos are directly preloaded against a top plate and either act independently or in a push–pull configuration. For the three piezostack version, two tilting axes are arranged orthogonally to each other, while an optional third axis is oriented at 45°. A diagram of such a stage is shown in Figure 13.15.

X, Y, and Z shown as dashed lines above represent the tilting axes when the individual PZT stacks are actuated in the housing. When all three are given an equal applied voltage, the top plate will move in the *z* direction. Temperature changes will move the stage in *z*, but not affect the tilting angle. This particular stage design may offer tilting of 1 to 4 mrad with resonant frequencies in the kHz range. Disadvantages of this design are the location of the pivot point and the fact that it does not offer plus–minus tilting—each axis will tilt in only one direction.
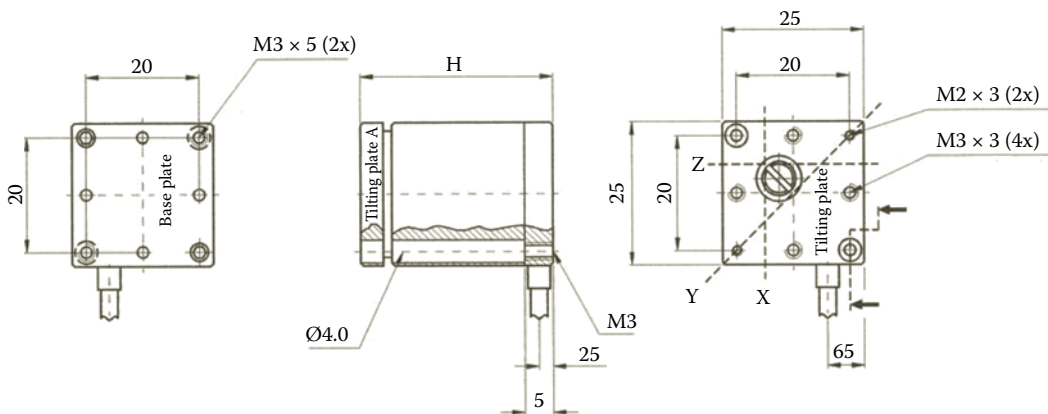
FIGURE 13.15
Three-axis mirror tilt scanning stage.

(a)

(b)

29.5

22.0

17.0

9.0

17.0

22.0

Ø25.0

4.5

M3-6H × 4 (4x)
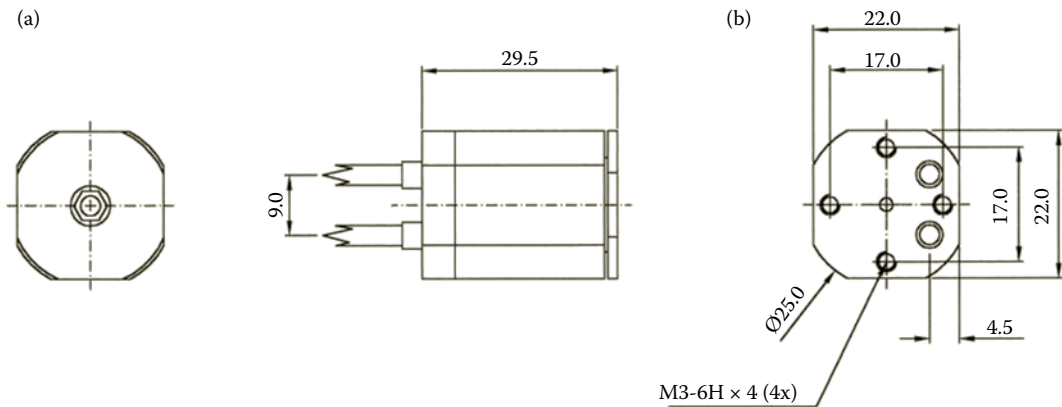
FIGURE 13.16
Two-axis mirror tilt scanning stage with push-pull design: (a) top (b) bottom.

When a tip-tilt stage with a central pivot point with plus–minus tilting is desired, a four stack push–pull approach is necessary. Once again, the stacks act directly against the top plate in order to maintain a high stiffness and high resonant frequency for the design. A diagram of such a stage is shown in Figure 13.16.

A larger tilting angle is possible with this push–pull approach and these stages are capable of up to 10 mrad (+/−5 mrad) of tilt angle with resonant frequencies in the kHz range.

13.9 LINEAR STAGE DESIGN

Today there is a demand for nanopositioners capable of a variety of specialized tasks. This wide variety of applications has resulted in a number of innovative new designs based on flexure techniques. Powerful computer algorithms utilizing FEA optimization have resulted in systems specialized for high resonant frequency under load, and minimized tilting and cross-talk errors over relatively large motions (large for a piezo stage is a mm).

13.9.1 Cross talk

Single dimension flexure-stack translation stages use a parallel guidance design. They use flexing monolithic solid metal hinges to guide and amplify the motion of the stack they contain. Usually the metal structure is designed so that they must be flexed slightly prior to mounting the piezo into them to provide an integrated preload and an epoxy is used to permanently bond the stack inside. A diagram and picture showing the design of such a structure and its associated motion are shown in Figures 13.17 and 13.18.

The motion in the diagram on the left has been exaggerated so that we can clearly see the problem here; in addition to the motion in the desired direction, we also have two other types of cross talk. There is a lateral component and a rotational component. A typical asymmetrical piezo flexure design such as this may have lateral cross talk on the order of 200 nm for a 100-μm stage. All four flexure hinges in the parallelogram design show this
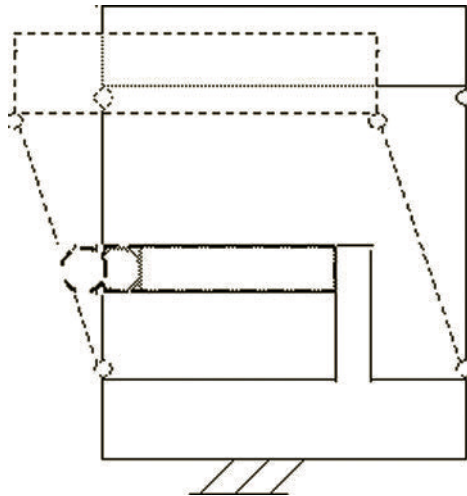
FIGURE 13.17
Guidance diagram for parallelogram stage.

FIGURE 13.18
Picture of a parallelogram stage F.

phenomenon. This is caused by a parallel deviation which is exerted around a rotational center. Also, as mentioned previously, the stacks themselves exhibit motion vectors in all six degrees of freedom. Although the flexure hinges direct these force vectors through very small areas, thus minimizing the impact of the unwanted motion, some forces from the stack operate outside these stress areas and contribute to additional cross talk not inherent in the design. Also a residual mechanical asymmetry caused by asymmetrical stack point of contact contributes and so the total rotational cross talk component is typically 15 µrad for 100 µm of motion. The major advantage of such a simple four hinge parallelogram guidance design is the low inertia in combination with a suitable guidance behavior; it is well suited for fast nanoscanning applications.

13.9.2 Minimizing Cross talk

It is possible to minimize this stack-induced cross talk by carefully selecting well-behaved stacks and actively aligning them with an interferometer prior to permanent epoxy bonding. This might reduce this type of cross talk by one-third, but cannot eliminate it altogether and as you might imagine, is time consuming and expensive.

By applying meander ordered hinges we obtain an elegant solution to the problem of cross talk by allowing the hinges to compensate each other. Such a system is shown in Figure 13.19.

Of course there are tradeoffs here for improved performance. In this case, the number of flexure hinges has been increased from 4 to 16. This significantly weakens the mechanical structure of the stage. Stiffness has now dropped dramatically and therefore results in significant reductions in resonant frequency for any laterally applied load to the system. Reduced resonant frequency means higher noise characteristics since a lower resonant frequency is more easily excited by ambient noise. However, for applications requiring highly accurate nanopositioning with light loads this solution works well and is free of parasitic cross talk.

13.9.3 Increasing stiffness

So how can we obtain the benefits of a symmetrical dual flexure structure and maintain our excellent noise properties?

One weakness of parallelogram flexure structure is that piezo stack provides a strong compressive force and stiffness, but we rely on the weak spring memory forces of the metal for our reset forces. One way to increase the reset forces of the flexure structure would be to increase the thickness, for example, reset force of the flexure hinge. This has the effect of making the entire structure more stiff and, unfortunately, since the desired DOF is blocked this increases parasitic rotational errors. Another way to increase the stiffness of the stage is to apply a push–pull concept to the design. We integrate two stack actuators into a flexure structure in such a way that they act in concert with each other and provide both compressive and reset forces to the structure for the motion of the stage. For this design concept we can increase stiffness by up to ten times in the intended direction of motion and by up to two times for the whole system. Additionally the overdetermination of this design results in a high degree of inner preload that makes the stage much more robust against mishandling, off-center loads and high loads. All featured without increasing rotational error as per Figure 13.20.

A main advantage of such a push–pull concept is the separation of guidance and gearing. As opposed to traditional designs, the flexure hinges are exclusively responsible for trajectory tasks; they are not responsible for exerting reset forces. Also, the stacks, when
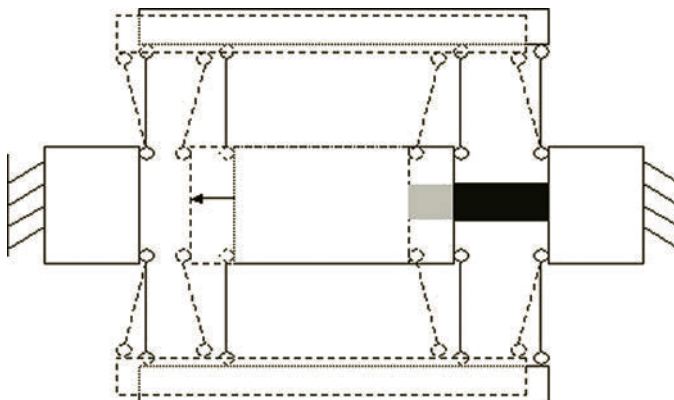
FIGURE 13.19
Guidance diagram using meander ordered hinges.
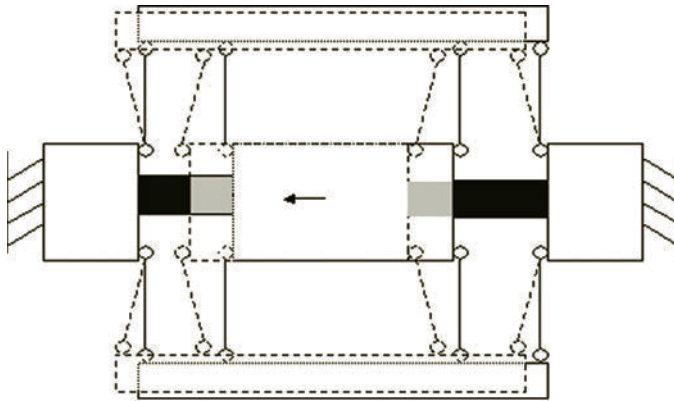
FIGURE 13.20
Guidance diagram of a push–pull drive mechanism.

arranged in alternating order, are able to accelerate and stop actively taking advantage of their huge pressure force potential. So this design offers a maximum of dynamical and guidance behavior. It is inherently temperature compensated and well suited for nanopositioning tasks as well as nanoscanning applications.

13.10 DAMPING

You can abstract a piezoactuator as a linear oscillating system consisting of spring, damper, and inertial mass, following the normalized mathematical description (force equilibration) (Figures 13.21 and 13.22):

$$kx + c\frac{dx}{dt} + m\frac{d^2x}{dt^2} = F_d(t) = 1$$

k = spring constant
x = stroke
c = damping coefficient
$\dfrac{dx}{dt} = \overset{\circ}{x}$ = velocity
m = inertial mass
$\dfrac{d^2x}{dt^2} = \overset{\circ\circ}{x}$ = acceleration
F_d = disturbance (force)
t = time

Solving the differential equation in respect to the initial values

$$x(t = 0) = 0$$

and

$$x(t = 0) = 0$$
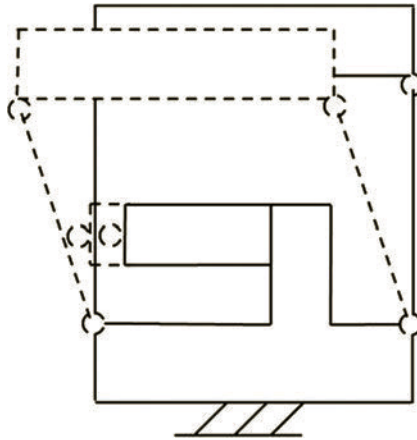
FIGURE 13.21
Schematic piezoactuator.
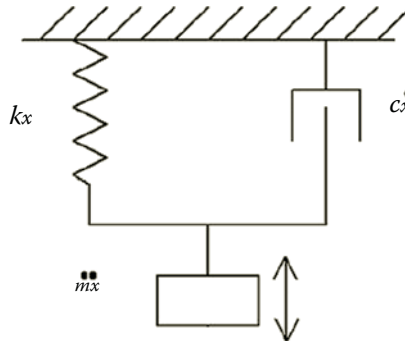
FIGURE 13.22
Analytic abstraction.

you obtain the response of the oscillating system to the external disturbing force (Figures 13.23 and 13.24). Depending on the damping ratio the behavior is described as:

undamped	$D = 0,$
underdamped	$0 < D < 1,$
critically damped	$D = 1,$
overdamped	$D > 1$

with $D = \dfrac{c}{2\sqrt{km}}$

D = damping ratio

An increasing damping ratio decreases the resonance magnification and sharpness. That means the resonance appears at a lower frequency but its suppression is much higher.[1–3]

Because of the marginal electrical losses of the piezoceramic that could reduce kinetic energy from the system by transforming into heat, in practice the underdamped piezoactuator is the usual one (Figure 13.25). The transient oscillation is described as

$$x(t) = 1 - e^{-Dw_0 t}\left[\cos wt + \frac{Dw_0}{w}\sin wt\right]$$
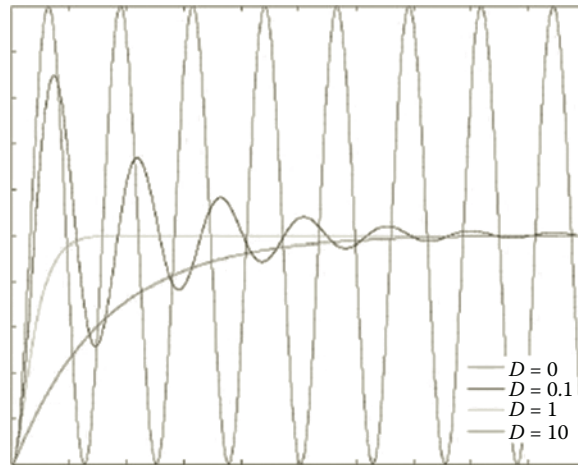
FIGURE 13.23
Disturbance response–time domain (amplitude vs. time) showing undamped, underdamped, critical damped, overdamped. The fastest is the critical damped case.
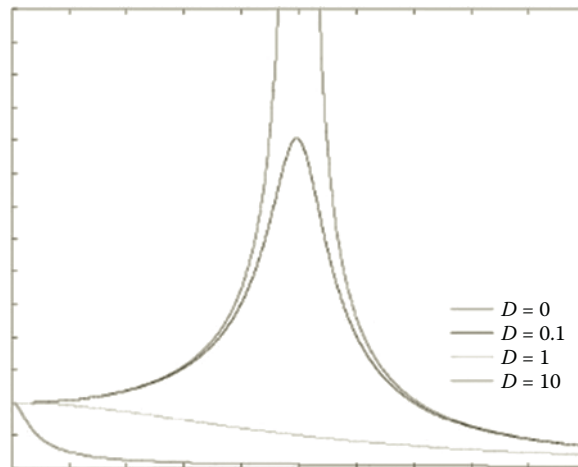
FIGURE 13.24
Disturbance response–frequency domain (magnitude vs. frequency) showing undamped, underdamped, critical damped, overdamped. Listed largest to smallest amplitude.

with $w = w_0 \sqrt{1 - D^2}$

ω = underdamped frequency
ω_0 = eigenfrequency (w/o damping)
and $x_{envelope}(t) = 1 \pm e^{-w_0 Dt}$
$x_{envelope}$ = envelope
describes the turning points (envelope).[1,3,4]

As said the damping behavior of a piezoactuator is marginal, well suited for ultrafast scanning applications and small loads. However, fast high load applications using the proprietary outstanding stiffness of the piezoceramic can be optimized by implementation of fitted passive viscoelastic damping parts.
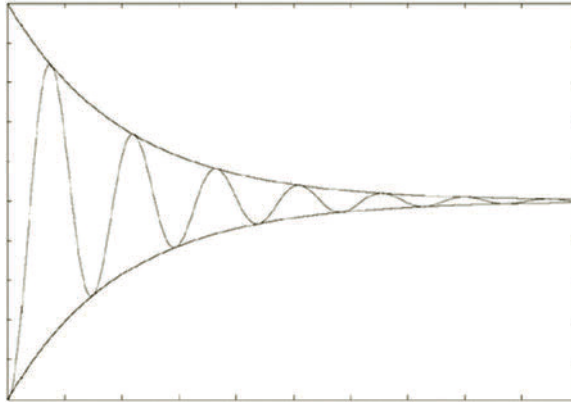
FIGURE 13.25
Amplitude vs. time graph: underdamped oscillation and its envelope.

The critical damping (aperiodic case) shows the shortest settling time and no overshooting, but the damping ratio depends on the inertial mass. Therefore you are able to optimize for one load configuration only. Higher loads will settle much faster than in the undamped case otherwise small load scenarios will behave unnecessarily slow. So it's recommended for versatile and robust use that the stage works in the underdamped range, close to the critically damped case.

But, how do you determine a well damped setup? For our assessment we use the damping ratio, that you can determine in such a way:

1. First, you generate a rectangular wave. Due to the fact that such a waveform consists of odd-numbered multiples of the fundamental frequency, the piezoactuator's resonances will be excited according to the following formula

$$x_r(t) = \frac{4h}{p} \sum_{k=1}^{\infty} \frac{\sin\left[(2k-1)w_f t\right]}{2k-1}; \quad k = 1..\infty;$$

   x_r = current stroke of rectangular wave
   h = degree of amplitude
   v_f = fundamental frequency, depending on the bandwidth of the function generator and amplifier

2. Next, you modulate the amplifier output signal and the piezoactuator oscillates in its resonances.

3. Transforming the oscillation into a frequency spectrum via, for example, fast Fourier transformation (FFT) allows you to determine the damping ratio per for example bandwidth method (Figure 13.26).

As mentioned before damping ratio and resonance sharpness correlate to each other. So you can calculate the damping ratio using the equation

$$D = \frac{f_2 - f_1}{2 f_r}$$
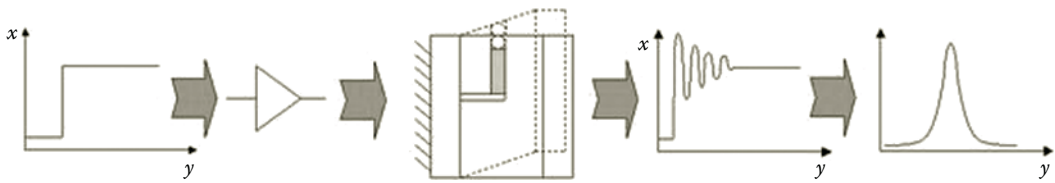
FIGURE 13.26
Scheme of determining the damping ratio of a piezoactuator.

f_1 = lower frequency @ $m_{max}/\sqrt{2}$
f_1 = upper frequency @ $m_{max}/\sqrt{2}$
f_r = resonant frequency
m_{max} = resonant magnitude

The longer the measured time (real or zero padded), the higher the frequency resolution and more accurate the resulting damping ratio calculation (Figures 13.27 through 13.30). Following the sample theorem (SHANNON/NYQUIST), the sample frequency has to be at least two times of the frequency that is to be measured.[4]

Using viscoelastic damping in open loop order offers a lot of advantages. Firstly, the settling time decreases significantly. The piezoactuator reacts much more robustly against high loads and dynamic forces, as well as mishandling and environmental oscillations. Also, the positioning noise caused by the noise of the amplifier is deeply suppressed. So you can perform subnanometer step scans in much smaller stepwidth, even at higher loads. However, the viscoelastic and piezoelectric intrinsic drift accumulates. Using viscoelastic damping in closed loop, the drift doesn't exist.[5,6]

13.11 CLOSED LOOP SYSTEMS

We've discussed a number of problems inherent to piezo technology for motion control. These include drift or creep, hysteresis, temperature dependence, nonlinear expansion, variance among individual stacks, cross talk, and rotational error. The simplest way to address many of these simultaneously is to characterize the piezosystem with an integrated measurement device and integrate a closed loop feedback mechanism to control the position. In order to take advantage of the piezosystem's capability for high resolution in the nm range and high dynamic capabilities it is important to select a type of sensor which is fast with nanometer accuracy. Clear choices that satisfy these requirements are capacitive sensors, strain gage sensors, inductive (LVDT) sensors, optical scales, and interferometric measurements. In order to somewhat limit the scope of the chapter we will concentrate on the use of capacitive sensors and strain gages in PID closed loop systems since these will be useful for a majority of applications which use piezos.

13.12 STRAIN GAGES

Strain gages can be applied directly onto a stack of PZT to measure strain that has been produced by the expansion of the PZT or onto the flexing hinges of the flexure structure.
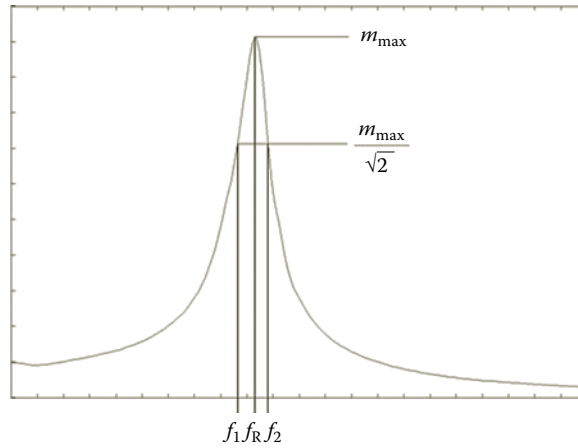
FIGURE 13.27
Magnitude versus frequency graph: determination of damping ratio *via bandwidthmethod.*
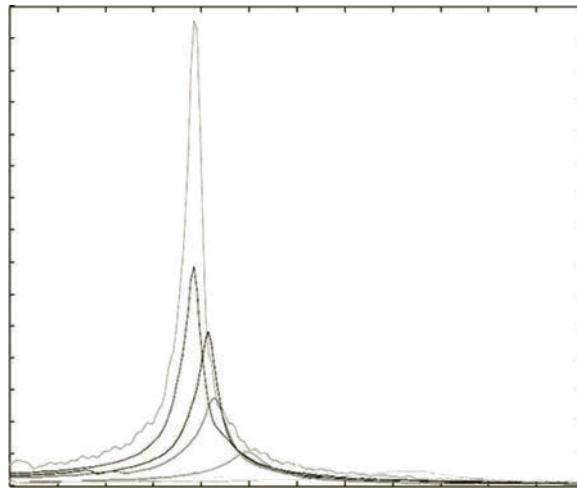
FIGURE 13.28
Disturbance response–time domain (amplitude vs. time): increasing damping ratio (left to right).

Since the expansion and therefore strain of a PZT stack are inhomogeneous and the strain gage is applied to just a small part of a stack, it is much more accurate (usually at least 2×) to apply the strain gage to the flexure hinges—usually more than one. This effectively makes a measurement summation of all motion "prior" to the flex hinge, while ignoring any motion which occurs "after" the hinge. Therefore all PZT motions including temperature-dependent ones are measured, but the strain gage system will not compensate for expansion of the top plate of a stage. Special care must be taken for a proper bond to the PZT or metal so that adequate heat transfer occurs to minimize Johnson–Nyquist noise since strain gages are resistive devices. A variety of strain gages are commercially available which are specifically tailored to various material expansion coefficients and the proper gage must be selected for the stage material (i.e., steel, aluminum, INVAR). A typical strain gage is either a full or half wheatstone bridge with 5 KOhm resistors arranged as in Figure 13.31.
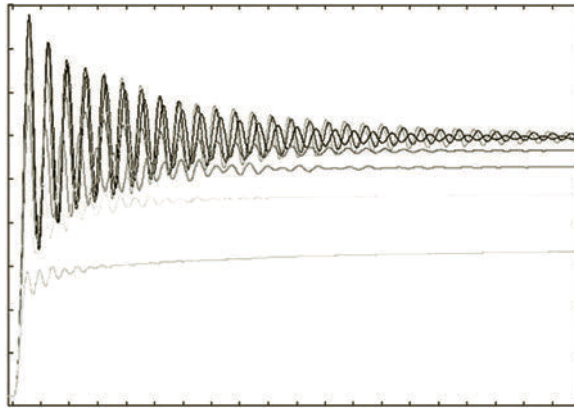
FIGURE 13.29
Disturbance response in open loop–time domain (amplitude vs. time): increasing damping ratio (top to bottom).
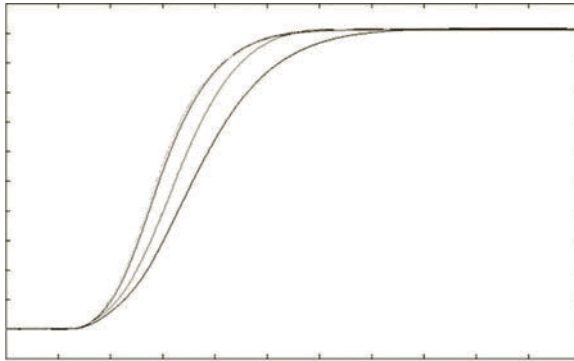
FIGURE 13.30
Disturbance response in closed loop–time domain (amplitude vs. time): increasing damping ratio (right to left).
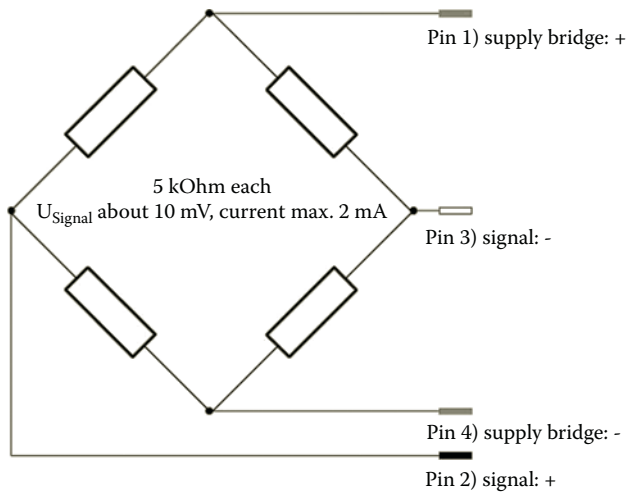
FIGURE 13.31
Electrical diagram of a strain gage.

Linearity-linearity
Max nonlinearity: Absolute = 46 nm   Relative: 0.058%
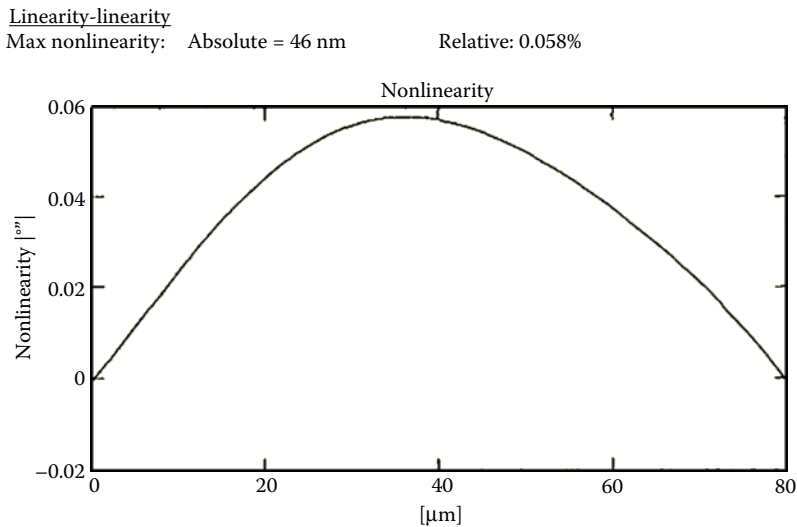
FIGURE 13.32
Nonlinearity of a strain gage sensor.

With a full wheatstone bridge expansion or contraction due to temperature variations do not affect the output signal of the sensor since all resistors will be affected equally. With a half-wheatstone bridge, expansion or contraction due to temperature variations will be measurable. I've run into very few situations where a half bridge is desirable. It's important to make this distinction since both designs are used for closed loop feedback in commercially available piezosystems and you should be aware of the type of system you're using. The typical safe bending radius for such a gage is about 1.5 mm which is far above the elastic limit of steel. Often a flexure can be damaged by applying excessive force and, as a result, it is bent beyond the elastic limit of the metal such that the strain gage still functions, but is permanently deformed so that the strain gage shows some permanent DC offset. In some cases it is possible to repair the stage by measuring the output of the strain gage while bending the flexure back until the DC offset is minimized and then performing a recalibration of the closed loop system. The response of a strain gage to motion is fairly linear. Figure 13.32 shows the voltage response of a strain gage when a flexure is actuated over a distance of 80 μm. The distance traveled was measured with an interferometer.

Often, nonlinearity curves follow parabolic functions and it is a fairly common practice to correct for these errors using polynomial fit algorithms or simple look-up tables.

13.13 CAPACITIVE SENSORS

Capacitive sensors use a charged flat metal plate of precisely determined area to make a noncontact measurement of distance based on formula for parallel plate capacitors. Although it sounds simple in theory, in practice this type of measurement can be quite complex due to the very small changes in capacitance that must be measured in order to achieve nm resolutions. Complex electronics must be modulated at specific frequencies

for various bandwidths and resolutions that are desired. Since changes in capacitances being measured are nearly on the same order of magnitude as cable capacitances, specially calibrated cables and connectors must be used. It is important to keep this in mind when selecting a capacitive measurement system for closed loop. Often it is not possible to change cable lengths or connectors for your system if the capacitive sensor electronics are built directly into the closed loop amplifier system. On the plus side, capacitive sensors offer the following advantages over strain gage systems: lower noise, higher bandwidth capability, and better linearity. Also, it is possible to arrange capacitive sensors such that they measure directly on the point of interest for the stage—or at least closer to it. Cross talk and temperature-dependent motion not measured by a strain gage are now measurable and correctable by a capacitive measurement system.

13.14 ELECTRONIC CONTROL ARCHITECTURE FOR CLOSED LOOP SYSTEMS

A simple diagram of a typical closed loop system is as shown in Figure 13.33.

Closed loop systems can be made completely analog, digital, or as a hybrid analog–digital system (for instance an analog PID control system with digital potentiometers for programmability).

Variables which must be calibrated based on the individual system are as follows:

1. Gain of the sensor
2. Offset of the sensor
3. Gain of the measurement interfaces (e.g., 0–10 V analog output, digital display, A/D interface for serial port)
4. Total travel of the system over the full voltage range (gain of the positioning system)
5. Gain parameters of the P, I, and D components of the system
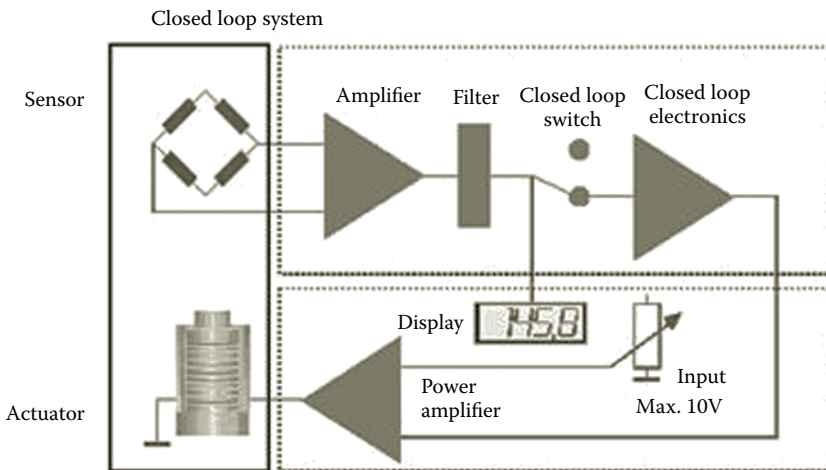
FIGURE 13.33
Closed loop system architecture.

Let's take the example of a stage that specifies its total open loop travel as 100 μm under light load conditions. Since, as mentioned previously, PZT stacks can vary quite a bit the actual motion from such a stage could vary as much as 20 μm. For this example let's assume the stage is capable of 110 μm for a voltage range of –10 V to 150 V. This intrinsic gain of the system will determine a number of factors in the final closed loop system since linearity, resolution, and repeatability are some function of the total motion amplification. Often, manufacturers will guarantee and characterize their system specifications as some percentage of total measured motion of the system in closed loop (i.e., a resolution of 0.05% for this system in closed loop would be 4 nm) even though it may actually be determined by the total motion the system is capable of making in open loop. In other words, taking a 400 μm stage and calibrating it for 80 μm of motion in closed loop instead of 360 μm will not increase the resolution of your system.

As we saw before, a piezosystem will overshoot and oscillate at some resonant frequency. Closed loop systems have the job of damping and controlling this behavior, but some overshoot and settling time, especially for step functions, is unavoidable. In order to allow for this behavior, closed loop systems are calibrated within a range of 80% of the full motion allowing 10% at the extremes of the travel range for overshooting (Figure 13.34). Since piezos vary from system to system this range is often further reduced to allow for PZT stacks that may have less travel than others. For this particular case a travel of 80 μm would be calibrated into the system since the minimum specified travel range is 100 μm.

The speed of a closed loop system is determined by the loading and resonant frequency of the system. Introducing a high load to a system calibrated for light loads may cause it to become underdamped and it may begin to oscillate. The PID can be adjusted in such cases to damp out the oscillation, but the rise time will suffer as a result. In some cases noise due to oscillation of an underdamped system is tolerated in the interest of speed. As an example, a stage which moves 100 μm in open loop calibrated for 80 μm in closed loop is loaded with a microscope objective which weighs 300 g. At rest the PID closed loop begins to oscillate at a frequency of about 130 Hz as shown in Figure 13.35.

Rise time for the system is about 11 ms for a step of 0.6 μm and P_P noise is about 6 nm. We can eliminate the oscillation by overdamping the system with the PID closed loop with results as in Figure 13.36.

As you can see the problem with oscillation has been reduced by a factor of 3, but the settling time to a final position has been extended from 11 ms up to about 25–30 ms.

Final tuning of a closed loop system is specific to the particular application. In this case, which is optical microscopy, the depth of field of this particular microscope objective was well in excess of the 6 nm and settling time was of primary importance. Another
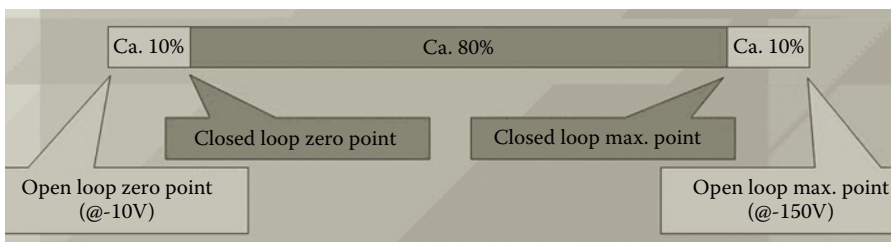
FIGURE 13.34
Closed loop versus open loop travel range of a piezo stage.
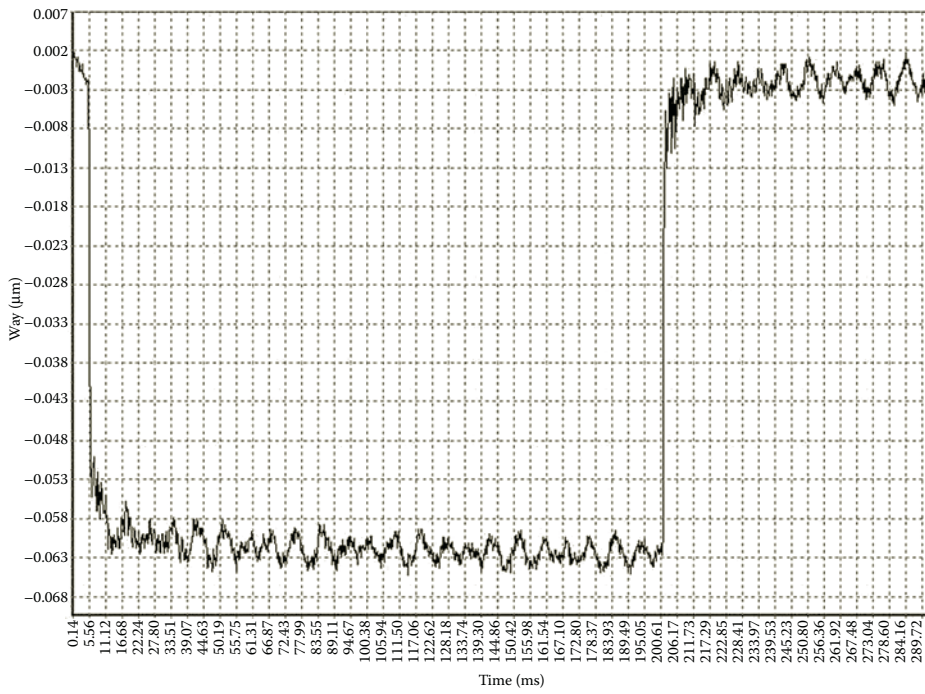
FIGURE 13.35
Closed loop system subject to additional loading.
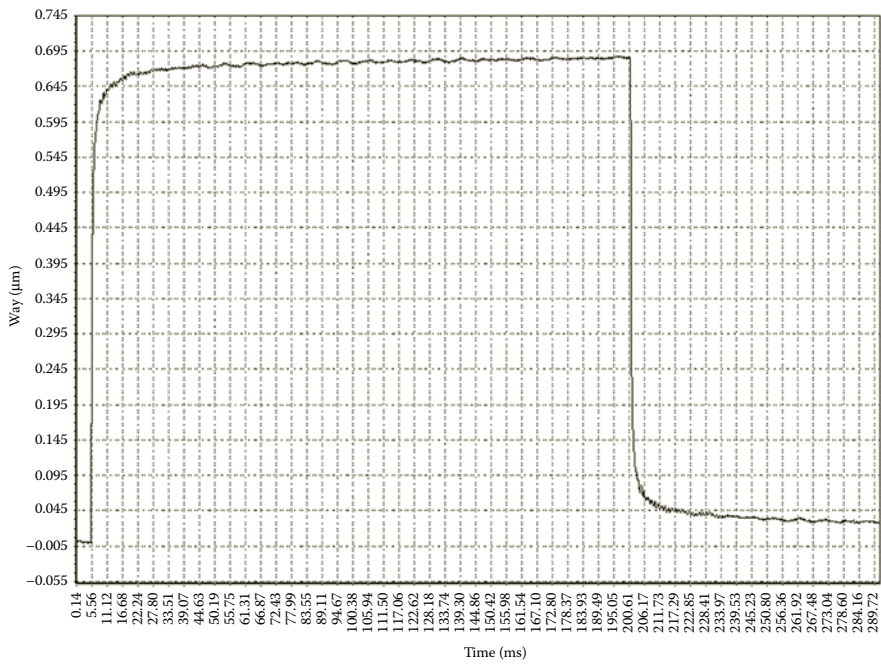
FIGURE 13.36
Closed loop system after tuning for additional load and minimum noise.

application (i.e., fine tuning the grating position for structured illumination) might require the lowest possible noise with no speed requirement. In most cases manufacturers can tune the system for a particular loading configuration, motion waveform, and desired resolution or speed.

Both types of sensors—strain gage and capacitive—have properties which will be unique to the stage system for which they are being used. Every strain gage and capacitive sensor system will have different gain and offset parameters even for identical stage systems. For strain gages these parameters are determined by resistance tolerances and prestressing of the metal flexure. Capacitive sensors are affected by offset distance and angular alignment.

For OEM systems this is often problematic since parts of a system (i.e., stage, cables, electronics) cannot be interchanged and any failure in one part of the system would necessitate complete replacement of all parts or at the very least, a recalibration for gain and offset parameters by an interferometer. Therefore modern piezoelectric closed loop systems offer integrated precalibrated stage systems with some calibration data stored in the stage or stage-cable assembly. This may be as simple as using potentiometers to regulate the voltage outputs to more complete data-chip storage systems which include PID parameters for the closed loop system. The two architectures discussed are more easily understood with the following diagrams:

Config. 1.

At motion site   At control site

Stage -------------------- voltage cable --------------------- Amplifiers+

+   PID+   Calibration Data

Sensor ------------------- sensor signal cable -------------- conditioning electronics

Config. 2.

At motion site   At control site

Stage

+

Sensor

+

Conditioning Electronics -------------------- Amplifiers + PID

+

Calibration Data for PID
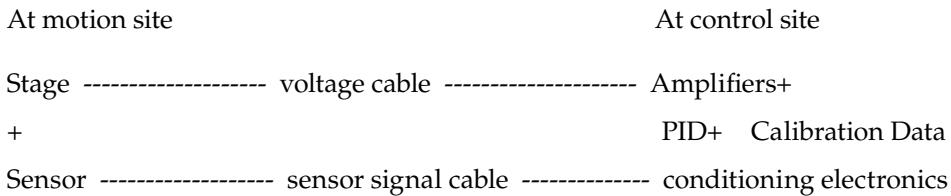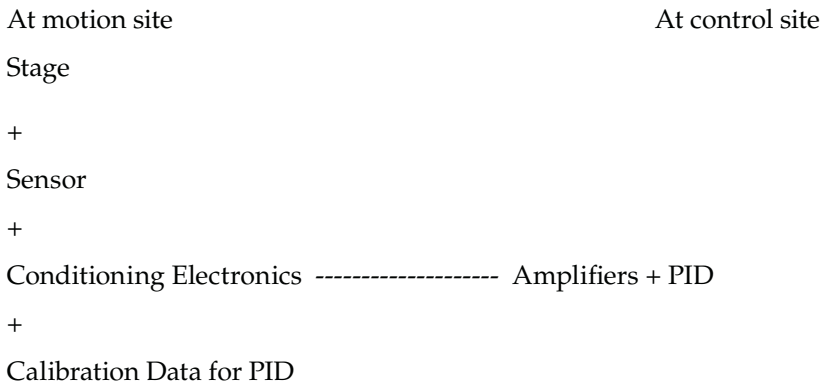
Config. 2 offers the advantage of the ability to replace the stage system or electronics without having to have the complete system for recalibration.

13.15 CONCLUSION

Piezoelectric flexure systems have unique properties that require special consideration when implementing them into a design. Specifically special attention should be given to hysteresis, drift, temperature influences, damping, and electrical drive considerations. They also offer numerous advantages over traditional motorized linear drives. These include:

High stiffness

High structural resonant frequencies

Sub-nm resolution

High speed

High load capacities

Large force generation

These advantages have made them indispensible in applications that require ultraprecise positioning accuracies. Some examples follow:

Configuration	Application
Stack type and ring actuators	Valve control, laser cavity tuning
Single-axis stages with mounting platforms	Rapid confocal microscopy Z-stack acquisition, high resolution focusing devices, grating positioning for structured illumination microscopy
Open frame x–y stages with large central apertures	Optical microscopy, atomic force microscopy, scanning electron microscopy
Compact multidimensional translation stages	Laser trapping and cooling techniques, optical tweezers, fiber optic alignment, storage device head alignment, CCD chip resolution enhancement, MEMS and microfluidic alignment tools
High speed scanning stages	Rapid prototyping machines
Mirror tilting systems	Laser alignment, interferometric systems
Piezo actuated slit apertures	Scanning electron microscope, proton beam, and laser aperture control, spectroscopy
Piezo actuated grippers	Cleanroom pick and place devices

As the need for smaller and smaller system continues at a rapid pace, the need for piezoelectric positioning solutions continues to grow.

REFERENCES

1. Schmidt, R.; Waller, H. *Schwingungslehre für Ingenieure—Theorie, Simulation, Anwendungen*; Wissenschaftsverlag Mannheim/Wien/Zürich, 1989.
2. Götz, B.; Martin, T.; Duparre, J.W.; Bücker, P. *Theoretische und experimentelle Untersuchung relevanter Parameter von Piezoaktoren*; Technischer Report, Piezosystem Jena, 1998.
3. Wittenburg, J. *Schwingungslehre*; Springer Verlag 1996.

4. Borchhardt, G.; Wehrsdorfer, E.; Karthe, W.; Hertsch, P.; Höfer, B. Displacement amplification mechanism for dynamic use; Technical report, Fraunhofer Institute of Applied Optics and Precision Engineering, 1998.
5. Müller, R. Verbesserung des Einschwingverhaltens wegübersetzter piezoelektrischer Aktoren durch Optimierungder mechanischen, dynamischen Parameter; degree dissertation FH-Jena-University of Applied Sciences, Piezosystem Jena, Jena, 2005.
6. Lorenz, M. Optimierung des Einschwingverhaltens piezogetriebener Einachsenmikropositionier-tische mittels Integration eines passiven Dämpfungsgliedes; degree dissertation FH-Jena-University of Applied Sciences, Piezosystem Jena, Jena, 2006.

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

14

Optical Disk Scanning Technology

Tetsuo Saimi

Matsushita Electric Industrial Co., Ltd.
Kadoma, Osaka, Japan

CONTENTS

14.1 INTRODUCTION

The aim of this chapter is to describe important aspects of optical disk recording and readout technologies, with a brief historical introduction and references for further study. The selected topics are based on the contemporary analysis and experimental results of general interest.

14.1.1 Progress in Optical Disk Technology

The fundamental concept of an optical disk dates back to 1961 when Stanford Research Laboratories developed a video disk using photographic technology. However, the low luminance of available light sources yielded reproduced images of low quality. Columbia Broadcasting System (CBS) announced the EVR (Electronic Video Recorder) system in 1967, but enormous costs ultimately forced them to discontinue development. The invention of the laser by T. H. Maiman et al. in 1960 provided the light source considered the most suitable for optical disks.

Lasers have good temporal and spatial coherence, which enables one to obtain the small, diffraction-limited beam spot necessary for high-quality information retrieval from optical disks. After many approaches were considered, the basic design of optical disks, the "bit-by-bit" recording method, was developed in the 1970s. The first optical video disk system for commercial use, the VLP (video long play), was released in 1973 by Philips of Holland and MCA (Music Corporation of America) of the United States. In early systems, the He–Ne laser was the preferred light source. The introduction of many new optical disk systems soon followed. The 12-cm diameter digital audio disk (DAD), later called the CD (compact disk), was announced in 1978. Standardized CD products from several

manufacturers became available in December 1982. CD players use semiconductor lasers to allow the design of small and lightweight players. In 1996, the digital versatile disk (DVD) for players was released. These playback-only systems marked the inception of optical disk products. Write-once optical disk systems were first introduced by Philips in 1978.

Development of rewritable optical disk systems accelerated in the 1980s as the performance of reversible media progressed. Magneto-optical (MO) disks that utilize a magnetic field reversal for recording and the Kerr effect for playback were commercialized in 1988 by Sony. In 1989, the first phase-change rewritable (PCR) disk, containing 470 megabytes user capacity and utilizing an amorphous-to-crystalline phase change[1] for recording and playback was commercialized by Matsushita. In 2000, rewritable DVDs (DVD-RAM, -RW) were released and the development for higher density DVD media started.

14.1.2 Characteristics of Optical Disks

Optical disks are now used in various applications, including audio, computer memory devices, picture files, document files, and video files. The advantages of optical disks over other known memory devices are:

1. Large capacity/high information density. The information capacity of a 120-mm diameter DVD disk is 4.7 Gbytes for single-layer ROM and RAM, and 8.5 Gbytes for double-layer ROM. The recording density of commercial products is about 3.3 Gbits/in^2 for DVD-ROM and 4 Gbits/in^2 for DVD-RAM. Recent developments for next generation products show that an information density of more than 16 Gbits/in^2 can be achieved by using a blue laser and an objective lens (OB) of high numerical aperture (NA).

2. Fast random-access library systems allow access to large mass memories. Changing mechanisms provide access within seconds to several petabytes of information.

3. Reliability. The information surface of an optical disk is covered with a protective layer, which ensures a long archival life. Information retrieval is achieved without physical contact between the optical pick-up and disk, which increases the reliability of stored information.

4. Replication. Mass production using injection molding or other high-volume techniques is possible. Replicated optical disks benefit from lower cost per bit than the rigid magnetic disk or tapes.

5. Removability/ROM-RAM compatibility. A large quantity of data can be handled easily by exchanging disks. Compatibility between replicated and recordable disks and interchangeability between standardized drives provide this capability. These advantages lead to the ubiquitous uses of optical disk products in consumer and computer applications.

14.1.3 Principles of Optical Read/Write[2–6]

In many optical disks, as in the normal audio disk, information is recorded in a spiral groove referred to as the "track." The information cells shown in Figure 14.1 are called "pits," or "marks." They are discontinuous small depressions, differential reflectivity patterns or phase-shifting patterns, all showing differential reflectivity. Information signals (SGs)are derived from changes in luminance caused by diffraction of the laser beam by the pits or marks (which are about 0.3-μm^2 diffraction cells). The laser beam emerging from
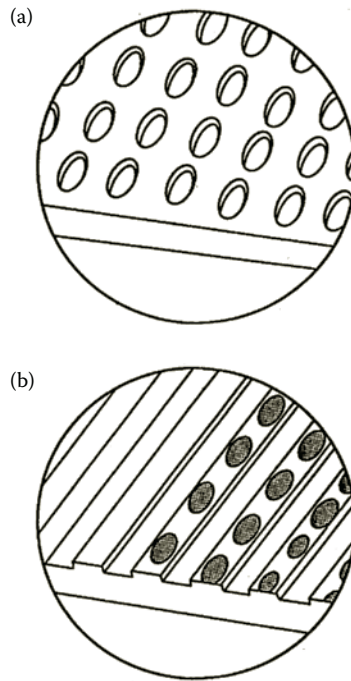
FIGURE 14.1
Pit patterns: (a) phase pit, (b) amplitude pit.

the OB is focused to a spot on the disk. The spot size is proportional to the wavelength λ of the laser beam and inversely proportional to the NA of the OB.

The NA is given by the sine of the angle θ between the optical axis and the marginal rays:

$$NA = n \sin q \tag{14.1}$$

where n represents the refractive index of the medium in object space. The full-width-at-half-maximum (FWHM) intensity diameter of the beam spot (D_s) on the disk is expressed as

$$D_s = k\frac{1}{Na} \tag{14.2}$$

where k represents a constant dependent upon the light amplitude distribution at the OB pupil. If the incident beam to the OB is plane wave, the value of k is 0.53. When the incident beam is Gaussian or contains some aberrations, k becomes larger. Since the information density of the disk is inversely proportional to the square of D_s, smaller k is more desirable. Supposing $k = 0.53$, $\lambda = 0.405$ μm (15.9 μin), and NA = 0.85, we obtain the beam spot diameter $D_s = 0.25$ μm (9.9 μin). More than 1.4×10^{11} information bits can be stored on one side of a 5.25 in optical disk using a beam of this size.

Figure 14.2 shows the playback optics for a reflective optical disk. The laser emission from the semiconductor laser (L) is reflected by a beam splitter (BS) and is incident on the
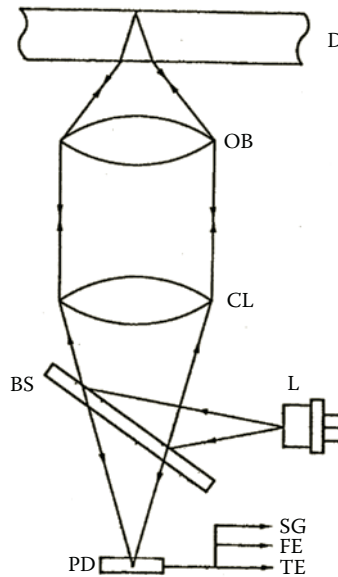
FIGURE 14.2
Playback optics.

OB through a collimating lens (CL). The wavelength λ generally used for CD is in the range 780–800 nm and for DVD, 635–660 nm. The NA of the OB is generally 0.45 when used for CD and 0.6 for DVD. Next generation optical disks will have a wavelength of 405 nm and a NA of 0.85. The OB aperture limits the spatial frequency response of the optical system.

The laser beam reflected from the disk is intensity modulated by the pits prior to a second pass through the OB. Part of the return beam is transmitted through the BS and is incident on the photodetector (PD). The SG, focus error signal (FE), and tracking error signal (TE) are generated from the PDs.

Reflective or transmissive mode systems can be constructed, but the reflective mode is used in most optical disk systems. In the transmissive mode, a second optical pick-up with the PDs must be positioned on the other side of the disk, complicating the design of the drive. Another problem is that the pits must be very deep and replication becomes more difficult, which leads to degradation of signals during read. A third problem with the transmission mode is difficulty in obtaining a good FE with a satisfactory S/N ratio. Simple FE detection methods are easily achieved in reflective mode.

14.2 APPLICATIONS OF OPTICAL DISK SYSTEMS

14.2.1 Read-Only Optical Disk Systems

Four types of standardized players are available for read-only optical disks: video disk, audio disk (CD), data file disk (CD-ROM), and DVD. Among the advantages of read-only optical disks are: (1) mass replication; (2) a relatively simple optical layout as compared to write-once or rewritable system; and (3) ease of commercialization due to its use as a

stand-alone unit. Signals in read-only optical disk systems are generally encoded with pulse width modulation (PWM), resulting in high recording density.

14.2.1.1 *Video Disk*

Optical video disk systems have been commercially available for many years. They have the international standard name of LV (laser vision), and use analog signal recording. Two types of disks having diameters of 30 and 20 cm are utilized. Rotational speed is constant at 1800 rpm for the constant-angular-velocity (CAV) mode, and a variable rotational speed of 600–1800 rpm is used for the constant-linear-velocity (CLV) mode. LV has relatively low recording density and is now being replaced by DVD.

14.2.1.2 *CD/CD-ROM*

The DAD system has been standardized using the term compact disk. The diameter of the disk is 12 cm (4.7 in), and the thickness of the polycarbonate protective layer is 1.2 mm. The linear velocity can vary from 1.2 to 1.4 m/s (3.9–4.6 ft/s) in the CLV mode. The maximum playback time is about 75 min, long enough to accommodate a fairly long classical music selection on a single disk. Audio signals are quantized using 16 bits, allowing a dynamic range of 96 dB in playback. CD is now dominant in package media and is widely used in the music market. Using the ordinary data signal coding of the CD system, computer compatible data can be stored for use as a read-only memory of a personal computer (CD-ROM). More than 650 Mbytes of information can be stored on one side of a disk, enough to store the entire text of *Encyclopedia Britannica* on one side of a CD-ROM disk. Most personal computers are equipped with a CD-ROM, although DVD drives are rapidly replacing the CD drives in this application.

14.2.1.3 **DVD**

The successor to the CD is the integrated DVD optical disk system. The DVD specifications for Read-Only Disk were issued in 1996, followed by Rewritable (ver. 1.0) in 1997, Rewritable (ver. 2.0), and Re-recording (ver. 1.0) in 1999, and DVD-R for General (ver. 2.0) in 2000. These DVD systems are all integrated into DVD for Multi. The DVD disk has a storage capacity of 4.7 Gbytes, and more than 135 min of MPEG2 video signal can be stored on one side of a 12-cm disk. The DVD has overtaken almost all use of LD and CDV in video and music videos.

14.2.2 **Write-Once Disk Systems[7–9]**

Write-once disks have been commercially implemented in applications such as archival data memory devices for computers, document storage, and picture filing systems. Polycarbonate (PC) is being used to form injection molded disk substrates. The recording mechanism of the disk may be (1) phase-changing; (2) hole-burning; and (3) bubble-forming. In Figure 14.3, the pits formed by these different recording methods are schematically shown.

Signal pits are recorded by irradiation with a semiconductor laser focused to a spot less than 0.3 up to 1 μm in diameter. This irradiation increases the temperature of the recording medium to about 200–600 °C (392–1272 °F), and the recording takes place as the result of the consequent physical or chemical change of the medium.
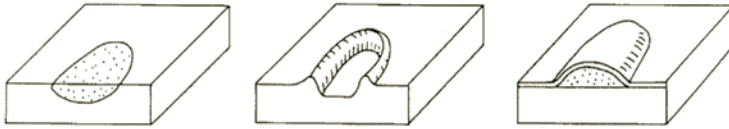
FIGURE 14.3
Pits formed by different recording methods.

14.2.2.1 CD-R

Write-once disk systems can be used for archival storage of large data files. The removability of optical disks and the standardization of products provide a broad range of application. CD-R is currently the most commonly used write-once disk. Typical specifications of CD-R disk systems are shown in Table 14.1.

14.2.3 Erasable Optical Disk Systems

Two major families of erasable media are available: PCR and MO. Data recording on PCR media is accomplished by inducing a transition from a crystalline phase to an amorphous phase. Differences in reflectivity of the two phases allow signal playback.

MO recording is accomplished by establishing the magnetization of a mark by heating it in the presence of a magnetic field. Read back utilizes the polarization change of the laser beam induced by magnetic modulation according to the Kerr effect. The principal characteristics of erasable disks are shown in Table 14.2. The overwrite mechanism of a PCR disk is easy to design. However, reversibility is better in MO disks. The MO disk drive requires a complicated system for applying the write and erase magnetic fields, which have opposite polarities.

14.2.3.1 PCR Disk[7]

Figure 14.4 shows PCR optical data file drives (DVD-RAM). Figure 14.5 shows the principle of the direct overwriting mechanism for the PCR disk. The laser intensity at the disk is modulated, in correspondence with the pit pattern to be recorded, between the maximum level (A) and the intermediate level (B) as illustrated in Figure 14.5a. At exposure level (A), the material reaches a melting temperature of over 600 °C. Rapid quenching forces the material to remain in the amorphous phase, giving low surface reflectivity. Exposure level (B) heats the material to about 400 °C, allowing rapid crystallization to proceed, giving an increased reflectivity. Figure 14.5b is a schematic illustration of the overwriting operation.

14.2.3.2 MO Disk[10,11]

In MO disks, a light beam is directed at a magnetic material to record or erase information. The underlying principle is the utilization of a temperature-dependent change of magnetic properties. There are several methods, including Curie point recording and compensation point recording, which can be used for recording. Figure 14.6 is an elementary illustration showing the principle of Curie point recording. In this example the initial magnetization of the recording layer is uniformly oriented in a given direction, as shown in Figure 14.6a. When a limited area of the recording layer is irradiated with light sufficient in intensity to heat it to a temperature above the Curie point T_c, the magnetization of the local area

TABLE 14.1

Specifications of CD-R Disk Systems

Items	Unit	Specifications
User data capacity	Mbytes	~650
Disk diameter	Mm	120
NA of OB		0.5
Wavelength	Nm	775–795
Wavefront distortion of pick-up	Λ	<0.050
Recording power	mW	$4 < Po < 8$
Playback power	mW	<0.7
Thickness of disk substrate	Mm	1.2
Rim intensities		Tangential 0.14 ± 0.04
		Radial 0.7 ± 0.10

TABLE 14.2

Characteristics of Erasable Optical Disks

	PCR	MO
Recording and erasing mode	Phase change	Change of magnetization
Read	Change of amplitude	Change of polarization
Material of medium	Te–Ge–Sb	Tb–Fe–Ni–Co
Overwrite mechanism	Simple	Complicated
Magnetic field	Not required	Required
Reversibility	Fair	Good
Required power	High	Medium

PCR, phase-change rewritable; MO, magneto-optical.
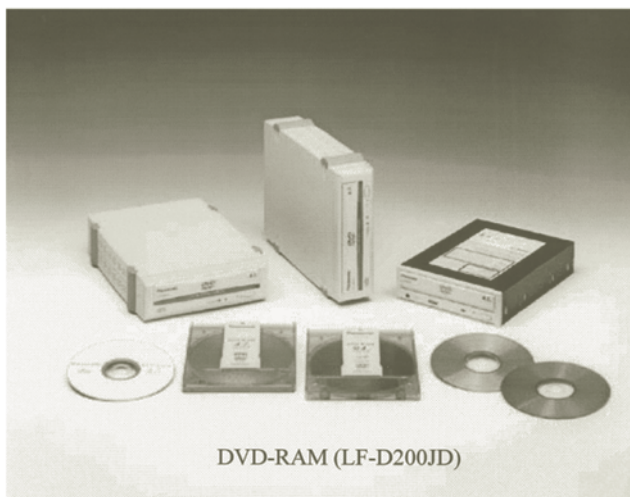
FIGURE 14.4
Optical data file drives (DVD-RAM).
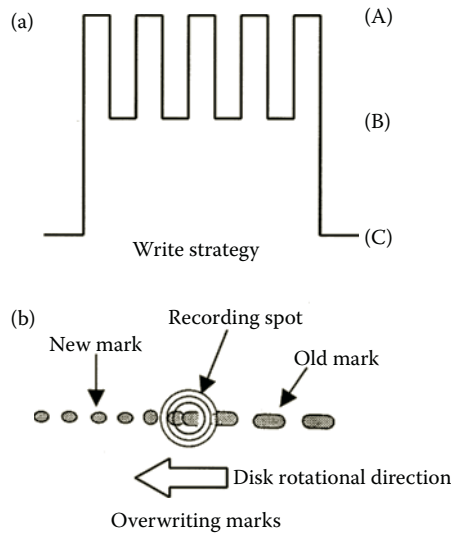
FIGURE 14.5
Principle of the direct overwriting mechanism for the PCR disk: (a) Write strategy, (b) Overwriting marks.

is lost, as shown in Figure 14.6b. When the exposure is discontinued, the temperature of the recording layer falls below T_c. The exposed area is remagnetized, but the direction of this magnetization coincides with the direction of the applied external magnetic field. Therefore, if the external magnetic field is applied in a direction opposite that of the original magnetization of the recording layer, as shown in Figure 14.6b, a magnetic domain different from the surrounding area remains, as shown in Figure 14.6c, enabling the recording of binary information. For reading the signal, the recording layer is irradiated with a laser light of low power. The polarization rotations of the reflected beam from the signal surface and the land surface are in opposite directions, as shown in Figure 14.6c. These beams are detected with a polarization analyzer to obtain the read signal. To erase the information, a selected area is again heated to a temperature above the Curie point, as shown in Figure 14.6d. The direction of the external magnetic field is reversed from that for recording.

In signal readout, the linear polarization angle of the incident beam is set at θ. The Kerr rotation angle is given by $\pm\Phi_k$. Referring to Figure 14.7, the differential output ΠI of the analyzer between the x and y directions is given by

$$\Delta I = I_0 R[\cos^2(q - \Phi_k) - \cos^2(q + \Phi_k)]$$
$$= (1/2)I_0 R\sin(2q)\sin(2\Phi_k) \tag{14.3}$$

where R is the disk reflectivity.

Since the linear polarization angle of the incident beam is $p/4$ and $\Phi_k \ll 1$

$$\Delta I \sim I_0 R\Phi_k \tag{14.4}$$

Thus, the playback signal level is proportional to the incident light intensity I_0, the disk reflectivity R, and the Kerr rotation angle Φ_k.
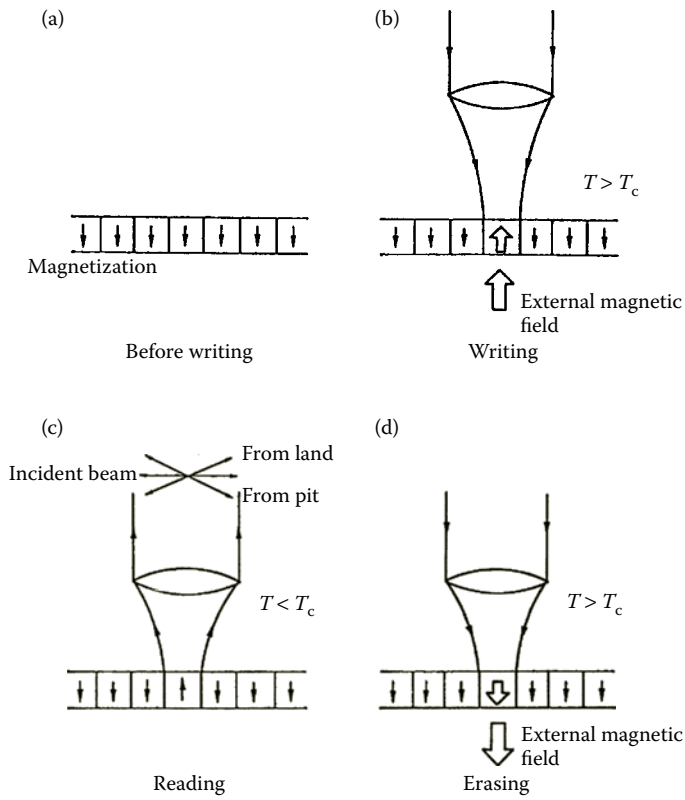
FIGURE 14.6
Magneto-optical disk method: (a) before writing, (b) writing, (c) reading, (d) erasing.
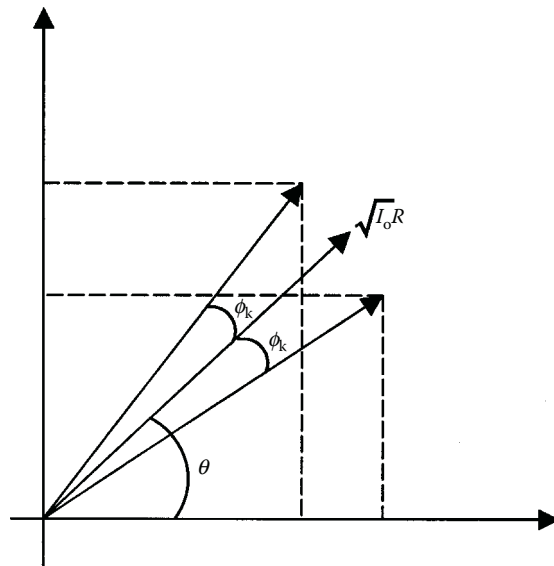
FIGURE 14.7
Readout of MO signal.

14.3 BASIC DESIGN OF OPTICAL DISK SYSTEMS

14.3.1 Pick-Up Optics

The many types of optical disks described in the preceding sections each have optimized optical pick-ups. The methods of design for the optics and mechanics of a writable optical pick-up will be described in this section.

The following factors determine the quality of read/write signals.

1. Frequency characteristics of signals
2. Cross talk from the adjacent tracks, which degrades read/write signals
3. Carrier-to-noise ratio (CNR) of read/write signals
4. Errors rate in read/write signals

Factors 1 and 2 are mainly dependent on the wave aberrations of the optics. Factor 3 is as much associated with the characteristics of elements such as the semiconductor laser, detector and electronics as with wave aberrations, and factor 4 is mainly dependent on defects in the disk.

14.3.1.1 Optical Layout

The schematic construction of the optics for a writable optical pick-up is shown in Figure 14.8. In this example, the astigmatic method is used for detecting the focusing signal and
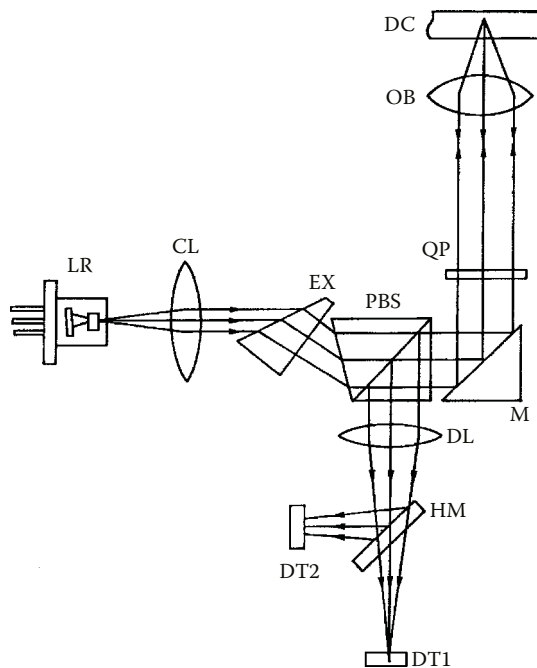
FIGURE 14.8
Schematic construction of a writable optical pick-up.

the push–pull (PP) method is used for detecting the tracking signal. The laser beam emitted from the semiconductor laser (LR) has a near-field pattern elongated in the direction of the active layer of the laser and is polarized in the same direction. The beam waist in this direction lies within the laser, and the beam waist in a direction perpendicular to the above direction is situated at the end facet of the laser active layer. The beam emergent from the laser is therefore anamorphic, and its far-field distribution is elliptical in cross section with an ellipticity of 2 to 3. To correct this elliptical distribution, it is necessary to use a one-dimensional afocal system after the CL, consisting of two cylindrical lenses or a wedge prism. With a single wedge prism, the designed incident angle of the laser beam must be approximately 69–72°. Since the wedge prism has chromatic dispersion, a change of the wavelength of the laser results in an angular deviation of the beam. Taking this angular deviation as $\Delta\alpha$ and the focal length of the OB as f_o, the beam spot moves approximately by $f_o \cdot \Delta q$ on the disk (DC). Using a single BK7 wedge prism with the incident angle 72°, an OB with focal length 4.5 mm and wavelength 0.78 μm, the beam spot displacement on the disk is approximately 0.073 μm for a change of 1 nm in wavelength. Therefore, the optics should be designed such that the direction of this movement will not cause a track offset. For this reason it is good practice to use two wedge prisms as illustrated in Figure 14.8. In the case of the playback-only optical pick-up, the influence of the elliptical and astigmatic beam can be small at the cost of beam utilization efficiency.

In Figure 14.8, the laser beam transmitted through a polarizing beam splitter (PBS) as a p-polarized beam passes through the $\lambda/4$ plate (QP) to become a circularly polarized beam, which is incident on the OB. The beam emerging from the OB is incident on the disk (DC) to form a beam spot for recording and reproducing the signals. The beam reflected at the disk enters the OB and again passes through the $\lambda/4$ plate (QP) to become an s-polarized beam and is reflected to the detection lens (DL) by the PBS. The beam emergent from the DL is partially reflected by a half-mirror (HM) and incident on the detector (DT2) for PP tracking signal detection. Because the convergent beam passing the HM is astigmatic, it is received by a quadrant detector (DTI) to give a focusing signal. The data signal is retrieved by sum of the output from both detectors (DT1 and DT2). The astigmatic focusing and PP tracking methods will be described in detail later.

14.3.1.2 Influence of Intensity Distribution

The intensity distribution of the beam incident on the OB is dependent on the beam divergence angle distribution of the semiconductor laser. With the OB aperture radius being standardized as unity and the intensity distribution of the incident beam assumed to be $\exp(-ar^2)$, the amplitude distribution is given by the Fourier–Bessel transform:

$$g(s) = \int \exp(-ar^2) J_0(sr) r \, dr \tag{14.5}$$

with $s = 2pnR/\lambda f_0$, where f_0 is the focal length of the OB and R is the polar coordinate in the focal plane. Integration gives (Appendix A, Equation 14.A4):

$$g(s) = \sum_{n=0}^{\infty} 2^n a^n e^{-a} \left(\frac{2J_{n+1}(s)}{s^{n+1}} \right) \tag{14.6}$$

Since $\square = 0$ for plane-wave incidence,

$$g(s)|_{a=0} = \frac{2J_0(s)}{s} \tag{14.7}$$

This is the well-known Airy distribution. When $a = 1$, the beam intensity distribution around the OB aperture is $1/e^2$. Figure 14.9 shows the intensity distribution of the $|g(s)|^2$ beam spot with various values of a. It is apparent from Figure 14.9 that when $a = 1$, the FWHM of the beam spot is increased by about 10% (relative to $a = 0$), and the peak of the side-lobe diffraction ring is made sufficiently small.

In order for the reproduced signal to have satisfactory frequency characteristics, the value of α in the signal direction must be in the range of $a \ll 1$. On the other hand, in the direction perpendicular to the signal direction, the cross talk from the adjacent track must be minimized. This cross talk can be small by using α close to 1. Therefore, the spot on the disk need not be truly round, but an improved frequency characteristic is sometimes obtained when the beam spot is elliptical with an ellipticity of about 10%.

14.3.2 Wave Aberrations[12]

When root-mean-square wave aberration W exists, the on-axis energy density Strehl definition (SD) of the beam spot is expressed by

$$SD = 1 - k^2 W^2 \quad \text{with } k = \frac{2p}{1}$$

(14.8)

where λ is the wavelength.

Figure 14.10 shows the relation between rms wave aberration and on-axis energy density. The on-axis energy density SD is a factor directly associated with reproduced signal SNR or record/reproduced signal SNR. The allowable rms wave aberration for the whole optical disk system is subject to Maréchal's criterion that the rms wave aberration is 0.070λ when SD has decreased about 20% from the level at no aberration. The validity of the criterion has been endorsed by read/write experiments. This allowable wave aberration for the whole system must be allocated to disk thickness error and tilt error, initial optical aberration, and defocus value.
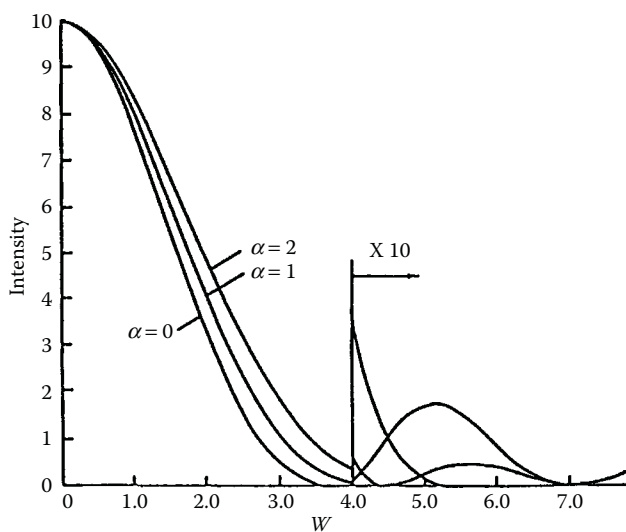
FIGURE 14.9
Intensity distributions of laser beam spot.
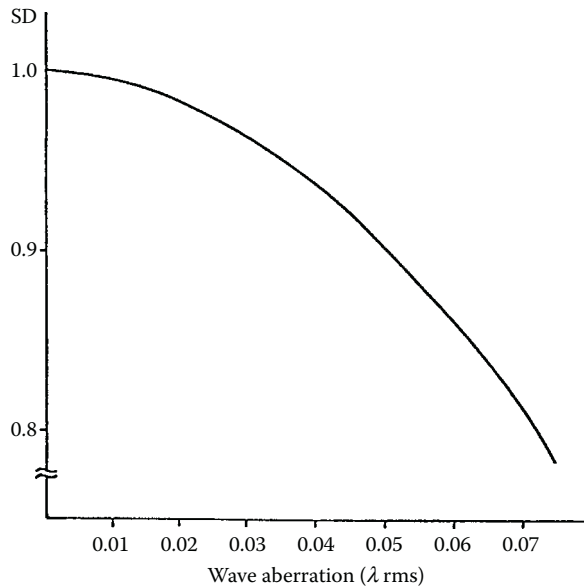
FIGURE 14.10
Wave aberrations vs. energy density on axis.

14.3.2.1 Aberration Derived from Disk Substrate

The aberration originating from the disk substrate is composed of the aberration W_{ST} due to the error $\sqcap t$ of substrate thickness t and the aberration W_{TL} due to the inclination θ of the substrate. These aberrations, when small, are expressed by the following equations (also Equations 14.A11 and 14.A14):

$$W_{ST} = \frac{\Delta t(n^2 - 1)(NA)^4}{8\sqrt{180}n^3}$$

(14.9)

$$W_{TL} = \frac{t(n^2 - 1)q(NA)^3}{2\sqrt{72}n^3}$$

(14.10)

where NA is the NA of the OB and n is the refractive index of the disk substrate. Figure 14.11 shows the relation between disk thickness error $\sqcap t$ and wave aberration W_{ST} with NA and wavelength λ as the parameters. Figure 14.12 shows the relation between disk tilt angle and wave aberration W_{TL} with NA and wavelength λ as the parameter. In the usual recordable CD, the practical values are NA = 0.5, t = 1.2 mm, n = 1.51, λ = 780 nm, Δt = 40 μm, and θ = 4 mrad. Substituting these values, we obtain W_{ST} = 0.011λ and W_{TL} = 0.017λ. In the DVD optical disk, the values are NA = 0.6, t = 0.6 mm, n = 1.51, λ = 650 nm, Δt = 16 μm, and θ = 3.9 mrad for the same values of W_{ST} = 0.011λ and W_{TL} = 0.0172λ. The plot for higher NA = 0.85 and short wavelength λ = 405 nm is shown as a reference. Thus the tolerance of the tilt angle for the DVD disk is almost the same as for the CD, as contrasted with the tolerance of the thickness being small.
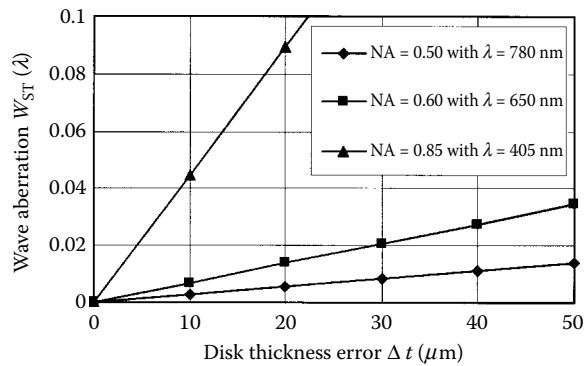
FIGURE 14.11
Wave aberrations versus disk thickness error.
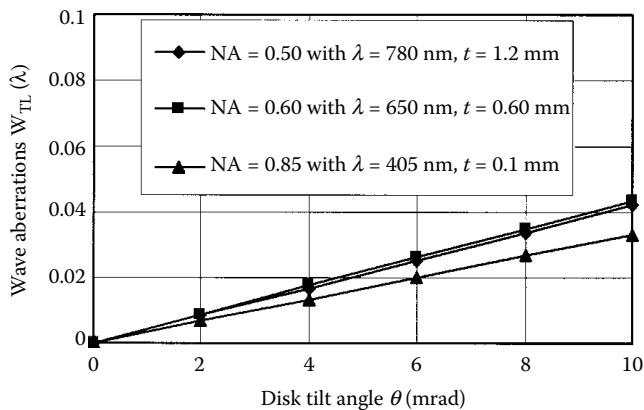
FIGURE 14.12
Wave aberrations versus disk tilt angle.

14.3.2.2 Wave Aberrations of Optical Components

Because mass-produced items are used for the disk optical components, the influence of variations in wave aberrations cannot be disregarded. Of all the components of the optical pick-up, the OB and the CL have the largest wave aberrations. Both the OB and the CL usually are aspherical pressed glass (APG) available on a mass production basis. Figure 14.13 shows an example of a mass-produced aspherical OB. Figure 14.14 shows the wave aberrations of a typical APG OB as measured with a Fizeau interferometer. The wave aberrations of prism systems are generally small, but as the number of prisms increases, the allowance for the entire pick-up is consumed.

14.3.2.3 Aberration Due to the Semiconductor Laser

Semiconductor lasers are generally astigmatic, and as this astigmatism is propagated and focused on the disk, variations in frequency characteristics may occur according to relative directions on the disk, or there may be reduced focusing latitude. In the record-playback optics, the electromagnetic emission from a semiconductor laser is passed through a CL
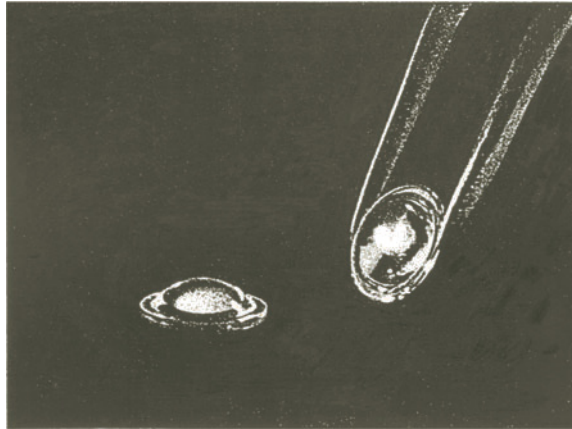
FIGURE 14.13
Mass-produced aspherical objective lens.
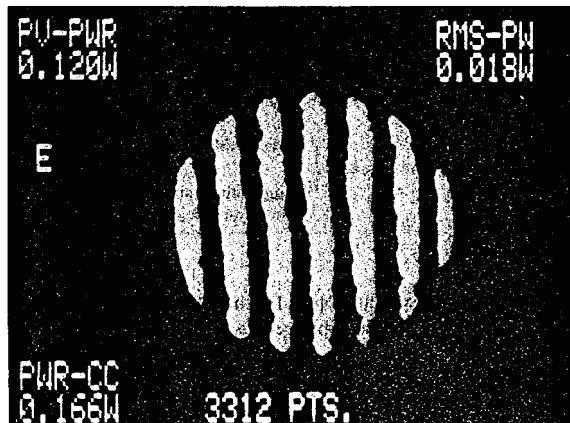
FIGURE 14.14
Wave aberrations of a typical APG objective lens.

to yield a beam of substantially parallel rays, which is then converted by an anamorphic beam expander to a beam having substantially isotropic distribution.

The correction for astigmatism is carried out concurrently in this stage. If stationary prism is used for correction, the astigmatism generated at an angle of 45° with the prism cannot be corrected. Therefore, when the semiconductor laser is mounted at an angle θ with the horizontal direction of the anamorphic expander prism, there occurs a residual astigmatism. The residual wave aberration W_{LA} due to this astigmatism is expressed as (Equation 14.A20)

$$W_{LA} = \frac{\tan q \Delta_L (NA_C)^2}{\sqrt{6} \cos^2 q} \tag{14.11}$$

where NA_C is the NA of the CL and Δ_L is the astigmatism of the semiconductor laser. Assuming that the allowable wave aberration dependent on the semiconductor laser is 0.010λ, the NA of the CL is less than 0.25 and the astigmatism of the semiconductor laser

is less than 8 µm (0.32 mil), and the allowable angle θ of the semiconductor laser is found from Equation 14.11 to be

$$q \le \pm 4°$$
(14.12)

14.3.2.4 Defocus

The factors responsible for defocus in an optical system may be classified as in Table 14.3. The relationship between the amount of defocus ε and the maximum optical path difference Δ_{DF} of the wavefront can be found from Figure 14.15:

$$\Delta_{DF} = \frac{e(NA)^2}{2}$$
(14.13)

Using the relationship between maximum optical path difference Δ_{DF} and wave aberration W_{DF}, wave aberration can be expressed as

$$W_{DF} = \frac{e(NA)^2}{4\sqrt{3}}$$
(14.14)

TABLE 14.3

Defocusing Factors

Defocus	Static defocus	Initial setting error
		Aging error
	Dynamic defocus	Servo residual error
		Temperature- and humidity-related error
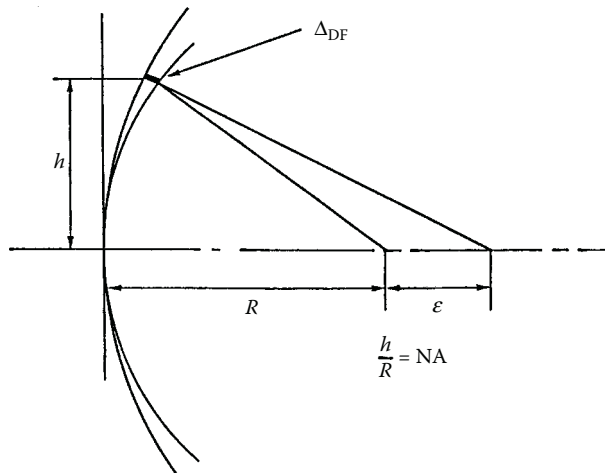
FIGURE 14.15
Optical path difference versus defocus.

For initial focus setting, the use of a diffraction grating having a spatial frequency near one-half of the cutoff frequency $2NA/\lambda$ of the disk optics is advantageous, for the influence of defocus is then most pronounced. Because the track pitch of the optical disk is usually the space frequency in the vicinity, the position of best focus is where the modulation by the track is maximal. By this adjustment, the setting error can be reduced to less than ±0.14 µm. Thus, with an optical disk of NA = 0.6 and λ = 650 nm and a defocus of +0.14 µm, the wave aberration is $W_{DF} = 0.011\lambda$.

14.3.2.5 *Allowable Wave Aberration*

Table 14.4 shows the typical wave aberration classified by causative factors. These wave aberrations can be integrated into the system allowance limit of 0.070λ rms as a totality. Since most of the factors responsible for wave aberrations are independent by nature, it is possible, in the actual design of an optical pick-up, that the allowable aberration value of each optical component is fairly liberal, as shown in Table 14.4.

14.3.3 **Optical Pick-Up Mechanism**

14.3.3.1 *Optical Pick-Up Construction*[13,14]

The optical pick-up generally consists of an optical base forming the optics assembly and an actuator for allowing the OB to follow the disk plane and tracking groove.

A typical optical pick-up construction for DVD is shown in Figure 14.16. Laser and PDs are combined on one silicon substrate. The beam emitting from the laser is initially parallel to the silicon surface and is then reflected by an engraved mirror to become perpendicular to the silicon surface. A polarizing hologram, shown in Figure 14.17, is used as a BS. It is transparent for the p-polarized beam emitted from the laser and diffractive for the s-polarized beam reflected from the disk. The p-polarized beam becomes an s-polarized beam as a consequence of the double pass through the quarter-wave plate. PDs formed on the silicon surface are located on both sides of the laser, and receive the beam diffracted by the polarizing hologram. The environmental resistance and reliability of the system are greatly enhanced when the number of reflective surfaces is minimized throughout

TABLE 14.4

Factors Responsible for Wave Aberrations and Amounts of Aberrations for DVD

System allowance limit 0.070λ		
Disk 0.028λ	Thickness error < ±1.6 µm	≤0.011λ
	Tilt < ±4 mrad	≤0.017λ
	Semiconductor laser	≤0.010λ
	Objective lens	≤0.035λ
Head 0.054λ	Collimating lens	≤0.025λ
	Wedge prism	≤0.014λ
	PBS	≤0.020λ
	$\lambda/4$ plate	≤0.020λ
	Perpetual change	≤0.011λ
Defocus 0.036λ	Initial setting error	≤0.011λ
	Servo residual error	≤0.023λ
	Temperature-dependent error	≤0.023λ
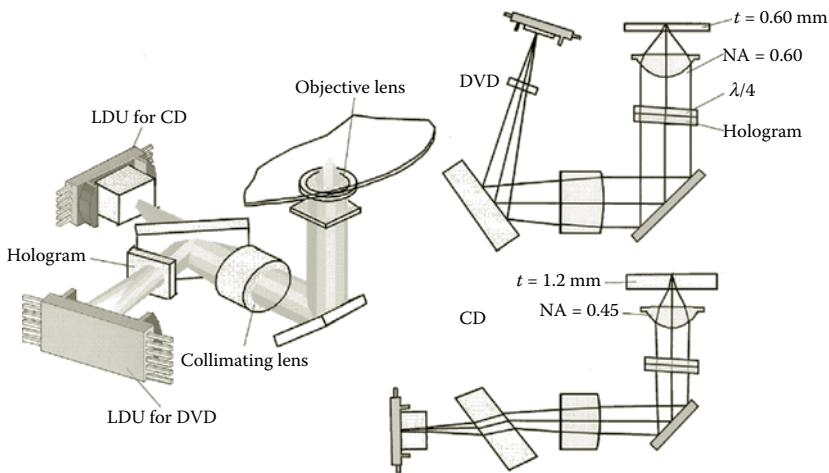
FIGURE 14.16
Optical pick-up for DVD.
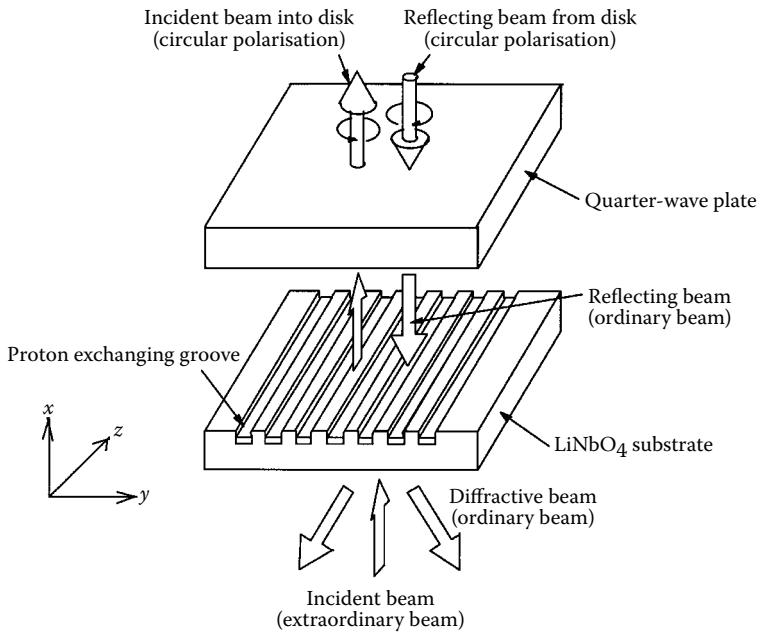
FIGURE 14.17
Polarizing hologram.

the optical path from the laser to the OB. Thus the construction with the integrated laser detector unit (LDU) is advantageous for reliability. Figure 14.18 is a view showing a LDU for a DVD player.

The optical base has a three-point support structure that enables a two-dimensional adjustment of the tilt angle of the optical pick-up. The axis of the OB is aligned perpendicular to the disk plane by mechanical adjustment, and the tilt of the optical base or the angle of inclination of the actuator must be adequately adjusted. The actuator and the optical
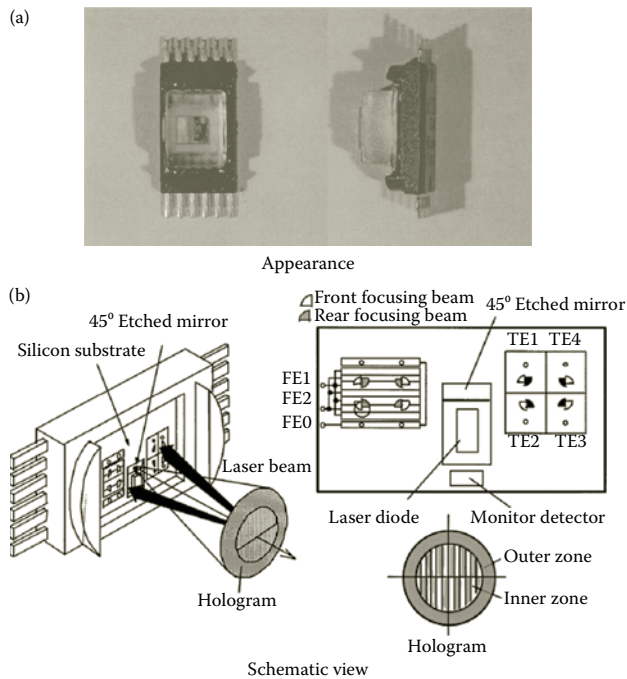
FIGURE 14.18
Integrated laser detector unit for DVD: (a) Appearance, (b) Schematic view.

base are provided with a convex and a concave spherical surface, respectively, whereby tilt correction can be made in two dimensions by means of couple of screws and springs. When the center of the sphere is aligned with the focal point of the OB, there is no transverse shift of the beam as it passes through the OB.

14.3.3.2 Actuator

The actuator has both a focusing drive mechanism for following the axial position of the disk and a tracking drive mechanism for following the track on the disk. The actuator must include a balanced combination of these two mechanisms, and must be designed in such a manner that there will be a minimum of interference between the mechanisms. The essential conditions that must be satisfied in the design of an actuator are:

1. Satisfactory frequency characteristics
2. High acceleration characteristics
3. High current sensitivity
4. Broad dynamic ranges for both focusing and tracking

Two actuator constructions satisfying these criteria are shown in Figure 14.19. The wire-suspended actuator[15] in Figure 14.19a is quite simple in construction and can be moved in the focusing and tracking directions by driving the center of gravity of its movable segment. Moreover, reliability is high because the four wires can be utilized as leads to the coils. The rotational actuator in Figure 14.19b is characterized by a small tilting angle of the critical axis and a large focusing dynamic range.
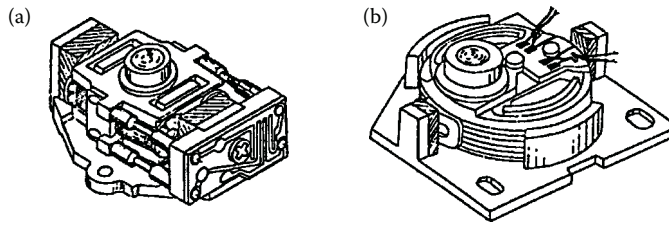
FIGURE 14.19
Two different types of actuator: (a) wire-suspended and (b) rotational.

The first-order resonant frequency f_0 of the wire-suspended actuator is

$$f_0 = 2p\sqrt{\frac{K}{m}} \tag{14.15}$$

where K is a spring constant and m is a movable mass. As a rule of thumb, the dynamic frequency range of an actuator is approximately from the level of the basic disk rotation frequency to the peak level of high-order resonant frequency. The larger this dynamic range value is, the larger is the servo gain that can be obtained.

14.4 SEMICONDUCTOR LASER

14.4.1 Laser Structure

14.4.1.1 Operating Principles of an Al–Ga–As Double Heterojunction Laser[16]

The energy band diagram of a double heterojunction semiconductor laser is shown in Figure 14.20. This laser consists of three layers having dissimilar energy gaps Eg, with increased energy gaps for the n-type and p-type cladding layers, which are on both sides of the active layer. As a photon hvg_2 corresponding to the active layer energy gap Eg_2 ($Eg_2 = hvg_2$) passes through the active layer, the electrons in the conduction band drop into the positive holes in the valence band to trigger a stimulated emission in phase with the incident photon. As a current I_p in the normal direction is passed through this diode, the probability of the presence of electrons in the active layer 2 is increased by the energy barrier $\Box E_c$, the conduction band. On the other hand, in the valence band, the probability of the presence of positive holes in the active layer 2 is increased by the energy barrier $\Box E_v$ to cause a population inversion in the active layer 2. In the active layer, therefore, the conduction band becomes full of electrons normally absent at thermal equilibrium, and the probability of recombination of electron-hole pairs with stimulated emission is increased. An incoming photon into the active layer is thereby amplified. The feedback mirrors at both ends of the active layer constitute a resonant cavity, and as the amplification surpasses the losses within the resonance cavity laser emission takes place.

14.4.1.2 High-Power Laser Technology[17]

Low-current and high-temperature operating laser diode of 650 nm AlGaLnP with a real refractive index guided self-aligned (RISA) structure is schematically illustrated in

Figure 14.21. The RISA structure is characterized by an AlInP current blocking layer, which leads to small internal loss in the waveguide and substantially reduces operating carrier density. The resultant operating current for 950 mW continuous wave at 70 °C is less than 100 mA. The RISA laser is produced by two steps of MOCVD growth. In the first MOCVD growth, an n-GaAs buffer layer, a cladding layer, the MQW active region, the optical confinement layer, the current blocking layer, and a nondoped GaAs (0.01 µm) are successively grown. Then, the stripe region for the current path is formed by chemical etching. In the second MOCVD growth, the cladding layer, buffer layer, and contact layer are grown. The cavity length is 500 µm. Front and rear facets are coated to obtain reflectivities of 4% and 90%, respectively.
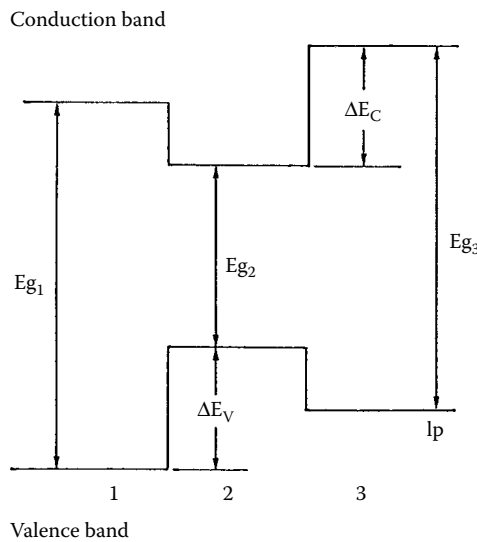
FIGURE 14.20
Energy band diagram.
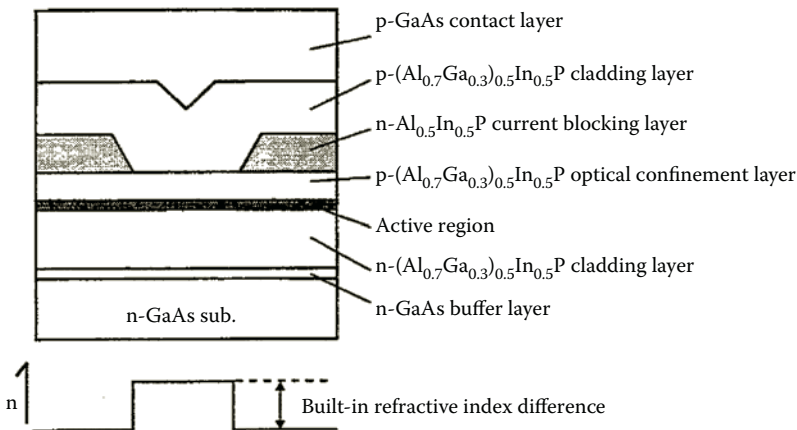
FIGURE 14.21
Schematic drawing of the RISA laser structure.

14.4.2 Astigmatism of the Laser

There are two categories of semiconductor lasers: gain-guided and index-guided. In a gain-guided laser the direction of beam propagation is not perpendicular to the wavefront. This mismatch causes relatively large astigmatism. Some lasers classified as index-guided also have weak evanescent waves. As a result, the beam waist in the horizontal direction is situated inwardly by Δ_L from the plane of beam emergence and thereby produces astigmatism. Whereas this astigmatism Δ_L is as large as 10–50 μm in the gain-guided laser, it is about 5–10 μm (0.2–0.4 mil) in the index-guided laser. Figure 14.22 shows a typical distribution of astigmatism in index-guided lasers. Generally speaking, the astigmatism of these lasers tends to decrease as the laser output increases.

14.4.3 Laser Noise[18]

Semiconductor diode lasers heat up during operation because of power dissipation arising from the injection current. The temperature increase induces mode hopping, which is a small shift in the wavelength of the output light. Figure 14.23 shows the temperature
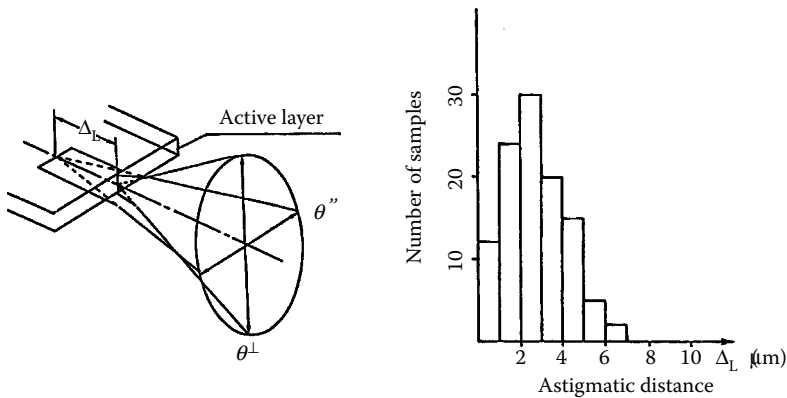
FIGURE 14.22
Distributions of astigmatic distance.
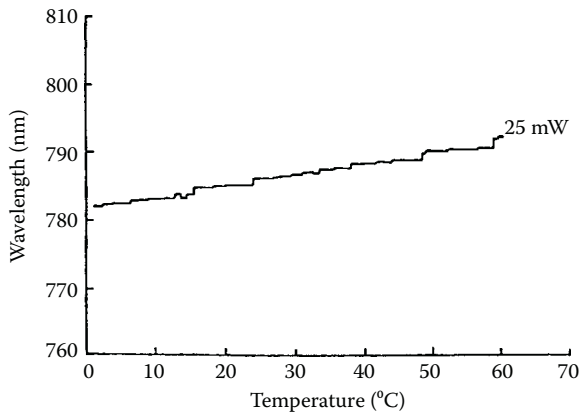
FIGURE 14.23
Temperature dependency of wavelength.

dependence of a semiconductor laser. As the temperature of the laser increases, the longitudinal mode of the laser is shifted toward longer wavelengths. Substantial noise accompanies this type of mode hopping. Figure 14.24 shows the relative intensity of noise (RIN) versus the laser heat sink temperature. Figure 14.24a shows the characteristic of the element itself, and the broken line represents the allowable noise level for an optical pick-up.

If a small amount of the output beam is directed back into the output aperture of the laser, the laser emission will become unstable, exhibiting both mode hopping and excessive amplitude noise. If the level of return light is about 0.5%, the relative intensity noise will be above the noise level allowable for an optical pick-up. Noise of this magnitude not only leads to a decrease in disk recording/reproduction SNR, but may also lead to instabilities in the focus and tracking servos.

An optical isolator consisting of a $\lambda/4$ plate and a PBS is typically used in the optical path to minimize light reflections back into the laser output operations. However, the return light cannot be completely eliminated due to the birefringence in the disk and variations in isolator performance. As a consequence, the generation of noise caused by return light is inevitable in semiconductor lasers of longitudinal single mode. Within the constraints of the basic aspects of the optical pick-up design, return light noise can be best controlled by broadening the emission spectral line width of the semiconductor laser to reduce the coherence of the light. Introducing the multiple longitudinal modes can broaden the emission spectral line width.

In index-guided lasers, the transverse mode behavior becomes single mode at an emission output of 1 mW or higher, due to a confining effect on the transverse mode. In gain-guided lasers, the transverse mode is confined by the gain corresponding to the carrier density and that generally results in multiple mode output. The influence of return light in a multimode laser is small and hence laser emission is unaffected. However, the inherent
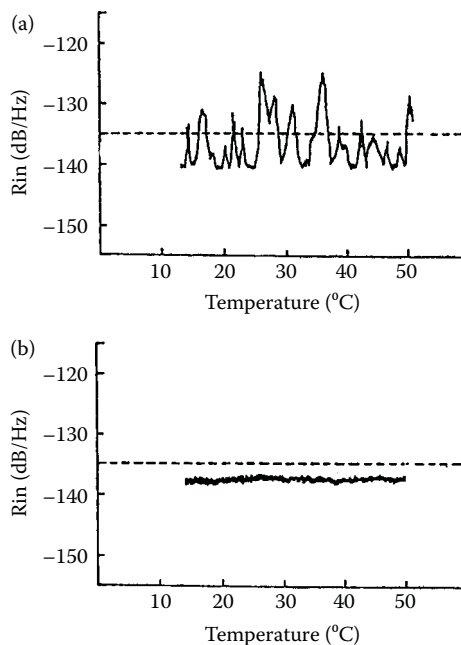
FIGURE 14.24
Noise characteristics of semiconductor laser: (a) inherent noise, (b) noise when modulated with high-frequency carrier.

noise level is higher than in single-mode lasers. As shown in Figure 14.24b, when the index-guided laser operating single mode is modulated with high-frequency carrier, the longitudinal mode becomes multimode with the result that the noise level is lowered. Figure 14.25 shows the return-light noise levels of various lasers. Noise level RIN is calculated as:

$$\text{RIN} = \frac{\langle \Delta P^2 \rangle}{P^2 \Delta f} \tag{14.16}$$

where $\langle \Delta P^2 \rangle$ is the mean square of noise power, P is the output power, and Δf is the noise bandwidth. When the high-frequency oscillation is set at 300–600 MHz and the modulation level is set below the laser emission threshold, the light output becomes a pulse emission providing a multimode operation. Figure 14.26 shows an example of (upper curve) the read signal obtained without high-frequency oscillation and (lower curve) the read signal with high-frequency oscillation. The addition of the high-frequency oscillation improves the CNR of the carrier signal by about 5 dB.

14.5 FOCUSING AND TRACKING TECHNIQUES

14.5.1 Focusing Servo System and Method of Error Signal Detection

The laser beam in an optical disk system is focused on the disk surface while the disk rotates at a high speed. An optical disk spinning at a high speed typically exhibits motion in the axial direction of tens to hundreds of micrometers. It is necessary that the OB follows this motion to keep the focus position of the beam on the signal plane of the disk
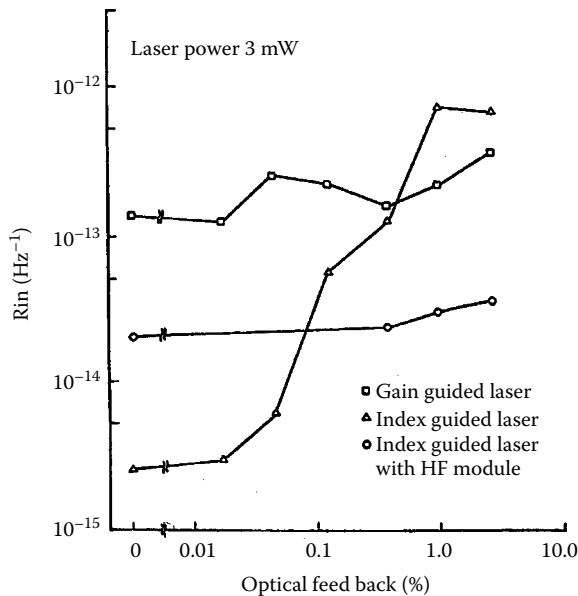
FIGURE 14.25
Laser noise versus optical feedback.
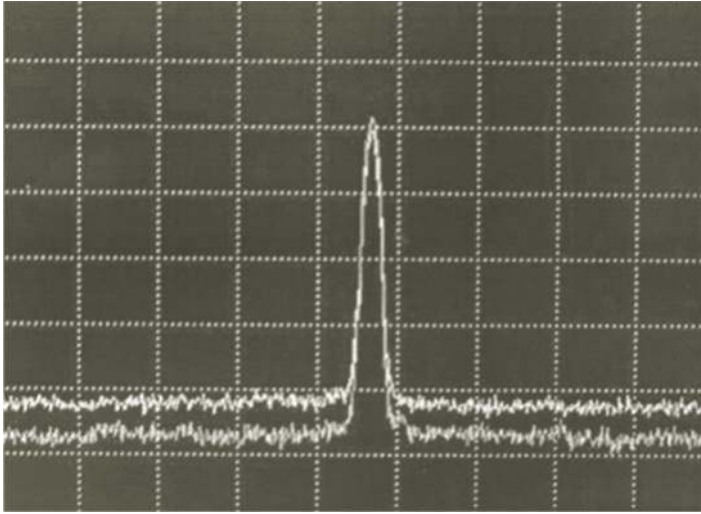
FIGURE 14.26
Reproducing carrier signals: upper baseline shows the noise level without high-frequency oscillation; lower baseline shows the noise level with high-frequency oscillation.

within the allowable limits of defocus of the optical system. The focusing mechanism generally used for this purpose is a moving-coil actuator employing a magnet and a coil. The required frequency response of the system is from several Hz to more than 10 kHz. Figure 14.27 shows a block diagram of the optical disk focusing system. The focusing servo loop comprises a focusing error signal detection unit, a circuit for amplification with phase correction of the detected error signal, and an actuator for driving the OB. This actuator is designed to follow the axial motion of the disk in response to a servo signal in the presence of external noise associated with the movement of the actuator and the interference from the tracking signal. In designing an autofocusing mechanism for the optical disk, external light noise must be reduced as much as possible. A balanced design must be employed and take into consideration: (1) interference from the tracking signal that occurs when the beam traverses the tracks; (2) mutual interference from motion of the focusing and tracking actuators; and (3) false focusing error signals associated with movement of the beam on the detector in the course of tracking.

Focusing errors are introduced by the axial motion of the disk, vibrations of the device and other causes. The focusing error information contained in the laser light reflected from the disk can be transformed into intensity or phase differences to derive an error signal. Each of the following beam characteristics can be utilized to generate a focusing error signal:

  1. Change of beam shape
  2. Movement of the beam position
  3. Phase of the modulated waveform of the beam

14.5.1.1 Beam Shape Detection Method

Two separate techniques can be used to detect the beam shape to obtain a focusing error signal. These are the astigmatic focusing detection method and the spot size detection method.

Figure 14.28 shows a basic optical system for the astigmatic focusing detection method using a tilted parallel plate. Although the conventional implementation includes use of a cylindrical lens, the tilted plate method is advantageous in the simplicity of the optics. The sensitivity of focusing error signal detection is dependent on the thickness and refractive index of the parallel plate assuming that the magnification of the OB is constant. The greater the thickness of the plate and/or the larger the refractive index, the larger the astigmatism and, hence, detection sensitivity. Figure 14.29 shows a typical focusing error signal in the optimum design. When the detection sensitivity is relatively low, defocus becomes large due to false signals caused by dropouts in the disk or movement of the OB during tracking. When the detection sensitivity is too high, the dynamic range of the focusing servo is diminished and the stability of the servo is decreased.

14.5.1.2 Spot Size Detection Method

Figure 14.30 shows the operating principle of the spot size detection method. The central part of the beam is received in front of and beyond the focal point of the beam by two three-segment detectors. A focusing error signal is derived from the intensity difference on the central and outer segments. The beam shape detection method generally has a large allowance for detector offset and has good temperature characteristics and aging stability. Recent progress of holographic technology enables use of a holographic optical element (HOE) to detect focusing signals.[13]
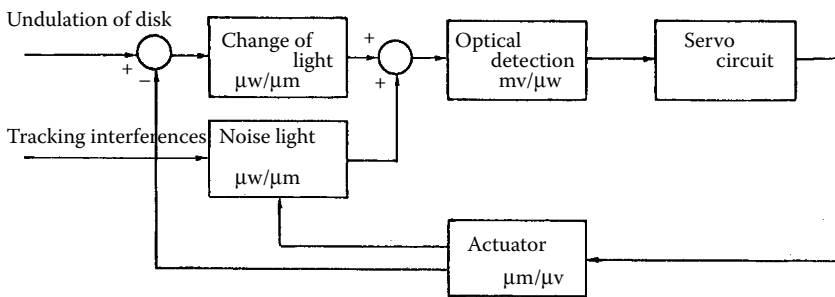
FIGURE 14.27
Block diagram of focusing servo.
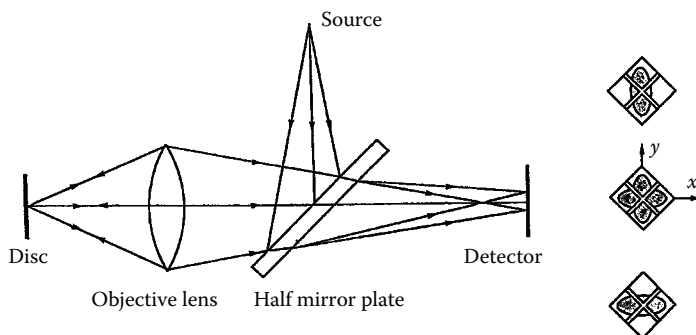
FIGURE 14.28
Astigmatic focusing method with plate.

Differential output
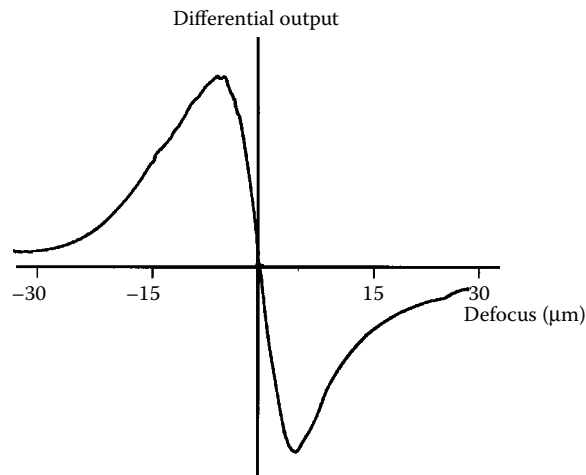
FIGURE 14.29
Typical astigmatic focusing error signal.
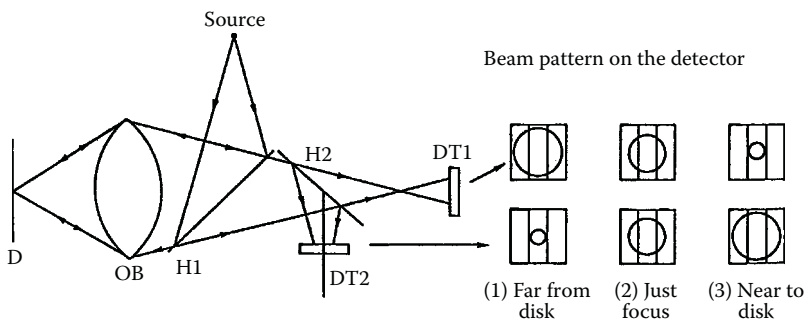
FIGURE 14.30
Spot size detection method.

14.5.1.3 Beam Position Detection Method

The beam position detection method converts the movement parallel to the optical axis—such as axial motion of the optical disk—to a beam movement in a plane perpendicular to the optical axis to obtain a focusing error signal. This detection method uses relatively simple hardware construction for detection and gives a broad focusing dynamic range.

Figure 14.31 shows a focusing error signal detection system using a bi-prism. This is an example of the Foucault focusing detection method. As the distance between the disk and the OB decreases, the intensity on the inner side of the respective split detectors increases and an increasing distance between the disk and the OB results in increasing intensity on the outer sides of the split detectors.

Figure 14.32 shows a focusing detection system using the critical angle of a prism. The beam of rays reflected from the disk enters the prism as a divergent beam when the distance between the disk and the OB is small, or as a convergent beam when the distance is large. When the prism is set at the critical angle, and the rays are not parallel, the beam will be partially transmitted by the prism to establish differential intensities on the
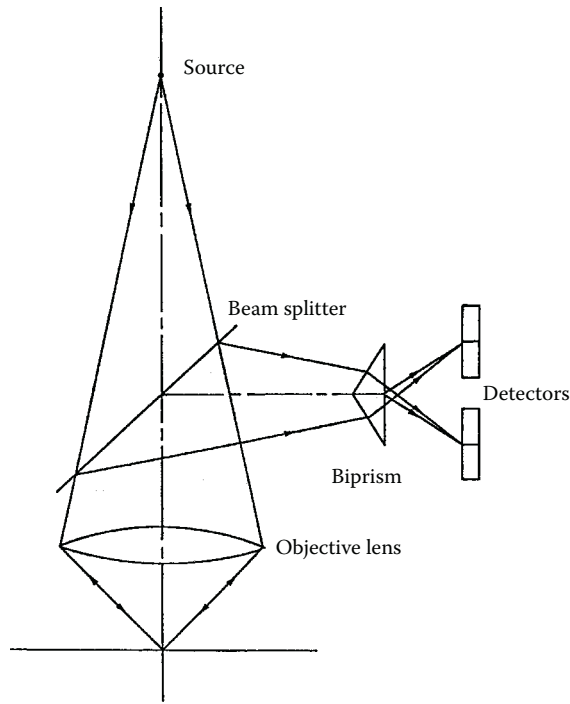
FIGURE 14.31
Foucault focusing method.
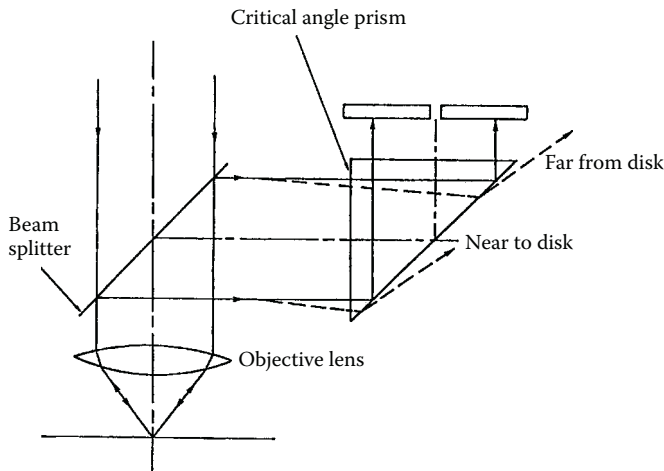
FIGURE 14.32
Critical angle focusing method.

detectors. In the case of a divergent beam, the near detector receives less light; in the case of a convergent beam, the far detector receives less light.

There are other methods for focusing detection, such as a skew beam focusing detection system, a system wherein the incident beam is eccentric with respect to the OB axis, a system using a single knife-edge, a beam rotation focusing detection system, and others.

14.5.1.4 Beam Phase Difference Detection

There are two methods to detect beam phase difference: the spatial phase difference detection method and the temporal phase difference detection method. In the spatial phase difference detection method, illustrated in Figure 14.33, the phase of the beam located in the far-field pattern of the reflected beam diffracted by a given pattern in the optical disk (e.g., the pregrooved track pattern) is detected. This method is dependent on beam wavelength, and the dynamic range of focusing error signal is narrow. The temporal phase difference detection method is also known as the wobbling method. In this method, the focal point of the beam irradiating the optical disk is modulated along the optical axis with a wobbler. The phase of the modulated signal obtained with a detector is compared with the phase of the modulated drive signal of the wobbler to obtain a focusing error signal proportional to the phase difference.

14.5.2 Track Error Signal Detection Method

14.5.2.1 Detection Methods

The signal tracks on a DVD disk have a pitch of 0.74 μm, and the signal tracks on a conventional CD disk have a pitch of 1.6 μm. The beam spot for reproducing the signal must follow this track within an accuracy of 0.04–0.1 μm. This tracking performance is achieved by driving the OB in a lateral direction with a voice coil actuator. The following methods are commonly used for optical detection of the TE:

1. Detection using two auxiliary beams generated by grating (3-B or 3-beam method)
2. Detection of the far-field distribution of the read/write beam reflected from the disk (PP method)
3. Detection from the difference between two signal levels obtained with sample pits disposed at an offset of ±1/4 pitch from the track (SS or sampled servo method)
4. Detection of the phase difference between the differential output from diagonal playback signal of quadrant detector and the sum output signal of quadrant detector (DPD or differential phase detection method)
5. Detection of the phase difference between the playback signal obtained by a slight induced displacement of the beam in the direction perpendicular to the track and the phase of the corresponding drive signal (wobbling method)
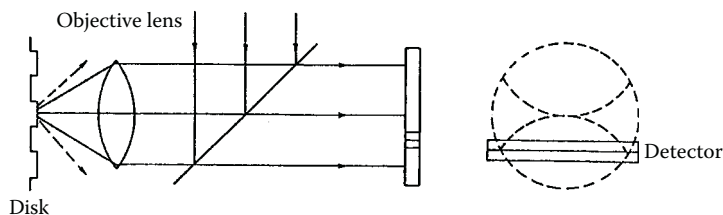
FIGURE 14.33
Spatial phase detection method.

14.5.2.2 3-Beam Method

The 3-beam method using auxiliary beams is shown in Figure 14.34. The two first-order beams obtained by passing the laser beam through a diffractive grating are aligned to positions on the disk of about plus and minus one-quarter of the track pitch apart from the track center (B1, B2). The reflected two beams are received by two detectors (D1, D2) to obtain a track error signal. The fundamental beam (B0) is used for SG detection by a central detector (D0). While this method is well suited to the read-only optical pick-up like CD, it must be carefully designed for use in read/write optics. In this case the beam intensity is increased during the writing mode, introducing the risk of harmful recording by the auxiliary beams.

14.5.2.3 Wobbling Method

In the wobbling method, the track error signal corresponds to the phase difference between a signal to the transducer that induces a slight displacement of the beam in the direction perpendicular to the track and a signal from the beam modulated by track edge diffraction. This method has only been implemented in certain limited applications. This is partly due to the poor stability of the wobbling frequency and partly due to the 0.1 μm wobble displacement required to obtain a TE with satisfactory S/N ratio. Wobble displacement of 0.1 μm is close to the maximum allowable tracking error value.

14.5.2.4 Differential Phase Detection (DPD) Method

This method uses the quadrant detectors to detect the phase difference between the differential output from the diagonal playback signal (D1 + D4) − (D2 + D3) of the quadrant detector and the sum output signal (D1 + D2 + D3 + D4) of the quadrant detector. The DPD tracking error detection is shown in Figure 14.35. This is the method recommended for playback TE detection in the DVD specifications.[19]
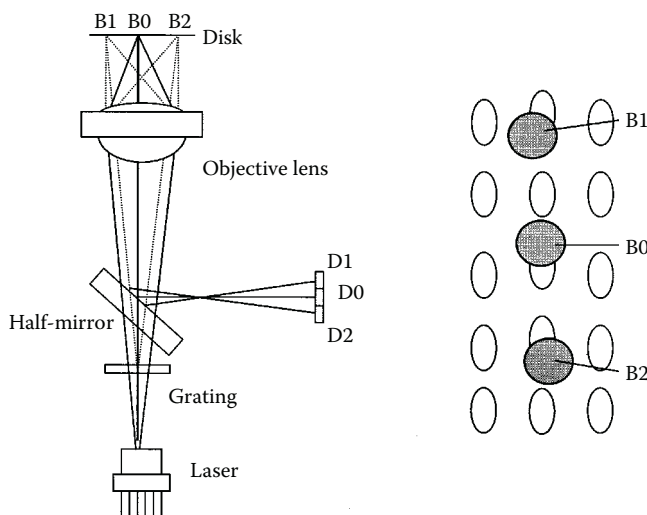
FIGURE 14.34
Tracking error signal detection with 3-beam method.
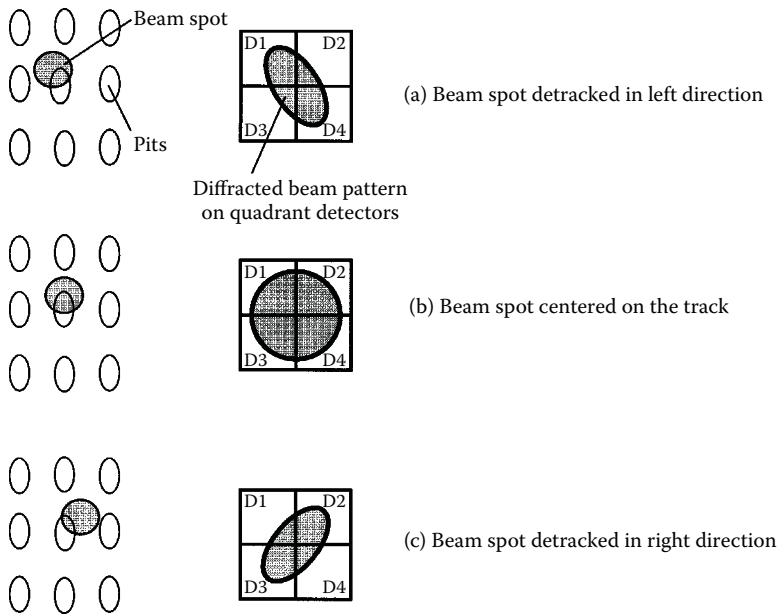
FIGURE 14.35
Differential phase detection method.

14.5.2.5 Push–Pull Track Error Signal Detection Method

The simplest method for obtaining a track error signal in read/write is called the "push–pull method." A split detector is inserted in the far-field of the beam in such a manner that the line of division of the detector is lined up with the track.

Figure 14.36 shows a basic optical system for signal detection using the PP method. The intensity pattern incident on the split detector is a combination of the zero-order and first-order beams (due to diffraction caused by the track on the disk). The track error signal is derived from the difference in the signals from the two detectors. Figure 14.37 shows the far-field beam distributions according to the beam spot position on the track. The asymmetry of the far-field beam intensity distribution and the track error signal level are maximum when the track groove depth is $\lambda/8$. (When the depth is $\lambda/4$ multiplied by an integer, asymmetry disappears and no track error signal is obtained.)

14.5.2.6 Slit Detection Method[20]

When the beam spot is located in the center of the track, the phase difference ψ between the zero-order beam and the two first-order beams in the PP system is dependent on the diffraction by the track and on the defocus ΔZ and can be expressed as[21]

$$
\psi = \frac{p}{2} + \frac{2p}{l}\left[\sqrt{1 - \left(\frac{l}{p} - \sin a\right)^2} - \cos(a)\right]\Delta Z
\tag{14.17}
$$

where α is an angle between the optical axis and an arbitrary point in a far-field image.
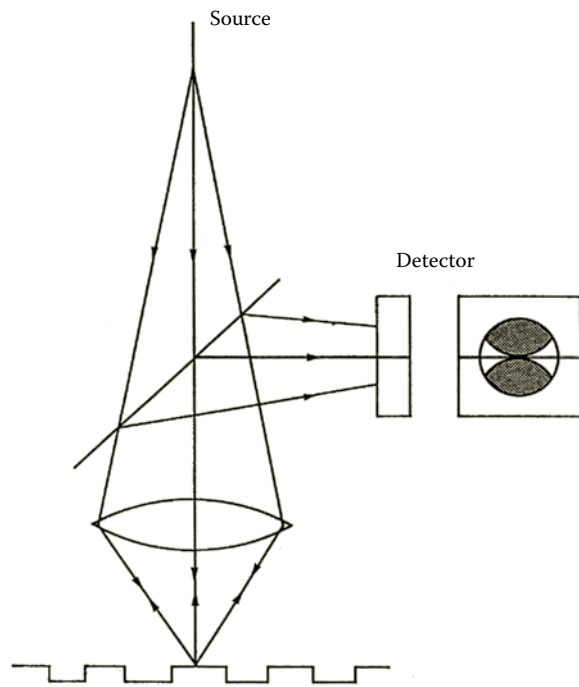
FIGURE 14.36
Push–pull method.
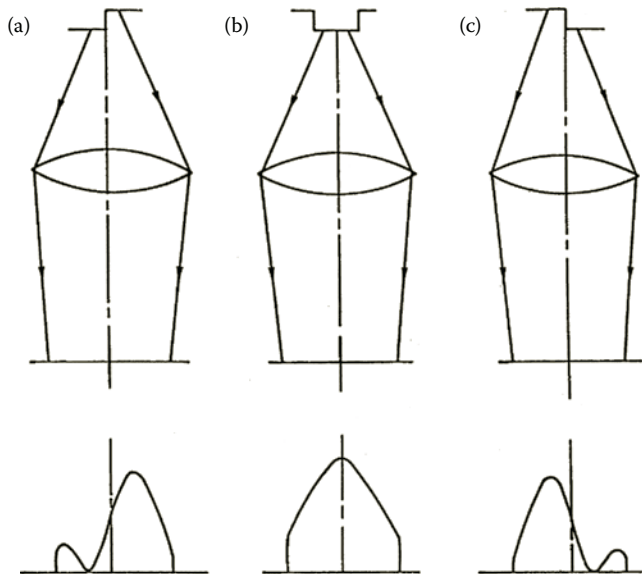
FIGURE 14.37
Far-field beam distribution at various locations on the track.

From Equation 14.17, the phase difference is constant for $a = \sin^{-1}(1/2p)$ in the far-field image and independent of defocus but exclusively dependent on diffraction by the track. Figure 14.38 shows a typical far-field beam distribution in the presence of defocus. By utilizing this property in the far-field, the control range of tracking with respect to defocus can be expanded. Thus, the defocus characteristic of the track error signal can be improved by providing slits symmetrically in the centers of overlaps between the zero-order beam and the two first-order beams as illustrated in Figure 14.39.

Figure 14.40 shows the change in track error signal level according to defocus at various slit widths. If the slit width is about 20% of the far-field pattern, the proportion of change in the track error signal due to defocus is improved by about a factor of 2. If the slit width
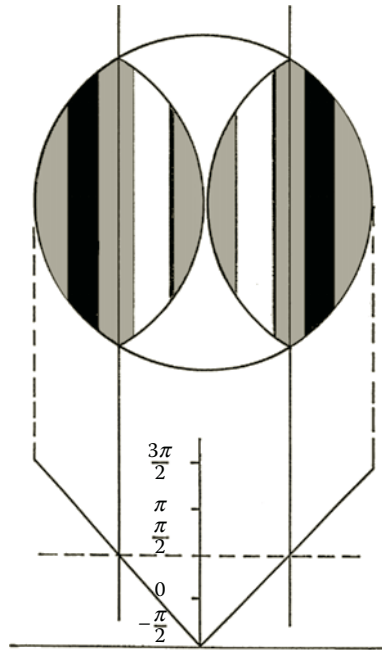
FIGURE 14.38
Far-field pattern when defocus occurred.
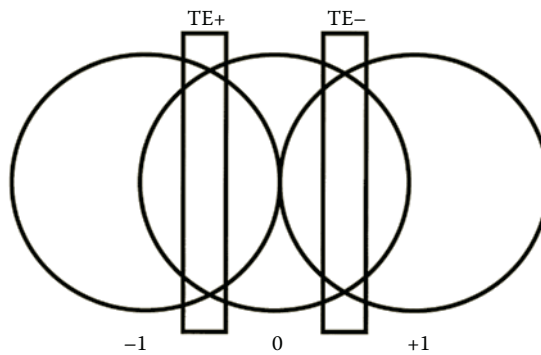
FIGURE 14.39
Slit detection method.
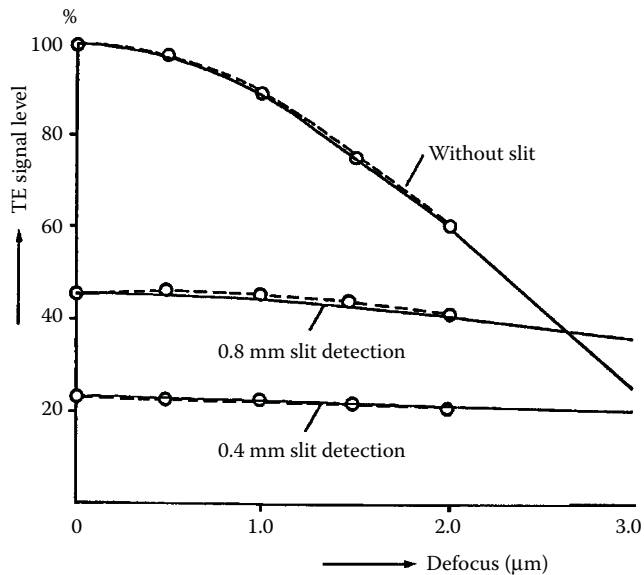
FIGURE 14.40
Defocus versus tracking error signal level.

is made too narrow, the S/N of the track error signal will decrease. The optical parameters in these experiments and theoretical calculations are as follows:

1. Objective lens: NA = 0.5
2. Laser wavelength: L = 830 nm
3. Track pitch: t = 1.6 µm
4. Slit width: W = 0.8 and 0.4 mm

14.5.2.7 Sampled Tracking Method[22]

In a sampled tracking system, track error signal detection pits are periodically provided in lieu of the continuous groove in a conventional grooved disk. The sample pits consist of two pits displaced by about ±1/4 of the track pitch from the track center and one pit centered on the track. Figure 14.41 shows the principle of sampled track error signal detection. When the beam spot is (a) off-track upwardly from the track center, the first pit output is large and the second pit output is small. In the on-track condition, the first pit output level and the second pit output level are equal (b). If the beam spot is downwardly off-track, the first pit output is small and the second pit output is large (c). The off-track condition is diagnosed by comparing these pit output levels. The third pit is used for making a sampling clock, and the track error signal detection is constructed from these outputs. Figure 14.42 shows a block diagram of the detection circuit.

In the sampled tracking system, each set of pits provided for the track error signal detection uses a track length equivalent to that used to store 1 byte of information. The sampling frequency must be higher than about ten times the cutoff frequency of the tracking servo. This reduces the size of the usable data area, but the overall system performance improves because there is less degradation of data signals and interference effects on the focusing
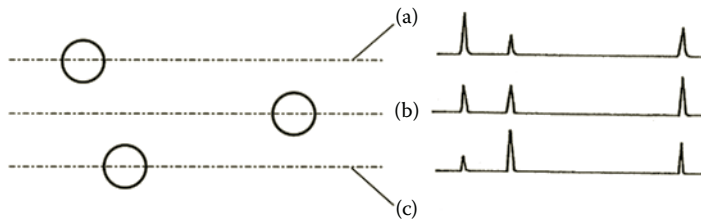
FIGURE 14.41
Tracking error signal detection from sampled pits: (a) off-track upwardly from the track center, (b) on-track, (c) off-track downwardly from the track-center.
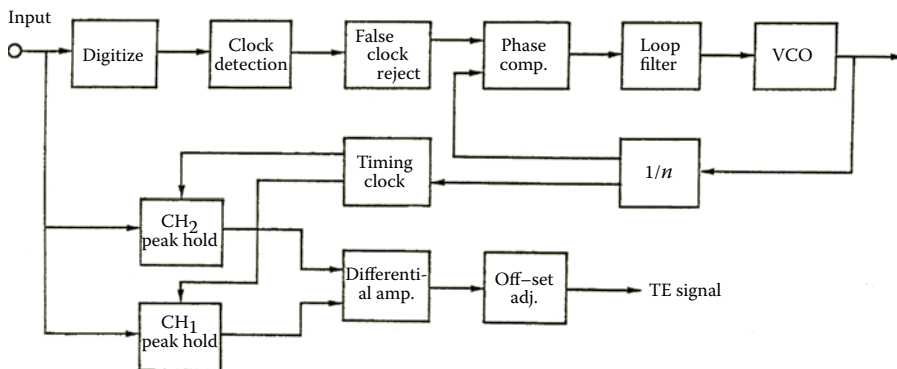
FIGURE 14.42
Block diagram of sampled servo tracking method.

servo caused by the track groove. Further, an inclination of the disk induces a track offset in the PP system, but not in the sampled tracking method. For example, a disk inclination of 0.7° of a 1.2-mm-thick substrate with NA = 0.5 and λ = 830 nm causes a 0.1 μm lateral shift in the quiescent operating position of the PP tracking servo. This type of systematic track error is decreased by about a factor of 5 when the sampled tracking method is used.

14.6 RADIAL ACCESS AND DRIVING TECHNIQUE

14.6.1 Fast Random Access

A critical aspect of an optical disk memory system is fast random access to the stored information. This random access is accomplished via two mechanisms: optical pick-up motion for rough positioning and the tracking actuator for precise positioning. A linear actuator is used as the coarse positioning means. To minimize access time, it is necessary to (1) develop a small and lightweight optical pick-up; (2) increase the resonant frequency of the linear actuator; and (3) develop a transfer mechanism with a minimum of friction. In the typical linear actuator designed for video recording applications, the transfer segment weighs only 78 g and has a thrust of 3.0 N/A. This linear actuator gives an average access time of 75 ms or less.[9] The optical pick-up base is provided with roller bearings so that it

freely moves on the guide rods. The low-frequency component of the track error signal is fed to the linear actuator so that the center of drive of the OB will lie at the center of the tracking mechanism motion range.

Figure 14.43 shows the modes of access by the linear actuator/tracking actuator combination[23] First, the tracking actuator is disabled. Then the linear actuator is accelerated and decelerated at the maximum speed to position the optical pick-up in the vicinity of the correct track (A to B). The tracking actuator is reactivated and the track address is read (B to C). The number of tracks between the actual and desired address is calculated, and access is completed by executing a multitrack jump. Figure 14.44 is a block diagram of the tracking servo circuit and linear actuator circuitry. The track error signal obtained by the procedure described in Sec. 5.2 is fed to an amplification circuit, a switching circuit, and a drive circuit in succession to drive the tracking actuator. The tracking drive signal is also used to drive the linear actuator. During random access, the drive signal is removed from
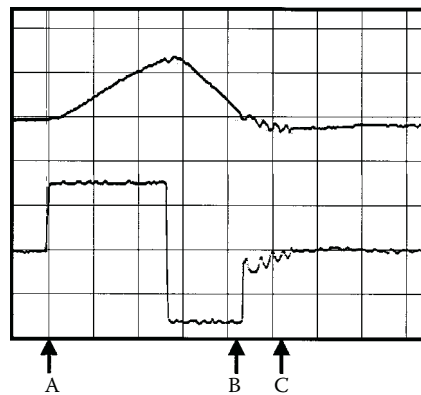
FIGURE 14.43
Modes of access by the linear actuator and tracking actuator.
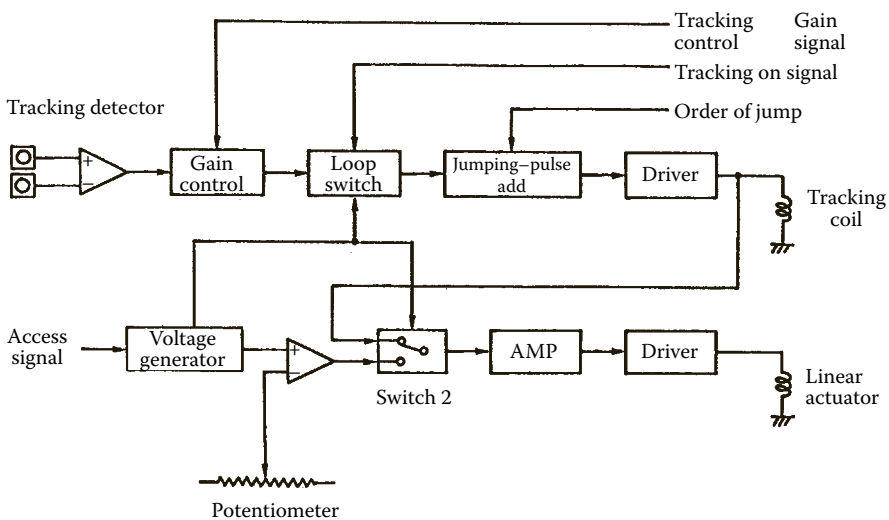
FIGURE 14.44
Block diagram of tracking and accessing servo.

the tracking actuator and a voltage corresponding to the access signal is generated and supplied to the linear actuator drive circuit. Two methods are available for ascertaining the actual position with respect to the target track. The first method involves calculating the number of tracks between the current and desired positions, then detecting and counting optically each pregroove as it is passed over in the radial scan. The scan is stopped when the correct number of tracks has been crossed.

Since this method counts the tracks themselves, the distance to the target track can be accurately computed. The track detection bandwidth must be broad enough to prevent miscounts of the tracks during the peak speed of the linear actuator. This method is less applicable when using sampled format disks; even track addresses can complicate the track-counting process in continuous format disks. A second method provides the optical pick-up with a position sensor for detecting the current position. This provides a stable position signal, and the access servo can be damped using the output signal from this sensor. Examples of position sensors include linear scale sensors, optical position sensors, and slide resistance sensors. Figure 14.45 shows a typical optical position sensor.

14.6.2 Optical Drive System

The optical disk system consists of hardware comprising the disk, optical pick-up, accessing circuit, signal processing circuit, error correction circuit, microcomputer, and so on, and software for processing the various signals. The height of the 5.25 in optical disk drive is either full height (82 mm), half-height (41 mm), 1 in height, or half-inch height, corresponding to standardized magnetic disk products. In the half-height drive, design goals include the use of a disk cartridge that is inserted and clamped and low profiles of the component parts for the access mechanism. The height of the optical pick-up must be less than 15–16 mm. For the thinner drive like those used in a notebook computer, the height of the drive is less than 12.7 mm. A height of the optical pick-up of around 7.5 mm is needed
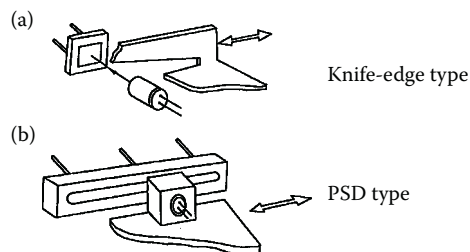
FIGURE 14.45
Optical position sensor: (a) knife-edge type; (b) PSD type.

FIGURE 14.46
Thin optical disk drive for DVD-ROM.

and an ingenious design is required. Figure 14.46 shows the thin optical disk drive for a DVD-ROM. The removability of the optical disk affects differences in the amounts of eccentricity and undulation each time a disk is mounted. Moreover, since the disk substrate is made of plastic, the amount of undulation increases with the age of the disk. The dynamic balance of the disk is affected by these factors, and vibrations are induced. It is necessary, therefore, to design the various actuators such that these vibrations do not degrade performance.

ACKNOWLEDGMENTS

Appendix A

When the amplitude distribution in the pupil is given by $f(r^2)$ and the radius of the pupil is one unit, the integral of Fourier–Bessel transform is written as

$$g(s) = \int_0^1 f(r^2) J_0(sr) \mathrm{d}(r^2) \tag{14.A1}$$

A plurality of solutions exist for this integral. However, the solution given by A. Boivin[24] is easily understood. Thus, the amplitude of the Fourier spectrum $g(w)$ is written in the form of a Bessel series:

$$g(s) = \sum (-1)^n 2^{n+1} f_{(1)}^n \frac{J_{n+1}(s)}{s^{n+1}} \tag{14.A2}$$

where $f^n(r^2)$ denotes the nth differential of the function $f(r^2)$. For calculating the Fourier spectrum of a truncated Gaussian, $f(r^2)$ is expressed by $\exp(-ar^2)$. Then the integral of Fourier–Bessel transform is written as

$$g(s) = \int_0^1 \exp(-ar^2) J_0(sr) r \mathrm{d}r \tag{14.A3}$$

and the result becomes[24]

$$g(s) = \sum_{n=0}^{\infty} 2^n a^n e^{-a} \left[\frac{2J_{n+1}(s)}{s^{n+1}} \right] \tag{14.A4}$$

Appendix B

Thickness variations, index changes, and tilts of the disk substrate all cause wavefront aberrations. Here, we calculate the optical path differences Δ_0 between two rays: the first is the on-axis ray, and the second is the outermost ray, which determines the NA of the OB. From Figure 14.47, the following relations can be easily calculated:

$$\sin(\psi - \theta) = n\sin(r_1) \tag{14.A5}$$

$$\sin(\theta) = n\sin(r_0) \tag{14.A6}$$

$$\Delta_0 = nt\{1/\cos(r_1) - 1/\cos(r_0)\} + t\{\cos(\psi - \theta)/[\cos(r_1)\cos(\theta)] - 1/\cos(r_1)\}/n \tag{14.A7}$$

Developing the power series of ψ and θ, the next quadratic terms are obtained:

$$\Delta = t(1-n)^2\{\psi^4 - 4\psi^3\theta + 8\psi^2\theta^2 + 8\psi\theta^3\}/8n^3 \tag{14.A8}$$

where ψ is the NA of the OB and t is the thickness of the disk substrate. Here, each term denotes the Siedel aberration. When only a thickness error Δt exists, the spherical aberration S_1 is generated:

$$S_1 = \frac{(n^2 - 1)\psi^4\Delta t}{8n^3} \tag{14.A9}$$

The relation between the wave aberration W_{ST} and the spherical aberration S_1 can be calculated from Maréchal's equation:[25]

$$W_{\mathrm{ST}}^2 = \frac{d^2}{12} + \frac{dS_1}{6} + \frac{4S_1^2}{45} = \frac{(d + S_1)^2}{12} + \frac{S_1^2}{180} \tag{14.A10}$$
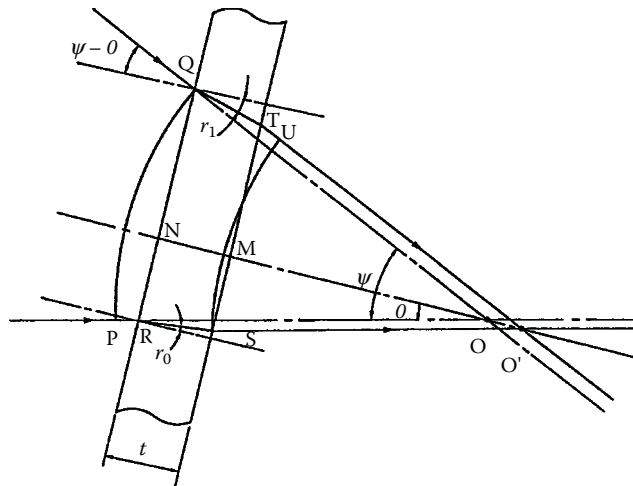
FIGURE 14.47
Optical path differences between two rays.

where the wave aberration W_{ST} is minimal when the defocus d is equal to the third spherical aberration S_1. Thus, the wave aberration due to the spherical aberration becomes

$$W_{ST} = \frac{S_1}{\sqrt{180}} = \frac{\Delta t(n^2 - 1)(NA)^4}{8\sqrt{180}n^3} \tag{14.A11}$$

For a small tilt θ of the disk substrate, high orders can be disregarded and the only important aberration is the coma C_1:

$$C_1 = t(n^2 - 1)y^3 q/2n^3 \tag{14.A12}$$

The relationship between the wave aberration W_{TL} and the coma C_1 can be calculated, again from Maréchal's equation[25]

$$W_{TL}^2 = K^2/12 - KC_1/6 + C_1^2/18 = (K - C_1)^2/12 + C_1^2/72 \tag{14.A13}$$

where the wave aberration W_{TL} is minimal when the tilt of wavefront K is equal to coma C_1. Then the wave aberration W_{TL} due to a tilt of the disk substrate becomes

$$W_{TL} = \frac{C_1}{\sqrt{72}} = \frac{t(n^2 - 1)y^3 q}{2\sqrt{72}n^3} \tag{14.A14}$$

Appendix C

The allowable limit of laser-mounting angle is calculated. When the wavefront of the beam emerging from a laser is astigmatic and its axis is not in agreement with the x, y-axis of the optics, there exists a residual astigmatism. With the y-axis as a reference, the phase difference $\psi(x)$ along the x-axis within the pupil plane of the CL with an astigmatic distance of Δ_L and a focal length of f_c is expressed by

$$y(x) = \Delta_L x^2/(2f_c^2) \tag{14.A15}$$

Assuming that the wavefront is inclined through an angle θ with respect to the y-axis, the phase difference $\psi(x, y)$ is expressed as

$$\begin{aligned} y(x,y) &= \Delta_L(x - y\tan q)^2/(2\cos^2 q f_c^2) \\ &= \Delta_L(x^2 + y^2\tan^2 q - 2xy\tan q)/(2\cos^2 q f_c^2) \end{aligned} \tag{14.A16}$$

By focusing the optics, the term $x^2 + y^2 \tan^2(\theta)$ can be zero. Therefore, the wavefront aberration assumes a maximum value in the direction of $x = y = h$:

$$y_o = \Delta_L \tan q h^2/(\cos^2 q f_c^2) \tag{14.A17}$$

Since h/f_c is the NA of the CL, the above equation may be rewritten as:

$$y_o = \Delta_L \tan q (NA_c)^2/\cos^2 q \qquad (14.A18)$$

Since Maréchal's equation gives the relationship between maximum astigmatism ψ_o and wavefront aberration W_{LA} as

$$W_{LA}^2 = \frac{y_o^2}{6} \qquad (14.A19)$$

we obtain

$$W_{LA} = \frac{\Delta_L \tan q (NA_c)^2}{\sqrt{6} \cos^2 q} \qquad (14.A20)$$

REFERENCES

1. Feinleib, J.; de Neufville, J.; Moss, S.C; Ovshinsky, S.R. Rapid reversible light-induced crystallization of amorphous semiconductors. *Appl. Phys. Lett.* 1971, *18*, 254.
2. Hopkins, H.H. Diffraction theory of laser readout systems for optical video disks. *J. Opt. Soc. Am.* 1979, *69*, 4–24.
3. Goodman, J.W. *Introduction to Fourier Optics*; McGraw Hill: New York, 1968; Chap. 6.3.
4. Braat, J. *Principles of Optical Disk System*; Adam Hilger Ltd.: New York, 1985; 7–85.
5. Firester, A.H.; Caroll, C.B.; Gorog, I.; Heller, M.E.; Russell, J.P.; Stewart, W.C. Optical read out of RCA video disk. *RCA Review* 1978, *39*(3), 392–407.
6. Mansuripur, M. Scanning optical microscopy part 1. *Opt. & Photonics News* 1998, May, 56–59.
7. Yoshida, T. Tellurium sub-oxide thin film disk. *Proc. SPIE Optical Disks Systems and Applications* 1983, *421*, 79–84.
8. Saimi, T. Compact optical pick-up for three dimensional recording and playing system. CLEO '82 Pheonix, April 1982.
9. Imanaka, R.; Saimi, T.; Okino, Y.; Tanji, T.; Yoshimatsu, T.; Yoshizumi, K.; Kamio, K. Recording and playing system having a compatibility with mass produced replica disk. *IEEE Consumer Electonics* 1983, *CE-29*(3), 135–140.
10. Hartmann, M.; Jacobs, B.A.J.; Braat, J.J.M. Erasable magneto-optical recording. *Philips Tech. Rev.* 1985, *42*(2), 37–47.
11. Deguchi, T.; Katayama, H.; Takahashi, A.; Ohta, K.; Kobayashi, S.; Okamoto, T. Digital magneto-optical disk drive. *Appl. Opt.* 1984, *23*(22), 3972–3978.
12. Born, M.; Wolf, E. *Principles of Optics;* Pergamon Press: Oxford, 1970.
13. Saimi, T. PD Head for "PD" System, *National Technical Report*, Dec. 1995; Vol. 41, No. 41.
14. Shih, Hsi-Fu. Holographic laser module with dual wavelength for digital versatile disk optical heads. *Jpn. J. Appl. Phys.* 1999, *38*, 1750–1754.
15. Nakamura, H. *Fine Focus 1-Beam Optical Pick-Up System,* National Technical Report, 1986; 72–80.
16. Finck, J.C.J.; van der Laak, H.J.M.; Schrama, J.T. A semiconductor laser for information readout. *Philips Tech. Rev.* 1980, *139*(2), 37–47.
17. Imafuji, O.; Fukuhisa, T.; Yuri, M.; Mannoh, M.; Yoshikawa, A.; Itoh, K. Low operating current and high-temperature operation of 650-nm AlGaInP high-power laser diode with real refractive index guided self-aligned structure. *IEEE J. Selected Topics in Quantum Electronics* 1999, *5*(3), 721–728.

18. Chinone, N.; Ojima, M.; Nakamura, M. A semiconductor laser below allowance of noise due to the optical feedback by adding the high frequency generating circuit. *Nikkei Electronics* 1983, *10*(10), 173–194.

19. ECMA Standardizing Information and Communication System Standard ECMA-267, December 1997.

20. Saimi, T.; Mizuno, S.; Itoh, N. Amelioration of tracking signals by using slit-detection method. *Proc. Conference of Japan Society of Applied Physics* 1987, *34,29a-ZL-7,* 743; Tokyo, March 1987.

21. Oudenhuysen, Ad.; Lee, Wai-Hon. Optical component inspection for data storage applications. *Proc. SPIE Optical Mass Data Storage II* 1986, *695,* 206–214.

22. Tsunoda, Y. On-land composite pregrove method for high tract density recording. *Proc. SPIE Optical Mass Data Storage I* 1986, *695,* 224–229.

23. Saito, A.; Maeda, T.; Tunoda, Y. *Fast Accessible Optical Pick-up, O plus E*; Shingijyutsu Communications: Japan, 1986, *76,* 84–87.

24. Boivin, A. *Théorie et Calcul des Figures de Diffractions;* Press de l'Université Laval: Quebec, 1964; 118–122.

25. Maréchal, A.; Françon, M. *Diffraction Structure des Images;* Masson & Cie: Paris, 1970; 105–112.

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

15

CTP Scanning Systems

Gregory Mueller

MacDermid Printing Solutions
San Marcos, California, USA

CONTENTS

15.1 INTRODUCTION

During the early 1990s a process called CTP (Computer to Plate) was developed to automate the production of printing plates and simplify the workflow of the printing plant. During the development phase of this technology, there were many companies producing CTP devices (platesetters). After the enthusiasm of early adopters waned and the technology matured, it became apparent that total market demand would not support 15 or 20 small platesetter companies and a consolidation of the suppliers took place. Ultimately between five and ten major suppliers of platesetters have had relative success. In recent years, CTP

has been widely adopted in all types of printing plants: from small commercial shops to the largest newspapers. Also, with the continuing rise of the digital delivery of information, what remains of the printing process will continue to be optimized—an understanding of existing platesetters and their architectures becomes increasingly important for those involved in procurement and use of these devices. Platesetters employ a light source (most often laser(s)) with an optical scanning system to expose photosensitive printing plates. This article will briefly describe the different types of scanning systems that are used, a methodology that will show how the different scanning systems are paired with appropriate light sources and printing plates, and examples of platesetters that are currently in the marketplace.

15.2 DESCRIPTION OF TYPES OF SCANNING SYSTEMS

The scanning systems used in platesetters are not typically novel in design—most find their origin in other nonprinting applications. As with all product designs, a series of tradeoffs is made in the attempt to optimize the functionality of the platesetter. With CTP's focus on streamlining the workflow of the printing plant, those tradeoffs often have occurred trying to balance the complication and cost of the scanning systems with those of automatic plate handling mechanisms.

15.2.1 A Note about System Resolution and CTP

It is very important to recognize that while platesetter resolution and quality of the printed matter are directly related (i.e., higher resolution generally results in higher quality), resolution and productivity are inversely related. Therefore, typical resolutions for higher quality printed material (e.g., for magazines) are 2400 dpi and 2540 dpi (1000 cm^{-1}) while newspapers commonly use 1270 dpi (500 cm^{-1}) because higher peak productivity is of greater importance than printing quality in the newsroom.

15.2.2 Internal Drum Scanners

The internal drum scanner is a post-objective scanning system in which the printing plate is mounted on the internal surface of a cylinder with the photosensitive side facing the center of the cylinder. Prior to CTP, this architecture was used in producing large format film negatives used in the exposure of printing plates (a process called Computer to Film [CTF]). This scanning system always uses a laser as its source because of the distance over which the beam must be focused to a small spot, usually with a FWHM equal to 1/Resolution of the system. The beam is first modulated temporally, and then focused through the objective. The beam then is aligned to a 45° mirror or right angle prism spinning on the axis of the cylinder. The reflector directs the beam toward the printing plate surface at normal incidence. These systems are sensitive to misalignment of the beam to the axis of the cylinder. Typically, the waist of the beam occurs at the surface of the printing plate.

The internal drum format's greatest strength is the relative simplicity of the optical system. Its disadvantages are that the use of multiple exposure beams is difficult to implement and this format is inherently inefficient in its use of optical power. Since the plate doesn't extend through the full arc of the cylinder, the overall efficiency is at most the angle
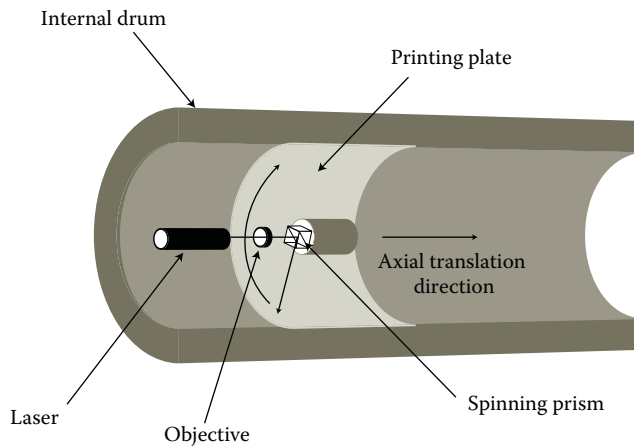
FIGURE 15.1
Internal drum platesetter.

subtended by the plate divided by 360°. (For example, if the plate mounted on the internal drum subtends a 180° angle, the overall efficiency of the system is less than 50% when other losses are factored in.) Figure 15.1 depicts the scheme used by internal drum platesetters.

15.2.3 External Drum

The external drum architecture finds use in the widest range of applications for CTP. Its applications range from engraving rubber flexographic printing plates (hundreds of microns thick) with a CO_2 laser to exposure of offset plates (with a 1-micron thick photo-sensitive laser) with a near-IR or violet diode laser.

The external drum system is not a typical post-objective scanning system—the exposure media is mounted on the external surface of a rotating drum and is scanned past the focussed beam. Temporal modulation of the beam can be by direct modulation of a laser source (e.g., a near-IR laser diode), by acousto-optic modulation of a laser source (e.g., a CO_2 laser) or by modulating a laser beam or some noncoherent source with a "light valve" or MEMS device. (This not only temporally modulates, but also spatially modulates the beam, splitting it into multiple exposing channels.) The light is then focused through an objective onto the printing plate.

One of the limiting factors faced by external drum systems is the maximum drum rotation rate (this is dependent on drum diameter, weight and thickness of the plate, etc.). The maximum data rate (and thus the productivity of the system) may be severely limited unless multiple exposing channels are employed. As the drum spins, the exposure "head" (typically composed of light source(s) and the objective) travels axially along the length of the drum. When multiple exposure channels are employed, they are typically spaced by one pixel in the axial direction. This creates a spiral "swath" that is *N* pixels wide (where *N* is the number of exposure channels). The number of channels employed in external drum systems can vary from one to several hundreds.

Because the drum diameter can be tuned for the specific size of the media being exposed, very little non-imaging time is experienced during the exposure cycle. Another advantage of external drum systems is that the optical system is typically very simple, typically comprised of the light source(s), a means of temporal modulation, and the objective.
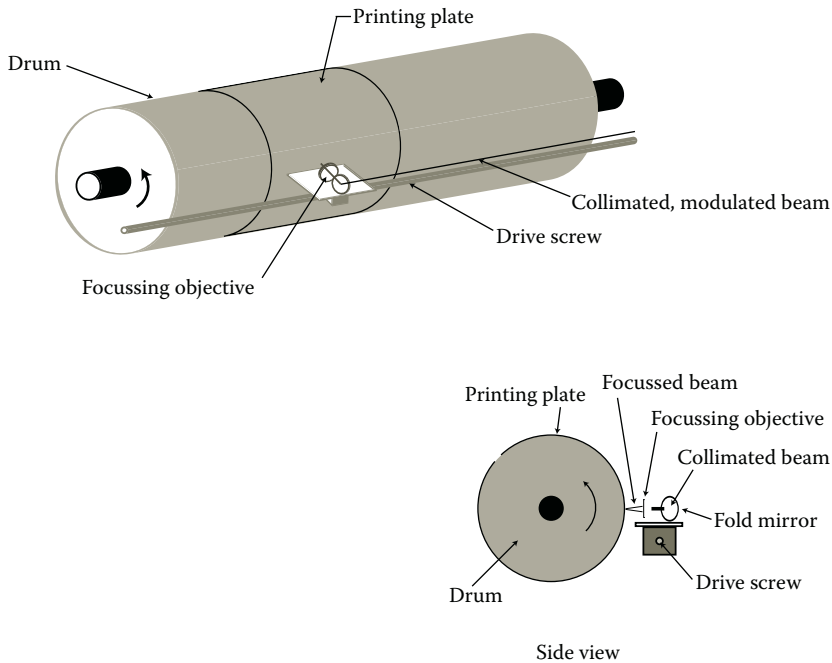
FIGURE 15.2
External drum platesetter.

The use of multiple channels somewhat complicates the manipulation of data and the optical system. Additionally, design of the plate handling system is challenging, since the printing plate must be mounted in register against the drum's external surface.

Figure 15.2 depicts the exposure scheme of a single beam external drum system where the light source is not mounted on the exposure head—an example of this would be a platesetter used to engrave a flexographic printing plate. Though atypical, this is a more general example of the format (in comparison to a system where the light source travels on the exposure head) since it illustrates the requirement that the laser beam must be accurately aligned parallel to the axis of the drum and the head drive screw. If this alignment is not very precise, significant imaging errors will occur along the length of the slow scan head travel.

15.2.4 F-Theta Scan Architecture

The F-Theta Scan system employs pre-objective scanning, that is, the scanning mechanism (usually a rotating polygon) occurs prior to the objective. The objective is designed so that the "focus" occurs in a plane, rather than on a curved surface. If the objective is telecentric, the objective must have a diameter that exceeds the "fast scan" dimension of the plate. (An important characteristic of the objective is that the position of the beam on the plate is proportional to the polygon angle of reflection, rather than to its tangent. The resulting velocity of the beam across the plate in the fast scan axis is nearly constant, which results in consistent plate exposure doses.) Because most CTP exposure media (i.e., offset plates) do not require a significant depth of focus, most systems that use an F-Theta Objective do not employ a telecentric design, resulting in incidence angles at the exposure plane that
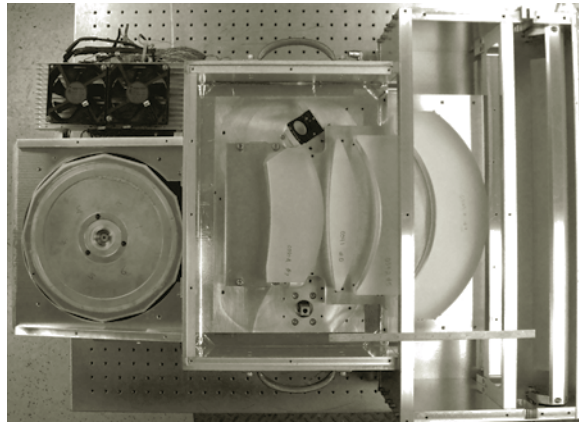
FIGURE 15.3
F-Theta scan objective. (Courtesy MacDermid Printing Solutions.)

may be in the tens of degrees. Because the light must be focused over significant distances, these systems always use a laser as a source. Figure 15.3 is a photograph of the F-Theta Objective (including the polygon) used in the MacDermid Printing Solutions Flexographic Platesetter. This nontelecentric implementation has a scan line length of more than 60 cm with a design resolution of 500 cm^{-1}. (Note: The 12-in scale in the lower right quadrant of the photograph gives some sense of the size of the optical elements in this objective.)

The most significant disadvantage in using an F-Theta scan system is the cost and complexity of the objective itself for high resolution, high efficiency systems. The MacDermid F-Theta objective in the photo above has a material cost of tens of thousands of dollars. Because capital cost is often a primary consideration for printers, this can be a significant marketing hurdle.

15.2.5 BasysPrint Platesetters

BasysPrint's platesetters are a unique example of post-objective scanning and employ a process somewhat analogous to projection lithography. The Texas Instruments DMD chip, a MEM commonly used in projectors and TVs, is used as a spatial modulator to project an image of the appropriate bitmap onto the printing plate through an objective lens. During exposure, the printing plate is stationary while the objective, or "optics head," is translated in three dimensions (translation in the z-axis is for focus). (Because the exposure head projects an image of the DMD onto the plate, the system is sensitive to defocus errors, so the design includes automatic focus correction.) Originally, a short arc lamp was used as the light source, but later versions have employed a 405-nm diode laser array coupled into a fiber bundle. Although the lasers are not coupled into single mode fibers (so the light emerges from the fibers incoherently), this option delivers significantly higher actinic power than the original arc lamp sources.

In its original version, the image was split into "minipictures"—the dimension of each minipicture was the size of the image of the entire DMD array. For a platesetter resolution of 1270 dpi (500 dpcm), the image of the 1024 × 768 pixel DMD had dimensions 2.048 × 1.536 cm. The exposure consisted of the head traveling to the appropriate spot, exposing for a specified amount of time ($D_{\text{BasysPrint}}$, in seconds), then traveling to the next spot approximately 1.5 cm away. This process was repeated hundreds of times per printing plate.

Desired Image

Fast scan direction →

Position 1 DMD state

Position 2

Position 3

Position 4

Position 5

Overall Exposure
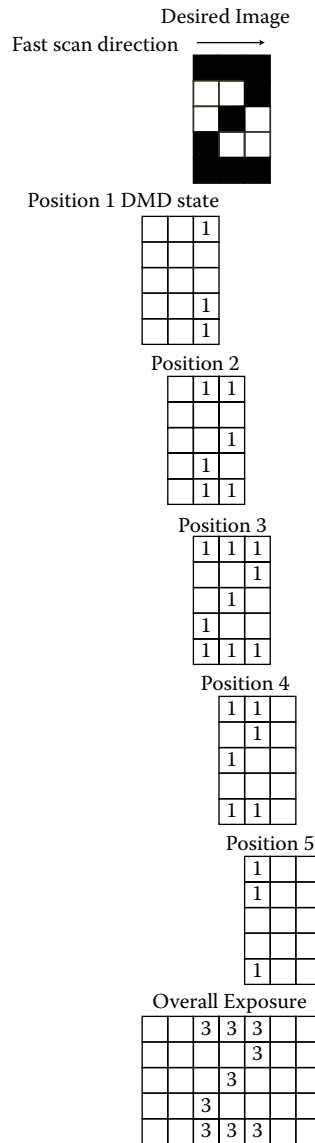
FIGURE 15.4
BasysPrint scrolling exposure.

The more recent versions of BasysPrint's platesetters use the same architecture and optical system, but the image delivery has been changed. The minipictures are now separated by just one pixel in the platesetter's fast scan direction. The process of scanning to the correct position, exposing, traveling to the next position (albeit only one pixel away) while the image to be exposed is being updated, then exposing the next minipicture is fundamentally the same as in the original version of the machine. The difference is that the exposure head no longer stops between minipictures in the fast scan axis, but rather, exposes "on the fly." The net result is that each pixel on the printing plate is exposed by the sum of the pixels used in the fast scan axis of the DMD. This type of exposure creates a "scrolling" effect and has

resulted in a significant increase in platesetter productivity since the time required to start and stop the head during the repositioning between minipictures is no longer required.

Figure 15.4 demonstrates how the scrolling effect is implemented. For the sake of the example, the DMD is reduced to an array consisting of five micromirrors in the slow scan axis and three in the fast scan axis. The image desired on the printing media (a "2") is depicted at the top of Figure 15.4. The other pictures depict the DMD state (0 = "off," 1 = "on") as the optics scrolls across the desired image position on the media, each successive picture separated from the previous one-by-one pixel in the fast scan direction. The bottom picture in Figure 15.4 depicts the integrated exposure of the image on the media.

The DMD has a maximum refresh rate of 40 kHz that is an inherent limit related to the mechanical nature of switching the mirrors on and off.[1] To further increase the productivity of these platesetters, use of new generation DMD chips that are wider (i.e., in the slow scan direction) is being considered.

Because the exposure head moves in all three dimensions over the plate during exposure these platesetters are able to skip over areas where the image is blank. This has the potential to significantly increase the productivity of the platesetter, though the benefit of this feature is totally dependent on image content. BasysPrint has focused the design of their machines on exposing printing plates that historically have been exposed through a film negative in a vacuum frame. These plates tend to be less expensive than other CTP media and BasysPrint's technology has been dubbed "Computer to Conventional Plate," or CTcP.

15.3 METHODOLOGY FOR DETERMINING CTP IMPLEMENTATION

Several equations are derived in the following paragraphs in order to provide a basis for comparing the implementation of the different scanning architectures.

15.3.1 Productivity (plates per hour [pph]), X

Platesetter productivity must meet the printer's peak plate production requirements—this is particularly important for printing plants for which the printed content is delivered just in time and the print run must be delivered under deadline—for example, newspapers. Productivity is a specification of great interest to printers and is typically specified for a given plate type of specific dimensions. The productivity, X (in plates per hour [pph]), is related to the exposure and plate-handling times by the equation:

$$X = \frac{3600}{t_{exp} + t_0} \tag{15.1}$$

where τ_{exp} (the time to expose a plate) and τ_o (the time between successive exposures) are expressed in seconds. The sum of τ_{exp} and τ_o is the platesetter "cycle time."

15.3.2 Plate exposure time, τ_{exp}

Product brochures do not typically specify the exposure time per plate, τ_{exp}, but this variable can be often be determined (with a few assumptions) by using Equation 15.1 and solving simultaneous equations for the different productivity scenarios listed in those brochures. For

systems which scan continuously (internal drum, external drum and F-Theta), one can solve for τ_{exp} by using the fact that the exposure time is simply the exposure width W_{slow} divided by the "velocity" of exposure in the slow scan direction (derivation of V_{slow} is found later):

$$t_{exp} = \frac{W_{slow}}{V_{slow}} \qquad (15.2)$$

15.3.3 Plate handling time, τ_o

The plate handling time clearly has a significant effect on platesetter cycle time (i.e., the elapsed time from the start of exposing one plate until the start of exposing the next plate) and thus, machine productivity. In fact, the time to eject the previously exposed plate and then load and position the next plate in register may exceed the time required to expose the plate. Once one has calculated τ_{exp}, Equation 15.1 leads directly to τ_o.

15.3.4 The Dose Equation

One simple equation relates the sensitivity of the exposure media, the required source power and the rate at which the media is exposed:

$$D = \frac{P_{plate}}{A'} \qquad (15.3)$$

D is the required plate exposure dose (in J/cm^2), P_{plate} is the optical power measured at the exposure media, and A' is the effective area scan rate (dA/dt). (It should be noted that D is a characteristic of the printing plate itself and, for our purposes, is better considered an estimate, since how the optical power is delivered [e.g., as a Gaussian beam or with a "top hat" profile] affects the final results on the printing plate.) Required exposure doses vary widely with the printing technology and plate type as is illustrated in Table 15.1.

TABLE 15.1

Photosensitivities of Various Printing Plates[2]

Printing technology	Plate type	Laser Type/Wavelength	Approx. Exposure Dose [J/cm²]
Offset	Silver Halide	Semiconductor—405 nm FD YAG—532 nm	10×10^{-6}
Offset	Photopolymer	Semiconductor—405 nm FD YAG—532 nm	100×10^{-6}
Offset	High Speed Conventional analog	Semiconductor—405 nm Lamp (365–420 nm)	50×10^{-3}
Offset	Conventional analog	Semiconductor—405 nm Lamp (365–420 nm)	100×10^{-3}
Offset	Thermal	Semiconductor—NIR (830 nm)	150×10^{-3}
Flexography	Photopolymer (MacDermid Laser Plate)	FT YAG—355 nm	30×10^{-3}
Flexography	Black Mask Ablation	Semiconductor—NIR (830 nm)	2.00
Flexography	Rubber Engraving	CO_2—IR (10.6 microns)	200

Source: Values for offset plates from Weber, Robert J. 2008. Computer-to-Plate White Paper. http://www.bob-weber.com/pdfs/CTP%20WP-Web.pdf. (Accessed April 1, 2010.)

15.3.5 Optical Source Power

Since P_{plate} is the power measured at the exposure media, the total optical source power must account for the optical transmission, T:

$$P_{\text{plate}} = P_{\text{total}} \cdot T \tag{15.4}$$

As optical systems increase in complexity, typically the number of optical surfaces increases, along with transmission losses due to imperfect optical coatings. For example, a system with 20 antireflection surfaces, with coating losses of 0.5% per surface suffers a transmission loss of approximately 9.5%.

Another significant transmission loss is due to the method of source modulation. Systems using diode lasers usually directly modulate them, totally mitigating this loss. Another common type of modulation scheme uses an acousto-optic modulator—losses due to these modulators are typically 10%–20%. According to their manufacturer, Texas Instruments, the DMDs used in BasysPrint platesetters have a transmission loss of approximately 35%.[1]

15.3.6 Area Scan Rate

The maximum plate area to be exposed, A (in cm^2), is as follows:

$$A = W_{\text{slow}} L_{\text{fast}} \tag{15.5}$$

The "slow" and "fast" subscripts refer to the respective "slow" and "fast" scan axes of the platesetter. During each fast scan cycle there is a period of active exposure and a period of "dead" time (e.g., when the beam is transitioning from one scan line to the next). These periods of time can be used to define the duty cycle, e:

$$T_{\text{scan}} = T_{\text{active}} + T_{\text{dead}}$$
$$e = \frac{T_{\text{active}}}{T_{\text{scan}}}$$

For systems employing continuous scanning, T_{scan} is simply the inverse of the scan rate, S, so that:

$$T_{\text{active}} = e \cdot T_{\text{scan}}$$
$$= \frac{e}{S} \tag{15.6}$$

To find the rate at which the area of the plate is scanned, we first define the maximum area exposed by a single scan in the fast axis:

$$A_{\text{scan}} = L_{\text{plate}} \frac{N}{R} \tag{15.7}$$

L_{plate} is the maximum plate length accommodated by the platesetter, N is the number of channels employed for exposure, and R is the resolution (in cm^{-1}). The value N/R is the

width of the scan swath in the slow scan direction. Combining Equations 15.5 and 15.6, the average area scan rate, A' (in cm^2/s), is as follows:

$$\frac{\mathrm{d}A}{\mathrm{d}t} = \frac{L_{\text{plate}}(N/R)}{T_{\text{active}}}$$

$$= \frac{L_{\text{plate}}(N/R)}{(e/S)}$$

$$A' = \frac{L_{\text{plate}} \cdot N \cdot S}{e \cdot R} \tag{15.8}$$

It is interesting to note from Equation 15.8 that A' is the product of a velocity in one axis and a width in the orthogonal axis:

$$V_{\text{slow}} = \frac{N \cdot S}{R} \tag{15.9}$$

while the value L_{plate}/e is the fast scan length accounting for the scan duty cycle.

On the other hand:

$$V_{\text{fast}} = \frac{L_{\text{plate}} \cdot S}{e}$$

is the average fast scan velocity while the value N/R is the width of the swath of exposure beams in the slow scan direction. As noted before, the value V_{slow} (from Equation 15.9) is especially useful in deriving the system productivity for platesetters employing continuous optical scanning (i.e., internal drum, external drum, and F-Theta systems).

15.3.7 BasysPrint Area Scan Rate

For the BasysPrint platesetter models, the area scan rate must be derived in a slightly different manner. The exposure head travels in a serpentine fashion over the plate: exposing as the head travels in the fast scan direction across the plate until it reaches the end of the image, stopping, moving to the next scan line, and then exposing while moving in the opposite fast scan direction. The distance between fast scan passes is the image height of the DMD in the slow scan axis. BasysPrint expresses the exposure dose as "light time" (in this chapter it will be referred to as $D_{\text{BasysPrint}}$). The BasysPrint fast scan velocity is simply the image height of the DMD in the fast scan direction divided by the "light time":

$$V_{\text{fast}} = \frac{(N_{\text{fast}}/R)}{D_{\text{Basys Print}}} \tag{15.10}$$

where N_{fast} is the number of fast scan pixels used in the exposure. The DMD's refresh rate can be expressed as:

$$f = \frac{N_{\text{fast}}}{D_{\text{BasysPrint}}} \tag{15.11}$$

One design feature of the DMD is that the refresh rate, *f,* can be increased by using only part of the array (i.e., Texas Instruments has formatted the DMD into smaller blocks that have a maximum refresh rate of 40,000 Hz). This does not hold the advantage it once did: at one point in time, the refresh rate using all of the pixels (768) in the fast scan axis was only 9700 Hz, producing a V_{fast} of 19.4 cm/s at a resolution of 500 cm^{-1}. Texas Instrument's most recent DMD brochure specifies a maximum refresh rate using all 768 of the pixels in the fast scan axis as 32,550 Hz, increasing V_{fast} to 65.1 cm/s (Texas Instruments Incorporated).[3] To use the higher refresh rate, the number of fast scan pixels, N_{fast}, is decreased. In the end, this is transparent to the user since the exposure dose is optimized empirically (i.e., the user does not request a dose of 0.1 mJ/cm^2; instead a series of exposures is made to determine the appropriate "light time.")

The BasysPrint area scan rate is simply the product of the image height of the DMD in the slow scan axis and the fast scan velocity, or:

$$
\begin{aligned}
A'_{\text{BasysPrint}} &= V_{\text{fast}} \cdot W_{\text{slow}} \\
&= \frac{N_{\text{fast}}/R}{D_{\text{BasysPrint}}} \cdot \frac{N_{\text{slow}}}{R} \\
A'_{\text{BasysPrint}} &= \frac{N_{\text{fast}} \cdot N_{\text{slow}}}{D_{\text{BasysPrint}} \cdot R^2}
\end{aligned}
\tag{15.12}
$$

This describes the area rate during exposure, and is used to calculate required source power. However, to calculate the platesetter productivity, one must also account for the time at the beginning of the scan to accelerate to the head exposure velocity V_{fast}, the time at the end of the scan to decelerate, and the time to reposition the optics head for the next scan. The value of this "dead" time ("dead" only in the sense that no exposure is being done) varies somewhat with $D_{\text{BasysPrint}}$ and, as will be demonstrated, significantly impacts productivity. For the calculations a value of $T_{\text{dead}} = 0.6$ s will be used.

To determine the productivity, one must determine the number of fast scan passes required, and the time required for each fast scan pass. The number of fast scan passes is the slow scan image length divided by the width of the DMD image, with *n* rounded up to the next integer since the platesetter cannot make partial passes in the slow scan axis.

$$
n = \frac{W_{\text{slow}}}{(1024/R)}
\tag{15.13}
$$

The time required for each fast scan pass is as follows:

$$
T_{\text{scan}} = \frac{W_{\text{fast}}}{V_{\text{fast}}}
\tag{15.14}
$$

where W_{fast} is the length of scan in the fast scan axis. From Equations 15.10 and 15.14, we see that

$$
\begin{aligned}
T_{\text{scan}} &= \frac{W_{\text{fast}}}{[N_{\text{fast}}/D_{\text{BasysPrint}} \cdot R]} \\
&= \frac{W_{\text{fast}} \cdot R \cdot D_{\text{BasysPrint}}}{N_{\text{fast}}}
\end{aligned}
\tag{15.15}
$$

Total plate exposure time is:

$$t_{exp} = n \cdot T_{scan} + (n-1) \cdot T_{dead} \tag{15.16}$$

The factor $(n-1)$ occurs since T_{dead} occurs only between scans. (Actually we need to add one acceleration time for the beginning of the first scan and one deceleration time for the last scan, but in the larger scheme these are negligible and will be ignored.) Productivity of the platesetter can then be calculated per Equation 15.1.

15.4 SPECIFIC PLATESETTER SYSTEMS

The following section will apply and discuss one specific CTP implementation of each of the architectures. In order to make reasonable comparisons, a platesetter resolution of 500 cm^{-1} will be used (though many platesetters have variable resolutions) and for productivity calculations the standard plate size will be 34.3 cm × 60 cm (a typical newspaper broadsheet).

15.4.1 Fuji Saber V8-HS (Fujifilm Graphic Systems) (Internal drum)[4]

Using two 60 milliwatt laser diodes (405 nm output, with each laser comprising a separate exposure head), this platesetter is aimed at offset printers (though not specifically newspaper printers) and uses printing plates with dose requirements of 10^{-4} to 10^{-5} J/cm^2 (see Table 15.1). The product brochure states that it has a scan rate of 60,000 rpm ($S = 1000$ s^{-1}). Using Equation 15.9, the slow scan velocity is $S/R = 2$ cm/s for each head.

   From the product brochure, the load time can be found by comparing the productivities at two different resolutions: 47 pph at 2400 in^{-1} and 70 pph at 1200 in^{-1} (for an undefined plate width). Combining Equations 15.1 and 15.2:

$$\frac{3600}{X} = t_{exp} + t_0$$

$$= \frac{W}{V_{slow}} + t_0$$

A reasonable assumption is that the plate handling is constant for both resolutions. Solving the two equations simultaneously and using $V_{slow} = 2$ cm/s, yields $W = 106.5$ cm and $\tau_0 = 26.3$ s.

   Returning to our comparison, $\tau_{exp} = 34.3$ cm/(2 cm/s) = 17.2 s, and using Equation 15.1 the productivity per head is determined:

$$X = \frac{3600 \text{ s}}{17.2 \text{ s} + 26.3 \text{ s}} \quad \text{(Exposure cycles per head per hour)}$$

$$= 82.8$$

Assuming that two exposure heads are employed and two plates are loaded per exposure cycle results in a productivity, *X*, of about 166 plates per hour.

The longest plate that can be exposed by this platesetter measures 960 mm in the fast scan direction. If the scan duty cycle, *e*, is assumed to be 50% (remember that the fast scan duty cycle for the internal drum architecture is small compared to both the external drum or the F-Theta scan systems), the area rate per head, from Equation 15.8, is as follows:

$$A' = \frac{96 \text{ cm} \cdot 1 \cdot 1000 \text{ s}^{-1}}{0.5 \cdot 500 \text{ cm}^{-1}}$$
$$= 384 \text{ cm}^2/\text{s}$$

If we assume that 90% of the total source power reaches the plate (the lasers are directly modulated and the optical system is quite simple), the applied dose is (from Equations 15.3 and 15.4):

$$D = \frac{P_T}{A'}$$
$$= \frac{0.060W \cdot 0.9}{384 \text{cm}^2/\text{s}}$$
$$= 1.14 \times 10^{-4} \text{ J/cm}^2$$

This agrees nicely with the dose requirements found in Table 15.1 for offset plates that are sensitive to 405 nm.

15.4.2 Kodak Generation News (Eastman Kodak Company) (External Drum)

The Generation News platesetter is aimed at thermal offset newspaper plates. Kodak's Thermal News Gold offset printing plates have a sensitivity of about 0.1 J/cm² for the near IR spectrum. The Generation News has two exposing heads and employs a diode bar emitting at 830 nm in each head.

The product brochure (Kodak Generation News System Brochure)[5] reports a productivity of 300 pph for 34.3-cm-wide plates and 140 pph for 89-cm-wide plates (the Z option figures). For the narrower plates, each head exposes one plate, so that there are 300/2 = 150 exposing cycles per hour. Equation 15.1 yields:

$$\frac{3600}{150} = t_{\text{exp}} + t_0 = 24 \text{ s}$$

For the wider plates:

$$\frac{3600}{140} = t_{\text{exp}} + t_0 = 25.7 \text{ s}$$

If we assume that τ_0 and V_{slow} are the same for both scenarios (using the same plate sensitivity and same plate handling system), we can solve simultaneous equations by substituting W/V_{slow} for τ_{exp} where *W* is the distance each head travels. This results in $V_{\text{slow}} = 5.95$ cm/s, $\tau_{\text{exp}} = 5.76$ s and $\tau_0 = 18.2$ s. (Note that the plate handling is the dominant factor in the platesetter's productivity.)

Using Equation 15.9:

$$N \cdot S = V_{\text{slow}} \cdot R$$
$$= (5.95 \text{ cm/s})(500 \text{ cm}^{-1})$$
$$= 2975$$

The product of NS is in units of channels/s and this value implies either a great number of channels or a very high drum speed. The Generation News (Z) actually uses 224 channels/head which results in a drum rotation rate of approximately 13.3 revolutions/s.[2]

To find the laser power required per head, the area scan rate must first be found. Remember that the drum circumference can be tuned to maximize the fast scan duty cycle—the only nonimaging space on the drum is the mechanism required for mounting the plate on the drum. For a maximum plate length of 66 cm, a reasonably assumed 2 cm for the required plate mounting mechanism results in a fast scan duty cycle of 97%. Using Equations 15.8 and 15.9:

$$A' = \frac{(5.95 \text{ cm/s})(66 \text{ cm})}{0.97} = 404.8 \text{ cm}^2/\text{s}$$

Substituting this into Equation 15.3 and using a dose requirement of 0.1 J/cm², one finds that about 40 W per head must be delivered to the plate to expose the Thermal News Gold plates.

15.4.3 MacDermid Flexo Platesetter (F-Theta Scanner)

MacDermid's Flexo Platesetter is aimed at newspapers using flexographic printing plates. These plates have a relatively thick layer of photopolymer (ranging from 0.25 to 0.40 mm in thickness) on a relatively thin (~0.17 mm) steel substrate. (The flexographic process uses raised areas on the plate to provide contrast between print and nonprint areas when ink is applied to the plate.) Because the uncured photopolymer layer is soft and the cut edges of the plate are slightly tacky, automatic plate handling is quite challenging—this factor pushed the design toward a flatbed implementation of CTP. The MacDermid platesetter employs a high power, quasi-continuous wave (modelocked) frequency tripled vanadate laser, emitting at 355 nm.

With a required dose of approximately 0.030 J/cm², and 8 W of laser power in the UV (the highest power available at the time the machine was introduced), the process is photon limited (i.e., the productivity of the machine is limited not by mechanics of the exposure process but by the UV power available.) There are approximately 50 optical surfaces and the system employs an acousto-optic modulator. This results in a total optical transmission of about 50% and Equation 15.3 yields an area rate of $A' = 133$ cm²/s.

The polygon has twelve facets and a fast scan duty cycle of approximately 85%. Using four channels to expose, Equation 15.8 yields a scan rate of:

$$S = \frac{A' \cdot e \cdot R}{L_{\text{plate}} \cdot N}$$
$$= \frac{(133 \text{cm}^2/\text{s})(0.85)(500 \text{ cm}^{-1})}{(60 \text{ cm})(4)}$$
$$= 236 \text{ scans/s}$$

(With twelve facets, the polygon rotation rate is thus approximately 20/s or 1200 rpm.) To find the machine productivity, we find the slow scan velocity, from Equation 15.9:

$$V_{slow} = \frac{N \cdot S}{R}$$

$$= \frac{4 \text{ channels/scan} \cdot 236 \text{ scans/s}}{500 \text{ cm}^{-1}}$$

$$V_{slow} = 1.89 \text{ cm/s}$$

The time to expose a plate (Equation 15.2) is calculated to be τ_{exp}=18.2 s and with a plate handling time of $\tau_0 \sim$ 15 s, machine productivity, X, is calculated to be about 109 plates per hour.

MacDermid has found that the relatively dirty newspaper environment presents a significant challenge in keeping optics clean. With peak UV intensities of approximately 50 MW/cm^2, purging the optics assembly with clean, dry air has proven useful in reducing required maintenance and extending the life of expensive optical components.

15.4.4 BasysPrint Series 6 Platesetters (Punch Graphix International)[6]

Most of BasysPrint's platesetters are aimed at offset printers, though flexographic newsprinters have also installed some of these machines. Assuming the exposure of a high-speed conventional offset plate, the required "light time" and source power will be calculated. For the productivity calculations, the product brochure does not provide enough information to determine the plate handling time, so for the sake of our comparison, we'll assume a plate handling time of 15 s.

From the product brochure, the platesetter productivity is 60 plates per hour for a plate measuring 60.5 cm × 74.5 cm. To be able to make a comparison with the other platesetters using a plate size of 34.3 cm × 60 cm, the required V_{fast} must be determined. Equation 15.1 allows a calculation of τ_{exp}:

$$t_{exp} = \frac{3600}{X} - t_0$$

$$= \frac{3600}{60} - 15$$

$$= 45 \text{ s}$$

To find the head velocity, the time for each scan, T_{scan}, must be found. Using $n = 30$, (60.5 cm/ (1024/500) rounded up to the next integer) and $T_{dead} = 0.6$ s, Equation 15.16 yields:

$$T_{scan} = \left[\frac{t_{exp} - (n-1) \cdot T_{dead}}{n} \right]$$

$$= \left[\frac{45 \text{ s} - 29 \cdot (0.6 \text{ s})}{30} \right]$$

$$T_{scan} = 0.92 \text{ s}$$

Assuming that the plate is oriented with the 74.5-cm dimension in the fast scan direction and that the platesetter uses two heads to expose this plate results in each head having a fast scan width of 37.25 cm. V_{fast} thus is 40.49 cm/s (37.25 cm/0.92 s). Using N_{fast} = 192 pixels in Equation 15.10, $D_{BasysPrint}$ is as follows:

$$
\begin{aligned}
D_{BasysPrint} &= \frac{N_{fast}/R}{V_{fast}} \\
&= \frac{192 \text{ pixels}/500 \text{ cm}^{-1}}{40.49 \text{ cm}/\text{s}} \\
&\approx 9.5 \text{ m s}
\end{aligned}
$$

(It should be noted here that the example from the product brochure does not represent the peak productivity of the platesetter, since the minimum "light time," $D_{BasysPrint}$, is 4.8 ms. [From Equation 15.11, with a maximum refresh rate of 40kHz.])

To compare the productivity of the BasysPrint with the other platesetters, T_{scan} can be determined using the correct fast scan width in Equation 15.14.

$$
\begin{aligned}
T_{scan} &= \frac{34.3 \text{ cm}}{40.49 \text{ cm}/\text{s}} \\
&= 0.85 \text{ s}
\end{aligned}
$$

Using n = 30, (60 cm/(1024/500) rounded up to the next integer) and T_{dead} = 0.6 s, Equation 15.16 yields:

$$
\begin{aligned}
t_{exp} &= 30 \cdot (0.85 \text{ s}) + 29 \cdot (0.6 \text{ s}) \\
&= 42.9 \text{ s}
\end{aligned}
$$

Using Equation 15.1 (and remembering that two exposure heads are used) the productivity can be determined:

$$
\begin{aligned}
X(/\text{head}) &= \frac{3600}{42.9 \text{ s} + 15 \text{ s}} \\
&= 62.2 \text{ plates/h/head} \\
X(\text{overall}) &= 124.4 \text{ plates/h}
\end{aligned}
$$

To find the required source power, the active scan rate must be determined using Equation 15.12:

$$
\begin{aligned}
A'_{BasysPrint} &= V_{fast} W_{slow} \\
&= (40.49 \text{ cm}/\text{s}) \cdot (2.048 \text{ cm}) \\
&= 82.9 \text{ cm}^2/\text{s}
\end{aligned}
$$

For the required dose of 0.050 J/cm², the required power at the plate is determined from Equation 15.3:

$$P_{\text{plate}} = A' \cdot D$$
$$= (82.9 \text{ cm}^2/\text{s})(0.050 \text{ J/cm}^2)$$
$$\approx 4.1 \text{ W}$$

Assuming an overall optical efficiency of about 40%, a total of 10.4 W of source power is required per exposing head.

15.5 SUMMARY

When one examines the equations that describe the CTP process it becomes evident that any of the architectures are adequate for the exposure of photosensitive printing plates. In most cases, the different platesetter architectures were developed because of specific requirements of the exposure media and the available source powers. The principle of keeping systems as simple as possible continues to rule, particularly when customer budgets are tight.

If one used a decision tree to choose successful CTP architectures one would probably end up on an External Drum branch. The External Drum architecture has a relatively simple optical design, with flexibility of choice both in the number of channels and the optical sources that can be used. As a result, one finds it in the widest variety of platesetters, from those exposing UV offset plates to those engraving thick flexographic printing plates with infrared wavelengths (10.6 microns).

By comparison, the Internal Drum format is simple, but is limited by the number of channels that can be used and requires the use of a laser with high beam quality that is used quite inefficiently. This tends to limit the use of the internal drum to low-power, low-dose applications as is found with exposure of 405 nm or 532 nm offset plates.

The F-Theta architecture tends to compete with the External Drum in printing plants where peak productivity is valued. Plate handling automation is easiest with this architecture, but because the source has to be a relatively high-quality laser beam, the architecture lends itself to low-power, low-dose applications. Both offset (sensitive at 355 nm, 405 nm, and 532 nm) and flexographic plates (sensitive at 355 nm) are exposed using this architecture. The cost of a high quality F-Theta objective is a significant hurdle to its wider acceptance in these markets.

BasysPrint's architecture finds its niche with printers desiring to expose lower cost media that has historically been exposed in vacuum frames through film negatives. The novel use of the Texas Instrument's DMD with either an arc lamp or semiconductor laser array as optical source enables this type of exposure. BasysPrint platesetters have also found application in flexographic newspaper plants.

REFERENCES

1. Dudley, D.; Duncan, W.; Slaughter, J. Emerging Digital Micromirror Device (DMD) Applications. *SPIE Proceedings* 4985.14, 2003.
2. Weber, R.J. 2008. Computer-to-Plate White Paper. http://www.bob-weber.com/pdfs/CTP%20 WP-Web.pdf. (Accessed April 1, 2010.)

3. Texas Instruments Incorporated. DLP Discovery 4100 Development Kit - DLPD4X00KIT - TI Tool Folder. http://focus.ti.com/docs/toolsw/folders/print/dlpd4x00kit.html. (Accessed April 23, 2010.)

4. Fujifilm Graphic Systems U.S.A., Inc. 2008. *SABER V-8 HS*. http://www.fujifilmgs.com/pages/8_up/64.php. (Accessed April 1, 2010.)

5. Eastman Kodak Company. 2009. *Kodak Generation News Systems*. http://graphics.kodak.com/US/en/Product/computer_to_plate/ctp_for_newspaper/Generation_News_System/default.htm. (Accessed February 15, 2010.)

6. Punch Graphix International NV. 2008. *Serie 6_downl_eng.pdf*. http://www.basysprint.com/en/products. (Accessed April 1, 2010.)

16

Synchronous Laser Line Scanners for Undersea Imaging Applications

Fraser Dalgleish, PhD

Ocean Visibility and Optics Laboratory
Harbor Branch Oceanographic Institute at Florida Atlantic University
Fort Pierce, Florida, USA

Frank Caimi, PhD., P.E.

Ocean Visibility and Optics Laboratory
Harbor Branch Oceanographic Institute at Florida Atlantic University
Fort Pierce, Florida, USA

CONTENTS

16.1 INTRODUCTION

The ability to observe objects underwater has been of great interest historically and many accounts exist in the quest to see "what lies beneath the surface of the sea." Unfortunately, the physical processes of absorption and scattering from water molecules and suspended particulates hinder the ability to see at great distances, and mandate the use of specialized image formation techniques to extend or even make practical the use of underwater optical imaging systems.

Undersea imaging systems are classified primarily in two categories: *conventional* and *advanced*. Conventional systems utilize either ambient or artificial lighting and a film or video camera. Advanced systems typically require specialized lighting, for example a laser, and some method for eliminating light that is primarily scattered within the medium and not returned from the object of interest in the illuminated field of view. Film cameras, CCD

cameras strobe lights, arc lights, and, more recently, HDTV cameras with LED lighting constitute *conventional* imaging systems. Advanced imaging system architectures have been proposed long before the advent of practical laser systems, and have been implemented using a variety of geometric configurations to mitigate the effects of scattering and absorption.

Conventional camera systems using adjacent broad spectrum light sources for illumination are useful for imaging surfaces at imager to target distances of one to two beam attenuation lengths. An attenuation length is the distance light must travel to be reduced to $1/e$ of its original intensity. It is typically 20 to 30 m in clear oceanic water, and can be less than 1 m in turbid coastal waters. It has been found that at imager to target distances of about three attenuation lengths, acceptable imaging can be provided by spatially separating the light source from the camera, that is, by using a flood light to illuminate the target region. However, as the scattering coefficient increases, the level of common volume scatter increases, and this creates a loss of signal-to-noise ratio (SNR), contrast and resolution ultimately leading to a contrast-limited image.

Advanced imaging systems concepts using laser sources typically provide imaging at distances greater than three attenuation lengths. These extended range imagers are generally of two classes: the synchronous laser line scanner (LLS) class and the laser range gated (LRG) class.

Synchronous LLS is a serial imaging system that provides scanning capability over wide swath (up to 70°) usually with a continuous wave (CW) laser source which is continually tracked on the target over a scan line by a narrow instantaneous field of view (IFOV) single element detector (shown in the left hand side diagram in Figure 16.1) in order to reduce the common scattering volume. Based on results of controlled experimentation and analytical modeling, synchronous scanners have been found capable of operation at maximum distances greater than five attenuation lengths, believed to reach a limit due to shot noise from multiple near-field backscatter in turbid waters, whilst becoming forward scatter limited in clearer water.[1–3] Such imagers have been under continued development for use aboard undersea vehicles, including towed bodies, manned submersibles, Autonomous Underwater Vehicles (AUVs) and Remotely Operated Underwater Vehicles (ROVs) to provide imagery for characterizing the sea floor for varied activities including military missions, scientific benthic survey, and inspection of oil and gas infrastructures (Figure 16.1).

Further improvement in imaging range benefits undersea operations by allowing increased vehicle speed and maneuverability and improved image resolution at greater
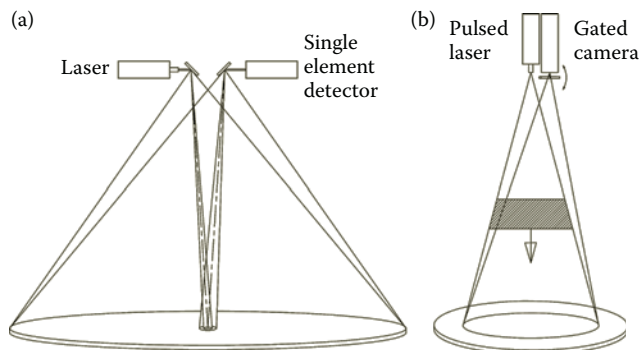
FIGURE 16.1
Geometric configuration of an underwater laser line scan (LLS) system using synchronous scan method for image formation (a) versus the laser range gated (LRG) system (b).

distances from target regions. The ability to more rapidly and reliably produce higher resolution images of targets and survey sites from greater distances will enable a more extensive and diverse set of applications for underwater vehicles. Depending on the size and complexity of surfaces in the target region, optical sensing may be the only effective means for characterizing features.

By way of example, in the exploration of unknown or dynamic environments, rapid topographical seabed variations can occur at rates greater than the vertical axis performance of the AUV. It is therefore necessary to "fly" the vehicle at a sufficient range above the seabed to avoid potentially catastrophic collisions. Consequently, the design of underwater optical scanning systems must accommodate varying distances to the object plane.

It has also been shown by both simulation and experimentation that the class of laser range-gated (LRG) imagers (shown in the right-hand side diagram of Figure 16.1), that is, those imagers utilizing a pulsed laser source, are also capable of adequate underwater performance for imaging target regions at distances greater than five attenuation lengths, albeit through a narrower total swath than LLS. These systems minimize introduction of energy due to scattered light with divergent laser pulses synchronized with gated intensified cameras.[4–7] One advantage of this approach over LLS is that it is not necessary to perform along-track image registration, since the entire FOV is imaged concurrently.

There have also been demonstrated efforts which utilize synchronous scanning with high repetition rate pulsed sources and single element gated detectors.[2,8] Although these imagers ultimately become limited due to forward scatter between the source and the target, as well as return path attenuation, they can be more compact than CW LLS systems because a spatial offset between the source and receiver is not required to reject scattered light.

Summarily, both classes of extended range underwater imagers are hindered in performance by the point spread and attenuation of the medium, as the laser beam travels from the source to the target region with only a small portion returning to the detector. Scattering and attenuation cause losses in contrast, resolution and SNR. These losses are particularly problematic at and near the range limit of operation.

As a consequence of the requirement to minimize the common scattering volume LLS system designs have a relatively small depth of field (DOF), typically no more than a few meters. This is particularly problematic when imaging in a dynamic undersea environment in which there is significant variation of optical transmission properties, seabed surface features or in which there is significant variation in platform altitude or attitude. Commonly, these factors can lead to unacceptable degradation in image quality or complete signal loss. As shown in Figure 16.2 the DOF is a function of the
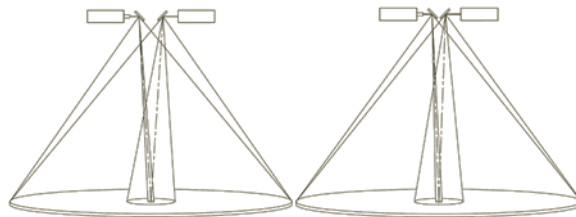
FIGURE 16.2
Geometry showing effect of source-receiver separation on DOF and common volume scattering for line scanner architectures.

source–receiver separation distance, the optical path length to and from the target (stand-off distance), beam divergence, and the acceptance angle of the receiver. Figure 16.3 shows how the receiving aperture of the LLS system may be widened to improve DOF. Alternately, a fine adjustment of the optical focus may be slaved in accord with an on-board altimeter.

The optical resolution achievable with a LLS system is dependent on the laser beam diameter at the reflecting surface in the target region, and is also dependent upon the precision with which the receiver can resolve intensity information from the return signal as a function of the scan angle. Minimizing the IFOV, for example, by minimizing the receiver spot size at the target, reduces the scattering volume, which can improve the SNR. That is, the imaging range of the system can be improved by reducing the size of the scattering volume, albeit with a reduction in DOF. This is shown conceptually in Figure 16.3.

Reducing the IFOV reduces the target area per pixel, commonly measured in cm^2 per pixel and, theoretically, improves image resolution. This is particularly desirable when imaging target surfaces having a high spatial frequency, as the combined effects of forward scattering and blurring, due to the limited DOF, further limit the achievable resolution.

The sections that follow describe the optical principles of two different synchronous LLSs. The first system, the concept of which was originally patented in 1973.[9] has been used since the early 1990s in various packages.[3,10] The model used in the raytrace analyses in Section 16.3 uses a source–receiver separation of 40 cm. The second system has a smaller source–receiver separation (23 cm) and was developed to bench test alternate system configurations of laser line scan that utilize pulsed and modulated-pulse laser sources which are believed to offer significant potential performance improvements over existing systems. The chapter is organized as follows. Section 16.2 provides a brief historical background to the development of laser line scan underwater imaging apparatus.

Section 16.3 introduces the optical design principals of two alternate types of laser line scan.

Section 16.4 continues a detailed technical discussion by exploring requirements for field-stop apertures in the focal plane to both maintain wide-swath operation and accommodate changes in focal distance.

Section 16.5 presents some recent test tank results generated with one of the featured synchronous line scanners to investigate performance trade-offs with alternate system parameters.
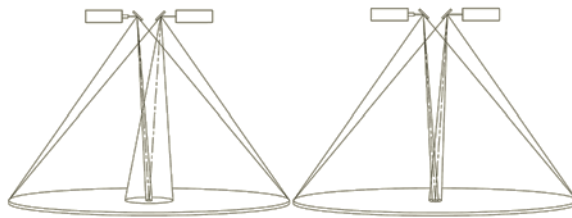
FIGURE 16.3
Geometry showing IFOV effects on common scattering volume reduction, DOF, and image resolution for line scanner architectures.

Section 16.6 concludes with an overview of the main optomechanical desirables for wide-swath extended range undersea laser imaging systems, together with a brief description of future possibilities.

16.2 LLS SCANNING SYSTEM HISTORICAL DEVELOPMENT

Although originally conceived in the early 1970s[11] synchronous laser line scan system development began in the 1980s and early 1990s with a series of Navy R&D contracts. Scientists from private contracting firms in the San Diego area were involved in the early phases of development, including Spectrum Engineering Incorporated (SEI). In 1988, SEI began development of the dual pyramidal polygon system designed as an underwater imaging tool, which achieved significant advancements in operational range, field of view, and image quality over conventional imaging systems. During the early 1990s several proof of concept tests were conducted in the field. In tests off San Diego the towed system imaged a sunken WWII torpedo bomber in a "shoot-off" with another laser system known as Wide Area Imaging System (WAIS) as well as several other state-of-the-art intensified camera systems. Commercially oriented tests were also conducted in March 1993 at British Gas Subsea Engineering Centre in Blyth, England in static and towed performance tests in a large test tank in a comparison with SIT, ISIT, and other CCD cameras.[12] The performance of the LSS was clearly superior to the other state-of-the-art imaging systems in this application, and the system was further developed to address a commercial market using an embedded microprocessor control and automated control functions. The commercial system was delivered in October 1992 and was known as the SM 2000. The system was housed in a pressure vessel 80 in long, 11 in in diameter and required 5 kW of power. Over the years it has been used with some success for a variety of habitat survey operations[10,13,14] and has also been utilized by the Navy for object identification purposes.

Another laser imaging device emerged from Navy contracts into commercial availability at this time. SPARTA, a high technology company with a laser systems laboratory in San Diego developed a range-gated laser imaging device.[5,6] The system was tested by the Coastal Systems Station in a head-to-head comparison with the LLS system with disappointing results, which may have limited its use by the Navy.[3]

In 2001, the Office of Naval Research sponsored an assessment of competing electro-optic identification (EOID) sensors.[15] In this assessment, a streak tube imaging lidar (STIL) and two LLS systems were mounted in a tow vehicle along with a suite of environmental sensors. As well as providing the opportunity to assess system performance in challenging environmental conditions, these exercises created a wealth of image data that was useful for performance prediction model validation.

For almost two decades LLS has widely been regarded as the optimal method for acquiring optical identification-quality imagery of the underwater environment. However, there was clearly a need to research performance improvements and to validate modeling and performance prediction software used by the Navy. In 2006 a dedicated underwater LLS test laboratory was built at Harbor Branch Oceanographic Institution (Fort Pierce, FL), now a campus of Florida Atlantic University. A more compact benchtop LLS system was developed in collaboration with Lincoln Laser (Phoenix, AZ), and the details of this system will also be covered in the sections that follow.

16.3 OPTICAL DESIGN PRINCIPALS FOR UNDERWATER LLS IMAGING SYSTEMS

16.3.1 Dual Pyramidal Line Scanner

The first design being considered is based on the original LLS, which, as described in Section 16.2, has been implemented in a cylindrical housing and used on various platforms for a range of seabed imaging tasks. It consists of two pyramidal polygons driven by a single shaft motor, as shown in Figure 16.4.

The laser beam is reflected by each facet of the rotatable polygon toward the target region and a portion of the returning flux is reflected by a second larger rotatable polygon toward the detector, which usually consists of a collection lens, field-stop aperture and a photomultiplier tube (PMT). The two pyramidal mirror assemblies are synchronously coupled along a common axis about which they are each symmetrically aligned. The common axis is coincident with shafts coupling each assembly to a scan motor system that rotates both the assemblies. The pyramidal polygon mirrors have four plane mirror faces or facets which are triangular and which each extends to an apex. As the assembly rotates, the beam incident on a mirror face is reflected at a variable angle that generates a scanning line in space. Thus, for each revolution of the first assembly, four scan cycles are generated, each over a 90° maximum scan angle. The system typically uses the downward 70° of each scan line for imaging.

Reflections created in the target region by the scanning beam form the diffuse portion that is collected from scanning reflections across the rotating faces of the second larger
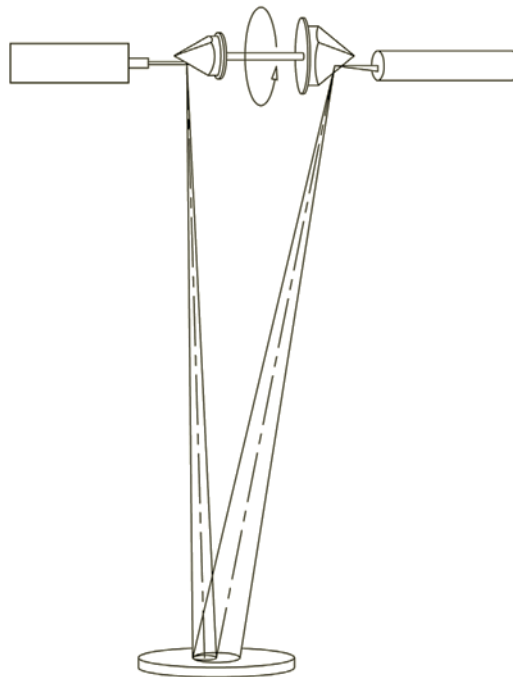
FIGURE 16.4
Dual pyramidal polygon LLS concept diagram.

pyramidal polygon mirror and condensing optics for input to the PMT. An aperture assembly resides at the focal point of the condenser lens to govern the angular extent of rays entering the PMT. The system of Figure 16.4 may include other common components such as cylindrical correction lenses in order to improve optical throughput.

A necessary characteristic to obtain high contrast images is minimal overlap of the illumination light cone with the receiver IFOV prior to the target plane. Therefore, the receiver must maintain a restricted IFOV, and it is not uncommon to require something substantially less than 5 mrad. In addition, the collection aperture must be large enough to collect sufficient number of photons to extend the range limit of the system. Since the number of photons collected is proportional to the square of the incident aperture diameter, a typical practical aperture may be on the order of 50-mm diameter or more. The mirrors required by the scanner must therefore be relatively large, and contribute to the overall size requirements of the deployed system.

One characteristic of this system is that the curvatures of the output scan field of the transmit polygon and the receive polygon scan field are in the opposite direction. This can require an undesirably large receiver IFOV in order to capture the reflected laser light as it scans across the receive polygon and is reflected back to the PMT. It can also lead to the collected irradiance spot walking over the focal plane of the collection lens and the sensitive area of the PMT, thus requiring a complex aperture assembly and PMT with a large photocathode, respectively.

However, this design is well suited for packaging in cylindrically shaped housings, which are typically tolerant to high ambient pressure and characteristic for some towed bodies and AUVs. In operation, a scan line is developed cross track to the vehicle motion, with the second scan axis being provided by the forward platform motion.

Due to the two-dimensional image being formed through forward motion of the vehicle, the scan rate must be sufficient to scan a large angular field before the vehicle advances too far compared to the along-track resolution requirement Δx. Vehicle maximum speeds, dx/dt are typically several meters per second, so that scan rates of hundreds of lines per second are required. An example calculation for a forward speed of 2 m/s and 1 cm resolution yields:

$$R_{\text{scan}} \approx \frac{dx/dt}{\Delta x} = \frac{2\text{m/s}}{0.01\text{m}} = 200\,\text{lines/s} \qquad (16.1)$$

For an $n = 4$ faceted mirror, this implies a rotation rate ω according to:

$$w = \frac{R_{\text{scan}} \bullet n}{4} = \frac{200\,(\text{lines/s}) \times 60\,(\text{s/min})}{4} = 3000\,\text{RPM} \qquad (16.2)$$

These rates are easily achievable, and clearly, even higher resolution is possible with higher scan rates. It should be noted that for the 70° scan swath utilized with these LLSs, the useful part of the scan duty cycle is 70/90 or 77.6%. A single line scan time τ_L is given by:

$$t_L = \frac{\text{Scan Duty Cycle}}{R_{\text{scan}}} = \frac{0.776}{200\,\text{lines/s}} = 0.00388\,\text{s} = 3.88\,\text{ms} \qquad (16.3)$$

Assuming a cross-track angular resolution δ or IFOV of 2 mrad (0.114 deg) corresponding to the receiver angular aperture, the imaging detector must be able to resolve each pixel in a time given by:

$$t_r = t_L \frac{d}{\text{FOV}} = 3.88\,\text{ms}\,\frac{0.114\,\text{deg}}{70} = 0.0064\,\text{ms} = 6.4\,\mu \qquad (16.4)$$

The minimum cross-track number of pixels N is given by:

$$N_{\text{crosstrack}} = \frac{\text{FOV}}{d} = \frac{70}{0.114} = 614 \qquad (16.5)$$

However, in practice higher sampling is used to allow for better resolution, and also to facilitate sample integration to improve SNR when necessary.

Typically, the postdetection optics is comprised of a telecentric collection system, where the angular acceptance angle is determined by the focal length of the system and a controllable aperture of diameter d. Ideally, the scanner should allow the collection system to operate with a fixed aperture, implying that the angular deviation of the collected photon bundle be constant regardless of scan angle or distance to the object plane. For a fixed aperture D, the instantaneous angular aperture δ would be given according to:

$$d = \tan^{-1}\frac{d}{2f} \qquad (16.6)$$

Assume for example, a scanner exit aperture of $D_{\text{exit}} = 50$ mm. A practical lens focal length f of 100 mm, that is an $F\# = 2$, would require an aperture given by:

$$d = 2f\tan d = 200\,\text{mm} \times \tan(0.002\,\text{rad}) = 0.4\,\text{mm} \qquad (16.7)$$

However, due to the known scan deviations with this design, the collected irradiance maximum will move with both scan angle and target distance variations. This effect will be examined in Section. 16.4.

16.3.2 Single Hexagonal Polygon Line Scanner

The second design to be considered consists of a single hexagonal polygon, which uses two symmetrical steering mirrors to adjust for changes in target distance (shown in Figure 16.5). It was designed, fabricated, and tested collaboratively between Harbor Branch Oceanographic Institute at Florida Atlantic University (Fort Pierce, FL) and Lincoln Laser (Phoenix, AZ). The broad design requirement was the need for a wide-swath scanner that is compatible with a small active area detector through various target distances.

This scanner system uses a hexagonal shaped polygon and two symmetrical steering mirror assemblies (also shown in Figure 16.9) to synchronize the laser beam transmit path with the return path to the PMT through an entire line scan. When one facet is positioned to direct the laser beam along the transmit path, another facet is positioned to reflect radiation along the detector signal path toward the telecentric collection optics and field-stop aperture which govern the receiver IFOV. The steering mirrors remain symmetrical about
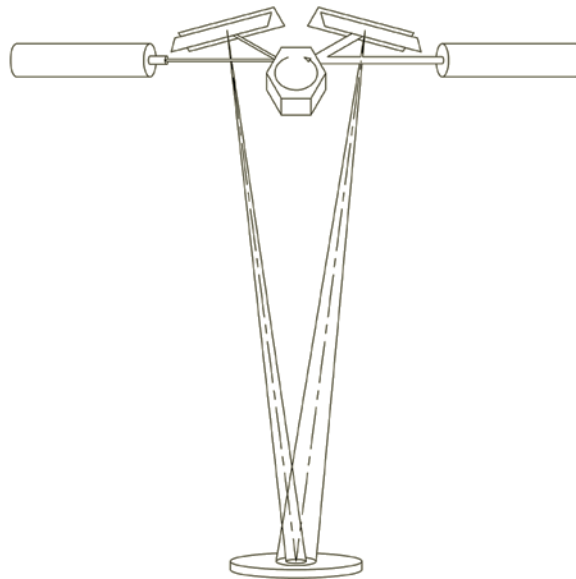
FIGURE 16.5
Single hexagonal polygon LLS concept diagram.

the center of the polygon axis, and only require adjustment to focus at a different target plane. The polygon assembly is coupled to a motor drive system for rotation at steady state speeds in the range of 1000 to 4000 rpm, although higher or lower rotational speeds may be selected based on required performance criteria, within the prevailing environmental and operational conditions. Rotation of the polygon transforms the transmitted beam from a static path into an outgoing scanning path. This rotation simultaneously transforms the static narrow IFOV receiver into a scanning narrow IFOV receiver, synchronized with the laser beam at the target plane through the entire line scan. Lines are scanned through a maximum of 120° from each polygon facet, limited by the length of the steering mirrors to utilize 70° around the nadir. This scan architecture was conceived to substantially reduce the scan deviations inherent in the previously developed systems, making it possible to use a simple aperture and small active area detector.

The outgoing steering optics comprises lower and upper steering mirrors, positioned orthogonally to reflect and thereby direct the scanning portion of the laser beam toward the target plane, as shown in Figure 16.9. The incoming steering optics comprises lower and upper steering mirrors, also positioned orthogonally to reflect and thereby direct the scanning portion of the return path from the same target region toward the polygon facet producing a reflection on a static path toward the PMT.

The prototype was tested with a wedged plane port interface into a test tank thereby limiting the total effective scan angle to be less than 70°. Imaging results using this scanner design are presented in Section 16.4.

16.3.3 Summary

Two alternate wide scan angle line scanner designs have been presented to illustrate different performance characteristics. Both systems have a collection area of approximately 20 cm², corresponding to the area of one polygon facet, to collect enough light for

image formation at extended ranges. The transmit and receive optical paths have been explained for each system up to the collection lens, and the theory by which each design has the potential to achieve synchronous scanning and reflected light collection over a wide total scan angle of up to 70° has been given. A key distinction between the two systems is that the hexagonal polygon system has the ability to adjust the focus distance at the incident aperture of the receiver path via steering mirrors, whereas the pyramidal polygon system uses a more sophisticated aperture assembly at the focal point of the collection optics.

The next section uses results from raytrace simulations of the two designs to examine the distribution of irradiance at the collection lens focal plane through a complete scan line. Furthermore the effect of variations in target distance is also simulated and the required field-stop aperture for each system is determined and discussed.

16.4 RAYTRACE STUDY: FOCAL PLANE APERTURE REQUIREMENTS

The primary optical requirement of both line scanner imagers for use in scattering-dominant waters, is to provide a narrow IFOV which is spatially synchronous at the target plane with the laser beam illumination, over a wide scan angle of up to 70°. This section uses optical raytrace simulation results using Lambda Research TracePro® to examine the spatial characteristics of the irradiance spot at the focal plane of the collection lens. Deviation of the collected irradiance spot due to changes in stand-off distance is also studied to examine the implications of the limited DOF of these designs to realistic environmental conditions. This information is critical to understanding the aperture design of the alternate systems to meet the primary optical requirements. Two key attributes of both scanning designs are analyzed:

1. The position in the focal plane of received light as a function of the scan position
2. The position in the focal plane of received light as a function of the stand-off distance between the imager and the target

Simulated irradiance profiles at the focal plane of the collection lens were produced for five optical scan angles (−35°, −20°, 0°, +20°, +35°) at three stand-off distances (5.2 m, 7.2 m, and 9.2 m).

16.4.1 Dual Pyramidal Polygon Line Scanner

A model for the dual pyramidal polygon line scanner was created which has a source–receiver path separation of 40 cm to be consistent with the existing systems. To improve ray collection efficiency the target was modeled as the inside surface of a cylinder having a radius of curvature equivalent to the target distance (in this case 7.2 m) centered on the scan output. The collected ray bundles were aligned at the center of the focal plane disk for the 7.2-m stand-off distance case. For the simulated cases at 5.2 m and 9.2 m, the target curvature was changed accordingly.

The model is shown in Figure 16.6 for nadir (0°) case as well as +35°. It can be seen that an additional aperture was used prior to the collection lens as a stop for stray light for the purpose of this study. In the implemented system the angular aperture is limited by a field-stop aperture in the focal plane.

A summary of the raytrace results for the dual pyramidal polygon LLS is shown in Figure 16.7.

It can be seen from Figure 16.7 and the composite overlaid images of Figure 16.8 that the collected target irradiance walks across the face of the focal plane disk due to both the scan angle variations (cross-track walk-off of almost 2.5 mm for 5.2 m case, shown in the vertical axes of the plots in Figures 16.7 and 16.8) and changes in target stand-off distance (more than 1.5-mm along-track walk-off between 5.2 m and 9.2 m cases, shown in the horizontal
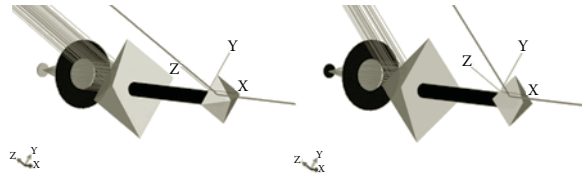
FIGURE 16.6
Dual pyramidal scanner system modeled with back end optics showing a ray trace scan angle of 0° (left) and 35° (right).
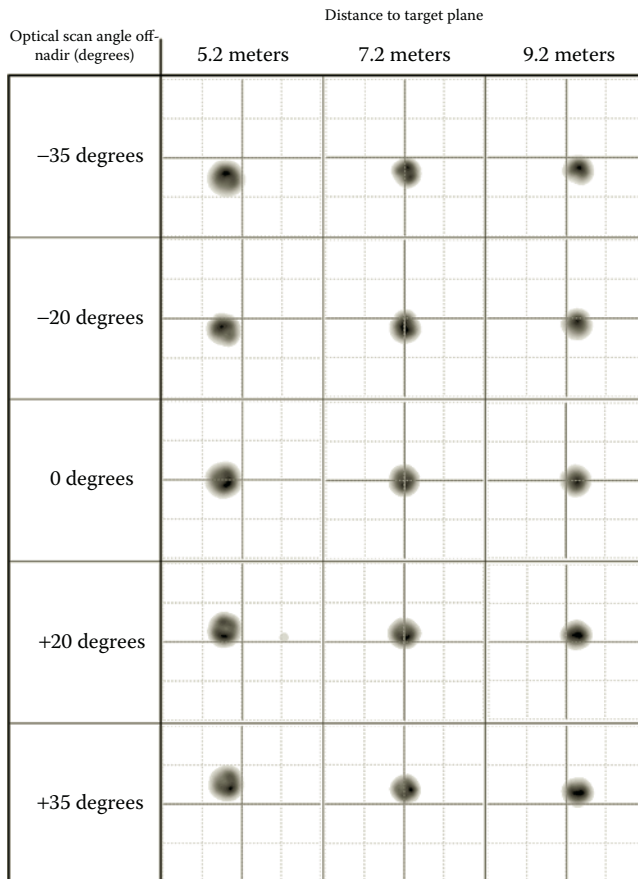
FIGURE 16.7
Dual pyramidal polygon line scanner: ray bundles at focal plane for a variety of scan angles and target distances. 15 plots in total. Each plot represents 5 mm × 5 mm on the focal plane.
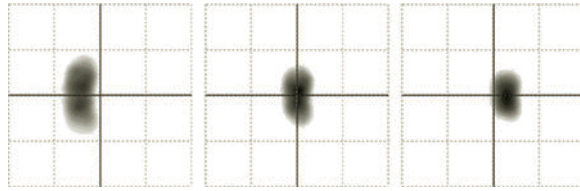
FIGURE 16.8
Composite overlaid images showing deviation of collected rays at focal plane over a 70° line scan for dual pyramidal polygon scanner (left: stand-off distance 5.2 m; center: stand-off distance 7.2 m; right: stand-off distance 9.2 m). Each plot represents 5 mm × 5 mm on the focal plane.

axes of the plots in Figures 16.7 and 16.8). The design of the system field-stop aperture needs to accommodate these deviations.

16.4.2 Single Hexagonal Polygon Line Scanner

A model for the single hexagonal polygon line scanner was created which has a source–receiver separation of 25 cm to be consistent with the fabricated prototype. Again to improve ray collection efficiency the target was modeled as the inside surface of a cylinder having a radius of curvature equivalent to the target distance (in this case 7.2 m) centered at the polygon center of rotation. For the cases at 5.2 m and 9.2 m the target curvature was changed accordingly. The collected ray bundles were aligned at the center of the focal plane disk for the 7.2-m stand-off distance case, and for the results in Figures 16.10 and 16.11, the steering mirror pair were aligned symmetrically at a focal distance of 7.2 m. The model is shown in Figure 16.9 for nadir case, as well as +35°. It can be seen that an additional aperture was used prior to the collection lens to limit the collection angle for the purpose of this study. In the implemented system this function is provided by the pinhole aperture residing at the focal plane.

A summary of the raytrace results for the single hexagonal polygon LLS is shown in Figure 16.10.

It can be seen from Figure 16.10 and the composite overlaid image of Figure 16.11 that the collected target irradiance walks minimally across the face of the focal plane disk due to scan angle variations (cross-axis walk-off of <0.5 mm, shown in the vertical axes of the plots in Figures 16.10 and 16.11) and the changes in target stand-off distance (along-track walk-off >1 mm between 5.2-m and 9.2-m cases, shown in the horizontal axes of the plots in Figures 16.10 and 11).

However, for the 5.2-m and 9.2-m target distance cases, the offset in the received irradiance can be nulled by refocusing the steering mirror pair at the focal distance. Figure 16.12 shows the effect of making steering mirror adjustments for the 5.2-m and 9.2-m cases; it is clear that the along-track walk-off has virtually been eliminated.

16.4.3 Discussion

Raytrace analysis of the two alternative line scanner designs highlights differences in aperture field-stop design to meet the optical performance requirements. More specifically, the results from the dual pyramidal polygon model indicate several second order requirements to achieve the desired level of performance. These include the need for a synchronous positioning of the aperture that defines the acceptance angle δ. This can be
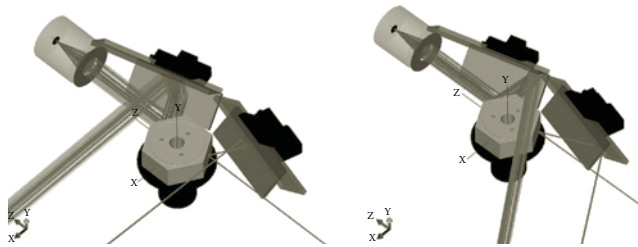
FIGURE 16.9
Single hexagonal line scanner modeled with back end optics showing a ray trace scan angle of 0° (nadir) and +35°.
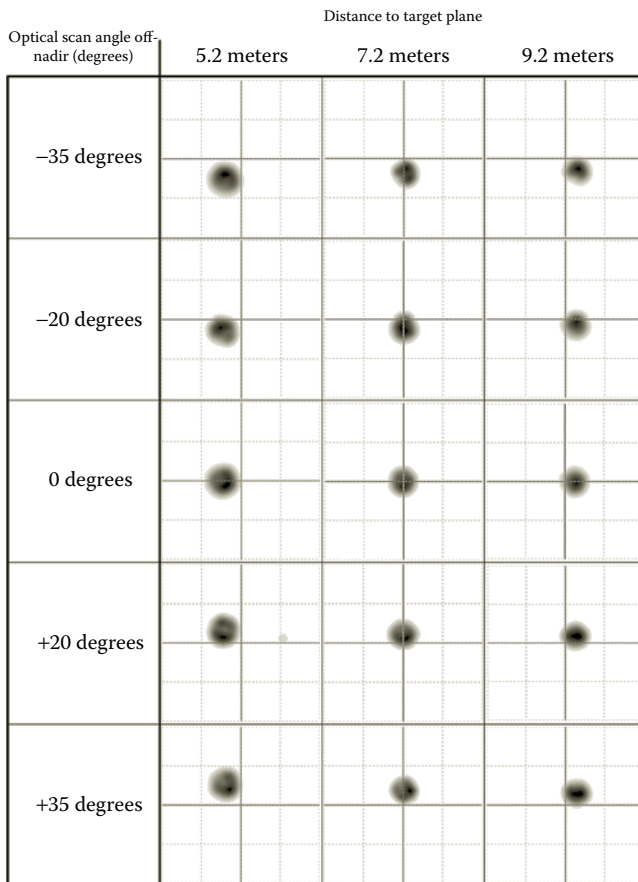
FIGURE 16.10
Single hexagonal polygon line scanner: focused ray bundles at focal plane for a variety of scan angles and target distances. 15 plots in total. Each plot represents 5 mm × 5 mm on the focal plane.

seen in Figures 16.7 and 16.8, where the focused ray bundle can be seen to deviate as a function of scan angle. The implemented system mitigates the cross-track deviation by utilizing a spinning aperture that is synchronized with the polygon scan motor to maintain a narrow cross-track IFOV through the entire scan line. To accommodate the along-track walk-off due to changes in stand-off distance, the implemented system also includes a
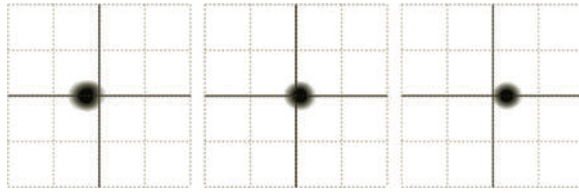
FIGURE 16.11
Composite overlaid images showing deviation of collected rays at focal plane over a 70° line scan for single hexagonal polygon scanner without refocusing adjustment to symmetrical steering mirrors (left: stand-off distance 5.2 m; center: stand-off distance 7.2 m; right: stand-off distance 9.2 m). Each plot represents 5 mm × 5 mm on the focal plane.
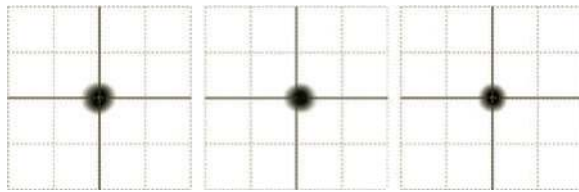
FIGURE 16.12
Composite overlaid images showing deviation of collected rays at focal plane over a 70° line scan for single hexagonal polygon scanner with refocusing adjustment to symmetrical steering mirrors (left: stand-off distance 5.2 m; center: stand-off distance 7.2 m; right: stand-off distance 9.2 m).

fixed aperture which can be set with knowledge of the desired upper imaging range and lower imaging range. A short distance behind this aperture assembly resides the photodetector, which is usually a PMT. The PMT needs to have a photosensitive region large enough to accommodate the two axes of walk-off in the collected irradiance profile, whilst being illuminated by an adequately large spot to avoid space-charge limitations. However, having a profile suited to a cylindrical housing, this class of LLS has major advantages in packaging for at-sea use. It has been used extensively for more than 2 decades and is the only known underwater synchronous scanning imager in operational use.

The results from the single hexagonal polygon model with adjustable steering mirrors have some advantages in reducing the complexity of aperture necessary to meet optical performance requirements. By means of both transmit and receive paths reflecting from a single polygon in alignment with the symmetrical steering mirrors, it provides parallel beam entry and exit axes at a selectable stand-off distance. This permits the use of a static pinhole aperture at the focal plane and allows for a smaller active area PMT placed behind the aperture. The purpose of the system was to experimentally investigate LLS performance trade-offs and to experimentally validate radiative transfer performance prediction models for the LLS class of underwater imager in highly controlled conditions at the Harbor Branch laser test facility. Ongoing use of this scanner involves bench testing future generation LLS system architectures that use pulsed and modulated-pulsed sources with gated PMTs. In general, high-speed PMTs needed for these studies have small photocathodes (<8 mm) so the receive path characteristics make the scanner suitable for such studies.

The next section summarizes some sets of test tank LLS imaging results performed with the single hexagonal polygon scanner to experimentally investigate performance trade-offs in both the choice of system geometrical parameters, and the use of alternate illumination and detection schemes. A description of these experiments, including test hardware details and image quality analysis results are given in Dalgleish et al. (2009).[2]

16.5 TEST TANK EXPERIMENTAL RESULTS USING SINGLE HEXAGONAL POLYGON LINE SCANNER

Experiments were conducted using the benchtop hexagonal polygon line scanner in the Harbor Branch laser imaging test facility at realistic stand-off distances in a variety of turbidity conditions ranging from very clear conditions to greater than seven attenuation lengths. Data were collected at 7-m stand-off distance (Z) for all imaging configurations presented here. The turbidity of the water in the tank was adjusted by adding a mixture of 50% laboratory grade magnesium hydroxide particles and 50% laboratory grade aluminum hydroxide particles. The resultant beam attenuation coefficient (c) and absorption coefficient (a) at 532 nm were measured with a Wetlabs ac-9 transmissometer. To allow for two-dimensional image formation, along-track motion between scanner and target was generated by mounting the target (1.2 m by 1.0 m dimensions) on a large rotating underwater drum (4 m circumference, 2 m length). Image data was collected with the line scanner configured with a 532-nm diode pumped solid state CW laser (maximum power 3 W) and continuous gain PMT. Using an adjustable pinhole iris as the field-stop aperture, the receiver IFOV was adjusted for each turbidity to vary both the common volume scattering volume and DOF. Image results from these tests, showing the effect of increased IFOV on image contrast are shown in Figure 16.13.

For the next sequence of images presented (shown in Figure 16.14), the line scanner was configured with a pulsed laser source and gated PMT. The laser source was a custom high repetition rate green (532 nm) pulsed laser was developed by Q-Peak (Bedford, MA). This solid state amplified master oscillator Q-switched YAG laser produced 7-ns FWHM green pulses at a fixed pulse repetition rate of 357 KHz. The average power entering the water was 1.3 W (4.6-μJ average pulse energy) with pulse-to-pulse energy instability of up to 40%. The pulse-to-pulse timing jitter was typically 10–20 ns. A small portion of the output was sampled by a reference detector and was used as a pulse monitor, both to trigger the receiver gate electronics and for normalization of pulse-to-pulse energy variations in acquired images. Correct alignment was verified, and the turbidity cycling and image acquisition sequence was repeated. System parameters (equal energy per pixel entering the water and far field beam divergence of 2 mrad for each laser) were adjusted to allow for a fair comparison with the CW LLS system images.

Despite obvious laser noise in the pulsed-gated LLS image set, the results in Figure 16.14 demonstrate that a significant contrast improvement is possible with pulsed-gated LLS. This is due to the ability of the gated-PMT to temporally reject the first 40 ns of backscatter from the emitted pulses, which contains no target information. The imaging performance of the CW LLS was limited mainly by multiple backscatter and shot noise generated in the receiver, eventually reaching a complete contrast limit beyond six beam attenuation lengths. The pulsed-gated LLS did not become limited due to multiple backscatter, instead reaching a limit (believed to be primarily due to forward-scattered light overcoming the attenuated direct target signal) beyond seven beam attenuation lengths. This result demonstrates the potential for a greater operational limit as compared to the existing CW LLS configurations. Furthermore, the pulsed-gated LLS does not require a significant source–receiver separation to produce quality images, and hence allows for more compact, simpler optical designs which have the potential to be more immune to changes in operating conditions, and hence more reliable than existing systems. Future advances in pulsed laser source technology should allow for these imagers to be suitably small for modern unmanned underwater platforms such as the man-portable AUVs.
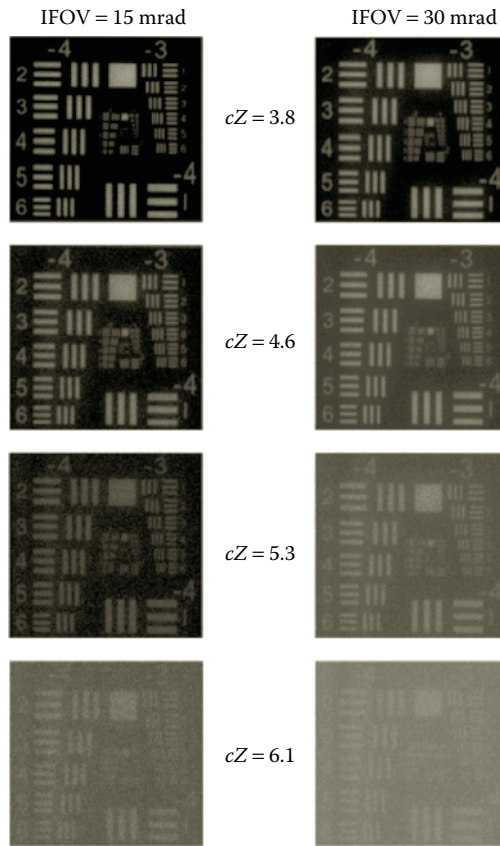
FIGURE 16.13
USAF-1951 white-on-black image sequences collected with CW LLS. Receiver angular aperture (in milliradians) and cZ, the number of attenuation lengths is also shown. Images show significant contrast reduction due to increased receiver angular aperture.

16.6 CONCLUSIONS AND FUTURE POSSIBILITIES

The line scanner designs discussed and analyzed in this chapter were both driven by the main optomechanical requirements for wide-swath extended range underwater viewing systems which are operable at variable stand-off distances. The ability to maintain synchronous tracking between laser beam and detector IFOV at the target plane through the entire scan swath of 70° allows a significant reduction in the common scattering volume between laser and receiver, thus spatially rejecting both forward and back-scattered light. To increase collection of light from the target region, a large collector area is required; each described design typically uses a 20-cm^2 collector area, defined by the area of a polygon facet.

The raytrace analysis of the two line scanner designs highlights differences in the required aperture field stop and consequently the minimum photosensitive region of
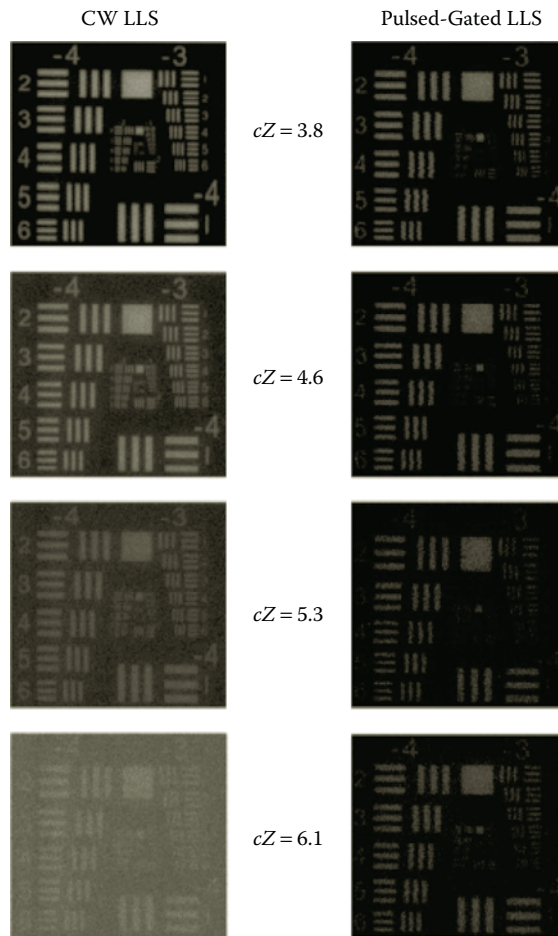
FIGURE 16.14
USAF-1951 white-on-black image sequences collected with CW LLS (left column) and pulsed-gated LLS using a 40-ns gate delay (right column). Receiver angular aperture was 15 mrad for both sets of images and cZ, the number of attenuation lengths is shown.

detector needed to meet the optical performance requirements. Perhaps the greatest difference between the two systems relates to dealing with variations in stand-off distance, and the implications at the focal plane where the field-stop aperture resides. The dual pyramidal polygon scanner, a mature system that has been operational for almost 2 decades, accommodates changes in stand-off distance at the focal plane via an adjustable field stop. Exhibiting significant along-track scan deviation, it consequently requires a large photosensitive region for detection. The hexagonal polygon scanner, a prototype system for benchtop experimentation of LLS alternatives, accommodates changes in stand-off distance via the use of symmetrical transmit and receive steering mirrors to direct the collected irradiance with minimal deviation at the focal plane. This results in a system that is compatible with a simple pinhole field stop and small photosensitive region for detection. Results in Section 16.5 describe the use of this benchtop scanner to demonstrate the achievable performance improvement of a pulsed-gated configuration of LLS over the existing CW LLS.

The design of the hexagonal polygon scanner was also motivated by the need to use a high bandwidth photodetector, which generally has small photosensitive regions. This is required for detection of pulsed or CW laser sources that use amplitude modulation, coupled with gating and coherent processing on the recovered waveforms. Such schemes, which are believed to offer measurable performance improvements over systems which use CW laser sources and possibly also over pulsed-gated LLS systems as described herein, have the additional potential for enhanced capabilities such as 3-D imaging and multiple-platform imaging.

Future LLSs for undersea imaging applications need to be even more compact to be compatible with current and future classes of man-portable AUV. The use of pulsed or modulated-pulse laser sources together with gated receiver modules can allow for more compact configuration of the described systems. These future systems may also make use of alternate laser scanning technologies, such as the optical microelectromechanical systems (MEMS) class of devices, which in recent years have been emerging in more diverse applications. These systems may be configurable without the need for precise optical path synchronization, instead using more flexible digital control and synchronization techniques and algorithms.

REFERENCES

1. Kulp, T.J.; Garvis, D.; Kennedy, R.; Salmon, T.; Cooper, K. Results of the final tank test of the LLNL/NAVSEA Synchronous-Scanning Underwater Laser Imaging System. *Proc. Ocean Optics* XI, 1750, 1992, 453–464.
2. Dalgleish, F.R.; Caimi, F.M.; Britton W.B.; Andren C.F. Improved LLS imaging performance in scattering-dominant waters. *Proc. of SPIE* 7317, 2009.
3. Strand, M.P. Underwater electro-optical system for mine identification. *Proc. SPIE* 2496, 1995, 487–497.
4. McLean, E.A.; Burris, H.R.; Strand, M.P. Short-pulse range-gated optical imaging in turbid water. *Appl. Opt.* 34, 1995, 4343.
5. Swartz, B.A. Diver and ROV Deployable Laser Range Gated Underwater Imaging Systems. Underwater Intervention '93 Conference Proceedings, New Orleans, Marine Technology Society and Association of Diving Contractors, 1993.
6. Witherspoon, N.H.; Holloway. J.H. Feasibility testing of a range-gated laser-illuminated underwater imaging system, *Proc. SPIE Int. Soc. Opt. Eng.* 1302, 1990, 414.
7. Fournier, G.R.; Bonnier, D.; Forand, J.; Luc and Pace, P.W. Range-gated underwater laser imaging system. *Opt. Eng.* 32, 1993, 2185.
8. Klepsvik, J.O.; Bjarnar, M.L. Laser-Radar Technology for Underwater Inspection, Mapping. Sea Technology, 1996; 49–52.
9. Funk, C.J.; Lemaire, I.P.; Sutton, J.L.; Marrone, F.A. Apparatus for scanning an underwater laser. US Patent 3,775,735, November 27, 1973.
10. Leatham, J.; Coles, B.W. Use of Laser Sensors for Search and Survey. Underwater Intervention '93 Conference Proceedings New Orleans. Marine Technology Society and Association of Diving Contractors, 1993.
11. Wells, W.; Hodara, H.; Wilson, O. Long Range Vision in Sea Water. Final Report ARPA order 1737, Tetra Tech. Inc., Pasadena, CA, 1972.
12. Gordon, A. Turbid test results of the SM2000 laser line scan system and low light level underwater camera tests. Underwater Intervention '94: Man and Machine Underwater, Conference Proceedings, Marine Technology Society, 305–311, Washington DC, 1994.

13. Carey, D.A.; Fredette, T.J. Use of Laser Line Scan System (LLSS) to locate and assess hazardous waste containers and geological features in Massachusetts Bay. GSA Abstracts with Programs, 1993; 128.

14. Carey, D.A.; Rhoads, D.C.; Hecker, B. Use of Laser Line Scan for assessment of response of benthic habitats and demersal fish to seafloor disturbance. In Special Issue, *Benthic Dynamics: In Situ Surveillance of the Sediment-Water Interface;* Solan, M, Germano, J.D., Raffaelli, D.G., Warwick, R.M., Eds. *Journal of Experimental Marine Biology and Ecology* 285–286, 2003, 435–452.

15. Taylor, J.S.; Hulgan, M.C. Electro-Optic Identification Research Program. *Proc. MTS/IEEE Oceans* '02, 2, 2002, 994–1002.

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

Index

A

Aberration, 47, 102–106, 113–114, 258, 507–508, 509, 513
  first-order chromatic aberration, correction of, 87–88
  third-order aberration, properties of, 88–91
  wave, 681–686
    allowable, 686
    defocus, 685–686
    from disk substrate, 682–683
    of optical components, 683
    due to semiconductor laser, 683–685
Absorption, 731, 732
Acoustic
  array, 568
  beam, 531, 546, 548
  cell, 545
  power density, 534
  power flow, 534, 535
  propagation, 530
  properties of bond layer materials, 577
  transducer, 553, 560–573, 583
  traveling wave lens, 551–553
  wave, 527–529, 536, 540, 542, 544, 582
Acousto-optic
  deflectors, 95
    design procedure, 548–549
  device fabrication
    cell fabrication, 573–574
    packaging, 577–578
    transducer bonding, 574–577
  devices, materials for
    general considerations, 555–556
    theoretical guidelines, 556–558
  interactions
    anisotropic diffraction, 536–543
    isotropic AO interaction, 528–536
    photoelastic effect, 527–528
  modulator, 276, 277, 721, 726
    design procedure, 549–551
    interaction bandwidth, 545–548
    multichannel, 554–555
    resolution and bandwidth considerations, 543–545
  scanner
    acoustic traveling wave lens, 551

    applications of, 578–591
    chirp lens, 553–554
    design considerations, 551–553
    materials for, 558–560
Across-scan, *see* Cross scan
Actuator, 442, 476, 688–689, 704–706
  electrostrictive, 482
  piezoelectric, 475, 638, 654–658
  rotational, 689
  stacked-type, 638, 642, 643, 653
  wire-suspended, 689
Additive
  color system, 160
  density, *see* Density, additive; Exposure, additive
Adiabatic driver, 625–627
Aerodynamic
  bearing, *see* Air bearing
  bearing comparison of, 322
  scanners, 324, 340, 341
Aerostatic
  bearing, *see* Air bearing
  bearing comparison of, 322
Air bearing, 264, 326–338
  comparison of, 322
  journal bearing, 327–330
  lobed bearing/shaft, 338–340
  scanning construction, 333–334
  spindle construction, 340–341
  spiral groove, 337–338
  thrust bearing, 330–333
Air lubricated bearing, 321
Airy disk
  distribution, 76, 584
  pattern, *see* Airy distribution
Airy distribution, 680–681
Aliased image, 152
Allowable wave aberration, 686
Along-array numerical aperture, *see* Numerical aperture, along-array
Along-scan, *see* In-scan
Amorphous, 557, 559, 560, 675
Amplifier, 628, 648, 649
  driver, 439–440
Amplitude distribution, 672, 680, 707
Analog drivers, 627–628