INTRODUCTION TO THE
Optical Transfer Function

INTRODUCTION TO THE

Optical Transfer Function

Charles S. Williams
Orville A. Becklund

SPIE PRESS
A Publication of SPIE—The International Society for Optical Engineering
Bellingham, Washington USA

Library of Congress Cataloging-in-Publication Data

Williams, Charles S. (Charles Sumner), 1911-
Introduction to the optical transfer function / by Charles S. Williams, Orville A. Becklund.
p. cm.-- (SPIE Press monograph ; PM112)
Originally published: New York : Wiley, c1989, in series: Wiley series in pure and applied optics.
"A publication of SPIE–the International Society for Optical Engineering."
Includes bibliographical references and index.
ISBN 0-8194-4336-0
1. Optical transfer function. 2. Optical measurements. I. Becklund, Orville A., 1915-
II. Title. III. Series.

QC367 .W55 2002
535–dc21 2002075891
CIP

Published by

SPIE—The International Society for Optical Engineering
P.O. Box 10
Bellingham, Washington 98227-0010 USA
Phone: 360.676.3290
Fax: 360.647.1445
http:// www.spie.org

Printed in the United States of America.

PDF ISBN: 9781510607828

Cover images: A representation of the two-dimensional, perfect OTF (Chapter 5, Ref. 6), and a plot of intensity distribution in the image of a sinusoidal test object showing (1) an ideal image, (2) a "perfect" image having reduced amplitude, and (3) an aberrated image having a further reduced amplitude and a phase shift (page 157, Fig. 5.7).

Dedication

We dedicate a second publication of this book to the memory of spouses—to Orville, deceased co-author and husband of Mrs. Alma Becklund; and to Dorothy, wife of co-author Charles Williams.

Orville was proud of the finished book; he was always pleased to say, "It took us 13 years to write that book."

Dorothy was too deeply ill with Parkinson's disease to appreciate whatever she did see of the printed book. Even so, let no one underrate the need for a spouse to the success of a book.—CSW, ABB

Contents

Preface to the Reprinted Edition

When Orville Becklund and I began writing our book, a powerful and rapid computer was not available to us. The best we had was a hand-held programmable calculator. We used it to calculate solutions in which each solution consisted of a series of series. Many times I had to program the calculator to run all night. I would turn it on to calculate until morning, and go to bed. Finally, I had a bunch of partial answers to put together. I think it should have been expected that we would always be uneasy about the accuracy of data that finally found its way into the text.

Computers, programs, and programmers have come a long way since then. One of the best for calculating problems relating to optics, is from the work of Dr. David F. Edwards of Tracy, California. Much of his work in optics programming was after his retirement from Lawrence Livermore National Laboratory as head of the Optical Sciences and Engineering Group. Our calculations are updated by Dr. Edwards's, and are found in Appendix D (p. 401).

Charles Williams
July 2002

Preface

An abundance of knowledge about the optical transfer function (OTF) has been published in many excellent articles during the past 35 years or so, but somehow a niche for this knowledge has never been found in the engineering and scientific structure. As a result, OTF publications are scattered throughout the archival literature of scientific and technical journals. Our book aims to bring together into one source much of this wealth of information.

Those concerned with grounding engineers and scientists in the procedures of optical evaluation have found that *spatial frequency*, *wave-front distortion*, and *optical transfer function*, though not particularly difficult concepts to understand, do not as easily become part of one's thinking, and therefore practice, as the concepts of *rays*, *ray tracing*, and *ray aberrations*. The word *ray* (geometrical optics), for example, in contrast with *spatial frequency* (physical optics) is used so commonly in our language that it is no longer an esoteric term reserved for optics. Actually, there are advantages peculiar to each of the two viewpoints, and an optical analyst is handicapped by a lack of facility with either. We hope that our book is articulate enough in the art to bring practitioners up to speed in the realm of spatial frequency and the OTF.

Specifically, our text dwells on such fundamental concepts as spatial frequency, spread function, wave aberration, and transfer function—how these are related in an optical system, how they are measured and calculated, and how they may be useful. In the early chapters we review the historical background for the OTF, the related concepts, and the necessary nomenclature and coordinate systems. We discuss in some detail the wave aberration function, which is a measure of an optical system's ability to produce an image that is a ''reasonable facsimile'' of the object and which, therefore, is a fundamental characterization of the system's excellence of performance. We derive the optical transfer function and related concepts mathematically, and we discuss some ways that the OTF can be used for assessing the quality of an optical system both during its design and during testing of the manufactured system.

We show how the OTF can be used: when specifications for the optical system are being drawn up, when the OTF is part of a merit function while the system is being designed by computer, and when the optical system is being tested to verify adherence to specifications. Finally, we show how the OTF can be calculated mathematically, both by analytical procedures and by numerical methods of integration.

In the appendixes some pertinent mathematical basics are reviewed, and we document a number of OTF calculations that other workers have made.

Our book makes liberal use of illustrations. For the reader who wishes to pursue studies beyond the scope of our text, we provide a full complement of references at the end of each chapter.

The reader of our mathematical chapters should have had courses in calculus; a course in transform theory would be helpful but not necessary because the mathematics in the appendixes provide a review of all the Fourier transform theory that the reader will need. Besides the professional nonexpert in physical optics, the level of our text is intended to suit undergraduates with limited exposure to optics, such as juniors and seniors in science, mathematics, and engineering.

We have purposely avoided certain OTF topics: We do not treat the geometrical approximation of the OTF, the OTF of sampled images, or the polychromatic OTF, because we feel that the state of the art concerning each of these topics is not quite ready to be included in a tutorial book on the optical transfer function.

We make no pretense that the ideas in this book are original with us. Our information has come through various paths and from many sources, and we have tried to give credit at the appropriate places in the text to the many whose work we have used.

Charles S. Williams
Orville A. Becklund

Dallas, Texas
May 1988

1

OTF Historical Background

INTRODUCTION

The Optical Transfer Function (OTF) is the frequency response, in terms of spatial frequency, of an optical system to sinusoidal distributions of light intensity in the object plane; the OTF is the amplitude and phase in the image relative to the amplitude and phase in the object as a function of frequency, when the system is assumed to respond linearly and to be space invariant. The OTF depends on and potentially describes the effect of diffraction by the aperture stop and the effects of the various aberrations.

The concepts related to the OTF, which are considered in some detail in the next chapter, evolved very slowly. In fact, our whole civilization developed so gradually that only rarely can we clearly mark the beginning or the end of a stage in the evolutionary process. Similarly, any historical stage through which our modern institutions evolved was so like the preceding and the succeeding stages that a date of birth can hardly be established. We think it of some importance, therefore, that in the world of optical design and evaluation a new era has emerged in which once familiar terms like *circle of least confusion*, *blur circle*, *resolution*, and *bar chart* have become obsolete and, instead, the optical transfer function is being accepted as a criterion for the performance of optical systems.

An analysis of any given optical system using the OTF must necessarily consider the shape of wave fronts at the exit pupil; that is, it must use wave optics rather than, or at least as well as, ray optics. The new emphasis on wave optics has undoubtedly been a handicap in the advancement of the OTF; there is evidence that a traditional dependence on geometrical optics and ray tracing for design and analysis has delayed the acceptance of the OTF. The people who made their living and built their reputations by tracing rays may have felt themselves too busy to explore the possibilities of the OTF. ''Prosperity, like the other creations of technology, is a tiger whose riders do not dare to dismount'' [1]. Among the senior practitioners of optics there has been a tendency to regard the OTF as interesting but too theoretical to be of much practical use for deciding either how to design an optical instrument or how to evaluate one. Perhaps the complementary new approach has somehow been seen as a threat to the old ways, and, as Robert Frost suggests in "Reluctance,"

Ah, when to the heart of man
Was it ever less than a treason
To go with the drift of things,
To yield with a grace to reason,
And bow and accept the end
Of a love or a season.

Nevertheless, substantial progress in the knowledge of the OTF and its potential usefulness seems to have begun during the mid 1950s; and the art of OTF has steadily continued to advance since that time so that now the OTF can be applied to the procedures of optical system design, specification, and evaluation. Therefore, this book must give special consideration to these topics along with the main topic, *the OTF*. The OTF is now recognized as a means of refining an optical system during the final stages of design optimization; its application has the potential of going beyond the optimum design that can be obtained with ray optics alone. In the sections that follow, a brief history is given of early optical design and evaluation; and the history of OTF concepts and of the OTF potential is outlined.

For a list of the principal contributors to OTF concepts, we offer the authors in the chapter reference lists of this book. Prominent among these contributors are H. H. Hopkins and his associates, including his pupils, at the Imperial College of Science and Technology of London. For guidance in assembling the brief historical background of OTF for this chapter, we are particularly indebted to Hopkins, Baker, and Smith [2–4].

THE EARLY HISTORY OF OPTICAL DESIGN AND IMAGE EVALUATION

The theory of optical instruments and the evaluation of their performance have been studied ever since the first useful systems were assembled early in the seventeenth century. Long before that time, Pliny and other ancient writers indicated that people knew about burning glasses, which were glass spheres filled with water. However, it was not until the thirteenth century that any mention was made of lenses deliberately made for a purpose, for example, spectacles. In about 1608, a Holland spectacle maker, Hans Lippershey, is said to have been holding a spectacle lens in each hand, and when he happened to align them before his eye with the steeple of a nearby church, he was astonished to find that the weathercock appeared nearer. Then when he fitted the two lenses in a tube to maintain their spacing, he had constructed the first telescope.

Galileo Galilei in Venice heard about the new telescopes in June of 1609, and immediately began to make telescopes of his own. His first had a magnification of 3, but his instruments were rapidly improved until he achieved magnifications of about 32. He manufactured hundreds with his own hands for observers and experimenters throughout Europe. Galileo's own application of the telescope turned to the heavens. He startled the world with observations of Jupiter's satellites, the spots on the sun, the phases of Venus, the hills and valleys on the moon, and the nature of the Milky Way.

Spherical aberration was soon recognized by the developers of the early telescopes as a reason for defective images, and considerable effort was spent in experimenting with various aspherical refracting elements to overcome this fault. In 1666 Isaac Newton discovered that refraction by a given lens depended upon color, and he correctly concluded that the most significant defect of the then current telescopes was what we now know as chromatic aberration. He hastily concluded that all glasses had the same relation between refraction and color, so he turned to reflectors to solve the color problem. This decision prevailed in the telescope art for almost seventy years. Then Chester Moore Hall, a gentleman from Essex, realized that the refracting humors of the human eye are balanced optically to avoid color separation in visual images. Why shouldn't an appropriate combination of optical glasses solve this problem in telescopes? In 1733 he demonstrated a refracting telescope essentially free of chromatic aberration.

Long focal lengths and parabolic mirrors were also early means for alleviating aberrations in the telescope art.

Inasmuch as the first optical systems were mostly designed for astronomical work, it is not surprising that the star (an almost perfect point source) became the standard test object. Although many other objects have since been used for testing optical systems, the study of star and other point-source images has persisted to the present time in evaluating the response of image-forming systems. As later chapters in this book indicate, the OTF represents the latest organized approach to judging systems by the nature of their point-source images.

Even systems that are almost free of aberrations produce highly complex star images. A skilled observer, particularly one who specializes in one kind of optical system, still finds the star test one of the most sensitive methods of evaluating aberration residuals; however, the image is extremely difficult to interpret quantitatively. In general, the star test provides so much data in its own peculiar code that considerable data reduction is required before a particular star image can be said to qualify the system for, say, a television or an image intensifier application. The precision required in data reduction can be appreciated when we realize that a star image rarely extends over more than one or two hundred micrometers in diameter and that the range of flux densities to be measured and recorded can extend over a few orders of magnitude. If we choose to apply the star test by calculating reference images for a sequence of design parameters for an assumed system, we find that calculation of the distribution of flux density in the star image is still a formidable problem.

As applications broadened from the study of sky objects to predominantly *extended* terrestrial objects, the difficulties of evaluating optical systems by interpreting star images led to the use of various extended test objects. In recent decades, a variety of test charts consisting of black and white bars of varying dimensions has been a favorite type. The bar chart has one notable advantage: Performance can be specified by giving a single number, which is the acceptable bar spacing (either the bar interval or the number of bars, or lines, per unit distance perpendicular to the bars) in the image of the chart. However, the method has a number of shortcomings. Results are highly dependent on the nature of the image detector (human eye, photographic emulsion, etc.). Near the resolution limit, the boundary between bars of the bar chart image becomes a gradual transition from maximum to minimum intensity (or reflectance) rather than a sharply defined boundary between black and white; so repeatable observations of the resolution limit for a given system are difficult. A phenomenon called *spurious resolution*, in which the color of bars is reversed, that is, a black bar becomes a white bar and a white becomes black, is often seen in bar charts. It is also hard to predict the results expected from optical design data or to describe the quality of reproduction of another kind of test object once the bar chart resolution is known. In contrast to star images, which seem to give too much information, bar chart images tend to tell too little.

As he reflected on the limitations of the bar chart in 1938, Frieser [5] suggested that the resolution test could be improved by substituting a one-dimensional sinusoidal variation of grays in the test object for the abrupt variation from black to white characteristic of the bar pattern. He saw an advantage resulting from the nature of optical systems to produce a sinusoidally distributed image from a sinusoidally distributed object over a wide range of spatial frequencies. This concept brought with it the first expression of a related idea, that the transfer function connecting the sinusoidal image with the sinusoidal object would be a good way to assess the performance of optical systems.

Since the beginning of optical system fabrication, designers have relied on the accepted concept of straight-line propagation of light with a change of direction (refraction) at boundaries between unlike media. Directed lines called *rays* depict light paths from object to image. Sometimes *pencils* of rays are traced to evaluate discrepancies between the configurations of object and image. Perhaps less frequently, ray density has represented light flux density, which allowed a comparison of image distribution of light flux with the corresponding object distribution. For the most part, the art of optical design has depended on simple algebra and plane trigonometry, generally referred to as *geometrical optics*.

Even modern computer three-dimensional ray-tracing programs, for all their speed, precision, and self-modification toward optimum design, are fundamentally the same process as the paper-and-pencil, cut-and-try procedures used for a couple of centuries.

About a hundred years ago, a study of intersection patterns on the image plane by pencils of rays from the object led to an analytical theory of aberrations. The image plane was usually defined as the plane of best focus, that is, the plane on which a point source would produce the minimum *circle of confusion*.

Geometrical optics, though recognized as only approximate almost from its inception, has remained the mainstay of optical design simply because it has produced excellent results. In fact, application of geometrical optics to the elimination of aberrations finally produced such fine systems that actual star images, in their departure from a simple point configuration, could no longer be explained by geometrical optics. In 1835, Airy [6], who was familiar with the wave theory of light, developed the formula for the diffraction pattern, thereafter known as the *Airy disk*, that constituted the image of a point source in an aberration-free optical system.

It would have seemed reasonable if the optical designers, impressed by Airy's accounting for the performance of their ultimate systems, had turned to wave theory for further advancement of their art. In the main, however, they did not. Their reluctance to incorporate the wave theory of light, included in what is known as *physical optics*, into their calculations could not be attributed to a lack of scholars and well-documented work in the field. Among the leaders were Grimaldi (1665), Huygens (1678), Young (1802), and Fresnel (1815). In fact, in 1818 Fresnel, by using Huygens' concept of secondary wavelets and Young's explanation of interference, developed the diffraction theory of scalar waves essentially in the form that we know it today. Nor could the resistance to wave theory by designers result from the adequacy of geometrical optics to handle *all* design problems. Besides the particular need satisfied by the Airy disk, general geometrical optics treatment of optical processes in the vicinity of the focus and near the edge of a beam in limited-aperture systems is particularly lacking.

The dichotomy of geometrical optics and physical optics has persisted to the present day. A frequent question for graduate students preparing to extend their optical background is, "physical or geometrical?" It is interesting to speculate on this division. The solution of diffraction problems associated with nonspherical waves, which are characteristic of nonideal optical systems, is a difficult mathematical problem. This fact, in combination with the scarcity of corresponding reliable, flexible test equipment, may have delayed the development of OTF techniques. Further, many of the best lens designers in the past seem to have had only sporadic training in physics and mathematics; on the other hand, the scholars of physics and mathematics have been quite unproductive of practical designs. Except for a curious glance over the fence now and then, each optical expert seems to have been diligently working in his or her own field.

LAYING THE FOUNDATION FOR OTF—1850 to 1940

Although the bulk of practical optical work until recently has been accomplished with little reference to physical optics, certain concepts, which seemed alien to each other but are fundamental to the optical transfer function, emerged early in the nineteenth century. By the late 1850s, Foucault was suggesting that a striped pattern as an extended object would be a better target than a point source for testing an optical system [7]. Rayleigh is said to have actually set up a periodic test object in 1896 [8]. Early in the 1870s, Abbe expressed the concept of spatial frequency, although he did not call it by that name, in his history of microscope imaging [9, 10]. This theory grew out of extensive experimentation that he performed to prove to his contemporaries that they were misinterpreting diffraction effects as resolved structures of the specimens they were observing under highly coherent light beams.

Parallel with the development of geometrical aberration theory (latter half of the nineteenth century) came a unification of geometrical rays with waves; this is expressed in the characteristic function of Hamilton [11; 12, pp. 200–202; 13, pp. 112, 133–142] and the eikonal of Bruns [14; 12, pp. 202–207; 13, pp. 112, 134–135]. Simultaneously Lommel [15; 13, pp. 435–438] solved the diffraction theory of light distributions produced in out-of-focus planes by aspherical wave fronts.

Also in the same period, Rayleigh formulated his well-known quarter-wavelength limit or criterion for optical quality [16]. This occurred while he was investigating the permissible inhomogeneities in large spectrograph prisms. Starting with the premise that the emerging image-forming wave front of a perfect optical system, having a point source object, is spherical, Rayleigh postulated that the image of an actual system would not be significantly different from that of a perfect system if the actual nonspherical wave front could be contained between two ideal spherical wave fronts no more than a quarter wavelength apart.

Another Rayleigh exercise that smacks of OTF is his analysis of the image of an equally spaced array of mutually incoherent point sources. He assumed an aberration-free system and expressed the intensity distribution in the image by a Fourier series. The resulting curve would now be described as the modulation transfer function, which is a component of the OTF.

Strehl, a contemporary of Rayleigh, has only recently been fully appreciated for showing how small aberrations modify the Airy disk [17]. In his book, *Die Beugungstheorie des Femrohrs*, he showed that the small aberrations reduce the intensity at the principal maximum of the diffraction pattern, that is, at the ''diffraction focus,'' and that the removed light is distributed to the outer parts of the pattern. As a measure of degradation of image quality, Strehl set up the ratio of intensity at the diffraction focus with and without aberrations and called it *Definitionshelligkeit*, which translates literally to *definition brightness*. The term without explanation is equally ambiguous in both languages. In English, it is sometimes called *the Strehl definition*; however, *the Strehl intensity ratio* or simply *the Strehl ratio* may be preferable.

The first lens designer to have come really to grips with combining physical optics and aberration theory to advance the art seems to have been Conrady [18]. Starting with the Rayleigh limit as the criterion for permissible wave-front errors, he first formulated theoretically and then applied in practice tolerances on focus errors and aberrations. He developed formulas by which he calculated optical path differences along various rays to arrive directly at the wave-front aberration for the optical system under investigation.

Both the Rayleigh criterion and the useful range of the Strehl ratio apply to highly corrected systems. The quarter-wavelength limit defined by Rayleigh corresponds to a Strehl ratio of about 0.8. Above this limit, A. Maréchal (1947) [19] showed that the loss of intensity at the diffraction focus is simply related to the root-mean-square departure, called the *variance*, of the wave front from a spherical shape. This relation was extended to simple formulas for finding the best plane of focus in the presence of aberrations. Maréchal also showed how best to compensate for higher order aberrations by the introduction of appropriate amounts of primary aberration. The variance of wave-front aberration in conjunction with a special *canonical* coordinate system is used in automatic optical design programs to estimate image quality.

When the Strehl ratio drops much below 0.8, the quality of the optical image deteriorates rapidly; in fact, the point-source diffraction image, especially off the optical axis, becomes so complex that it can no longer be practically analyzed to give a Strehl ratio or any other single-number figure of merit.

Relations in optical systems are generally nonlinear, so pioneering work had to be done to justify the linearity assumptions inherent in the optical transfer function approach. This requirement parallels the linearity assumptions that underlie most electrical system analyses. Work in this direction was done by Frieser [5] and Wright [20], who in the late 1930s suggested the use of an optical transfer function. Frieser applied the concept of a transfer function to photographic film and discussed the amplitude reduction of sine-wave targets by the film emulsion. Wright considered an optical lens as a component in a television network and suggested that the performance of the lens might be conveniently expressed in terms of the amplitude attenuation it introduced. Of course, he was thinking of frequency in time rather than space in his consideration of the whole network, but inherent in his treatment of the lens as a contributing component was the concept of spatial frequency.

Frieser's work with photographic films and their emulsions was connected with experimental sound recording on motion picture film using variable density in the sound track. (The competing technology in early motion pictures involved recording a variable opaque area in the sound track.) Ideally a pure note of a single pitch with no harmonics would record as a pure sinusoidal variation of transmittance along the sound track. To describe the properties of the recorded image, Frieser employed the notion of an optical transfer function.

THE APPEARANCE OF SOME IMPORTANT MATHEMATICS

During the 1930s and 1940s mathematicians were developing a branch of their discipline that is now known as transform theory. One kind of transform known as the *Fourier transform* became quite useful to electrical engineers who were working out the theories of sound recording and amplification. The Fourier transform, defined mathematically by an integral known as the Fourier integral, establishes a correspondence between a pair of mathematical functions; for example, one a function of time and the other a function of frequency. The functions, expressed in terms of either variable, could represent some physical quantity such as voltage or current. Operation of the integral can be made on either function; that is, the transform can go either way, from the time variable to the frequency variable or from the frequency function to the time function. Thus, there came into being the two concepts: a time domain and a frequency domain. It is easy to go from one to the other, and so the analyst can work in whichever domain is most fruitful and then go back to the other domain if necessary.

One particularly useful process in Fourier transform theory is an integral called the convolution integral. From it came the convolution theorem which states: If a function is given by the convolution of two other functions, the transform of the first function is given by the product of the Fourier transforms of the other two functions, when their Fourier transforms are taken individually. The convolution theorem is a powerful tool in frequency analysis.

The Fourier transform can easily be extended to two dimensions as over an area in the object plane, or image plane, of an optical system; and the pair of corresponding functions can be in terms of spatial coordinates or in terms of spatial frequency. Functions representing physical quantities, for example, light intensity over the object plane, easily meet the requirements for validity of the Fourier transform. Thus, we have a ready-made theory for treating functions of spatial frequency in the object and image planes of optical systems, provided we find a means for relating the frequency function in the image to the frequency function in the object; this means must be a function of the optical system alone. Such a ready-made function is the spread function, which is the distribution of light flux in the image when the source of light in the object is a single point source. It turns out that the image function in terms of spatial coordinates is the convolution of the object function and the spread function when the optical system is assumed to have linearity and to be space invariant. The Fourier transform of the spread function is one definition of the OTF.

And so, this book, reduced to its fundamental purpose, establishes, explains, illustrates, and extends the concepts of the previous three paragraphs. We have attempted in these paragraphs to describe concisely what the OTF is and how it can be treated, mostly by Fourier transform theory. The resulting formulas begin with the wave-front shape in the exit pupil or with a spread function: point spread function, line spread function, or the edge trace.

GROWING AWARENESS OF OTF—THE 1940s

By 1950, Schade, Selwyn, and Luneberg [21–23] were analyzing the image quality of the early television systems by Fourier methods. During World War II, Selwyn applied these methods separately to lens and film. Later, drawing upon his extensive experience in assessing images in terms of photographic resolving power, he showed that the light intensity variations in the image of a sinusoidal test object could be calculated from the observed variation of light intensity in the image of a line source.

In 1944, Luneberg published a detailed theoretical discussion about the resolution of objects having periodic structure.

Schade, an electrical engineer working on television lens evaluation, developed the concepts of contrast reduction through a sequence of cascaded system elements by applying communication theory. In a series of papers beginning in 1948, he discussed the calculation and evaluation of the electro-optical characteristics of imaging systems, and he introduced the transfer function of a lens and showed how it could be modified to increase image sharpness.

In a series of papers beginning in 1935, P. M. Duffieux first fully formulated the theory of the optical transfer function, including the role of the diffraction integral, for describing the image of a two-dimensional incoherent object. He was undoubtedly helped by progress in the techniques of mathematical analysis spurred by World War II research in acoustics and communication theory. His book, *L'Integral de Fourier et ses Applications a L'Optique* [24], in which he applied Fourier methods to optics, became a widely used practical source book, particularly for European workers in the field.

Duffieux showed that the Abbe theory of image formation of coherently illuminated objects may be usefully restated in terms of Fourier analysis. He also established the complementary fundamental theorem for the image formation of self-luminous, incoherent objects. His transmission factor is equivalent to the transfer function, and he showed how to compute it from the wave front in the exit pupil. He extended his theory of image formation to any form of aperture and aberration, making only the assumption of the Fourier transform relation between a diffraction aperture and the Fraunhofer diffraction pattern. However, for self-luminous, incoherent objects, the Fraunhofer diffraction formula assumption is an unnecessary restriction. By using the convolution theorem, Duffieux demonstrated that the Fourier transform of the function expressing the distribution of intensity in an image is closely approximated by the product of the Fourier transform of the distribution in the object by the transform of a point source image. This relation holds provided the point-source image (also known as the *spread function*) does not change significantly over the angle of field through which the Fourier components are transmitted to the image. The fractional contrast reduction of these components can be called the transfer function of the optical system; Duffieux found that this function depended on both the lens aperture and the aberrations.

Duffieux's new concept of image formation passed largely unnoticed by workers in the field of optics, many of whom were concentrating on making consistent measurements of resolving power that would correlate in some way with the predictions of lens designers.

The story of the 1940s would be incomplete without recognizing the advances made in analyzing diffraction effects caused by various types of wavefront distortion. For small aberrations where departure from a spherical wave front is only a fraction of a wavelength, Nijboer [25], partly in collaboration with Zernike and Nienhuis, showed the influence of small aberrations on the image in an extensive diffraction treatment. Nijboer was the first to use the concept of a reference sphere, and he also defined the *wave aberration function* as the difference between the wave surface and the reference sphere. The effects of large aberrations were studied by Van Kampen [26]. Finally, Hopkins' book, *Wave Theory of Aberrations* [27], defined the basic types of aberration in terms of the wave-front distortion. So the foundation was laid for the calculation of the OTF for specific types of aberration.

INVENTIVE OTF INSTRUMENTATION—THE 1950s

The ideas prevalent at the beginning of the 1950s are found in the published proceedings [29] of a symposium held at the National Bureau of Standards in

1951. An important paper by Elias and others [28] was published in 1952, but, as in Schade's work, the OTF ideas are in terms of input and output functions without any reference to physical optics. Another symposium, held in 1955 at the University of Rochester, resulted in a large collection of papers related to the OTF in the September 1956 issue of the *Journal of the Optical Society of America*. These publications document the growing awareness of the OTF as a new procedure in optical system analysis. During the 1950s, and especially during the last half of the decade, there seems to have been a mushrooming of concepts and instrument developments related to the OTF; these came from many sources, largely independent of each other. Along with the growth of OTF ideas there was a beginning of instrument construction and OTF measurement techniques. Apparently, Schade's work with Fourier analysis, and the acceptance of the term *transfer function*, inspired many engineers and scientists to invent a variety of new photoelectric instruments for measuring what is now defined as the OTF.

In contrast to the input–output approach of Schade, H. H. Hopkins based his OTF developments on physical optics to provide a comprehensive foundation for optical design and evaluation [27, 30, 31]. E. L. O'Neill assembled his work in a mathematical form particularly appealing to engineers and others familiar with communication theory as applied in electronics systems [32]. His summaries and the impact of his teaching were particularly significant contributions.

Rosenhauer and Rosenbruch published a review of OTF instruments [33], most of which were developed during the 1950s. This work reflected the enthusiastic equipment-building effort in most research activities interested in OTF. Although the resulting publications indicate that the equipment verified a few curves calculated from diffraction theory, the real interest of the participants seemed more in inventing and constructing fascinating equipment than in results that could be compared with theory or with results obtained by other instruments.

In the general wave of appreciation of OTF in the fifties, certain favorable features of this approach became better understood:

1. A sine wave (a sinusoidal variation of light intensity over the object) is imaged as a sine wave.
2. The OTF approach deals with an *extended* object.
3. Real-time testing of optical parts and systems is practical.
4. A wide range of systems, from those suffering from severe aberrations to diffraction-limited systems, can be tested. Also, different types of systems, including fiber optics, can generally be evaluated by OTF methods.
5. One-to-one comparisons can be made between measured OTF data and the corresponding data calculated from design information.

6. By resorting to Fourier analysis, one can predict the shape of the OTF curve required to produce an acceptable image of a given object.
7. The real-time and quantitative nature of OTF evaluation of optical parts and systems provides a means of specifying and quality control in optical manufacturing.

On the other hand, some of the limitations of OTF testing were also appreciated. For instance, the effect of scattered light or veiling glare can best be evaluated by methods not depending on the OTF.

A review of OTF equipment most promising at the end of the 1950s for real-time testing in industry indicates that they reduce to two basic types. Both types employ a narrow optical slit, a sinusoidally distributed pattern, and a photocell detector; but the two types differ in the sequence of these optical elements. In the first, the sinusoidal pattern, incoherently illuminated, is the object of the optics under test. Its image, which is also a sinusoidal pattern, is scanned by the slit, and the detector picks up the time-varying transmitted light signal, the amplitude and phase of which represents a point on the OTF function. In the second method, the slit is illuminated and becomes the object. The slit image is scanned by the sinusoidal pattern, now a transmitting screen, and the detector picks up the ''chopped'' signal. Application of Fourier transform theory shows the two equipment types to be equivalent.

These two theoretically direct techniques for measuring the OTF for an image-forming optical system were found to have certain practical difficulties:

1. Periodic patterns of constant contrast and truly sinusoidal distribution were extremely difficult to make.
2. Design limitations tended to restrict the range of spatial frequency to less than desired. Also, an equivalent test near zero frequency was difficult to attain.
3. Practical light signal levels at the detector were often low enough to be troubled by ambient light.
4. Illumination of the object, either the sinusoidal pattern or the slit, had to be truly incoherent. Even a small degree of coherence would produce a wrong result.

These and other difficulties usually required equipment considerably more elaborate than the basic principles might indicate.

By the end of the decade, there was a growing success at theoretically calculating the OTF for certain specific types of aberration. (See Appendix A.)

Since, as already mentioned, geometrical optics rather than physical optics was the field of the practical lens designers, one could expect attempts toward geometrical approximations for the transfer function. Indeed, Hopkins [34] and his colleagues, De [35], Bromilow [36], and Marathay [37], showed that good results are possible at low frequencies for aberrations greater than two wavelengths. Miyamoto [38] and Hopkins [34] each developed a general expression for a geometrical transfer function. This general method gave good results for very large aberrations, but the approximations become poor for small aberrations. In the intermediate range around one and two wavelengths of aberration, a mixed method of multiplying the geometrical transfer function by the transfer function of the aberration-free system was often found to give good results. However, even this approach had its shortcomings for certain levels of aberration [39, 40].

Ultimately the most exact method for computing the OTF is to calculate the self-convolution of the pupil function (defined in Chapter 5) or to calculate the diffraction integral of the pupil function over the exit pupil followed by calculation of the point spread function and, finally, the Fourier transform (discussed in detail in later chapters). Some of the results from this approach first appeared in the late 1950s and early 1960s in papers published by De [35], Black and Linfoot [41], Steel [42], Barakat [43], and Goodbody [45].

ADJUSTMENT TO PRACTICE—THE 1960s

Since the amplitude part of the OTF (called the *Modulation Transfer Function*) is based on measuring the contrast in the image of a periodic object, a relatively simple concept, electronics engineers found that it paralleled other transfer functions in their experience so closely that they had no hesitation in applying it to their own instruments. The concept was also readily accepted by optical physicists, who, with their understanding of the diffraction theory of image formation and with laboratories equipped for physical optical experiments, proceeded to make their own equipment for OTF measurements. Unfortunately for later correlation of these measurements, the optical physicists typically relied on optical test benches that had been designed only for *visual* assessment of image quality.

The performance of the earliest OTF measuring equipment was checked by the ability either to produce the Fourier transform of a known aperture or to indicate the OTF of a well-corrected microscope lens or collimater. Later experience was to show that these checks lacked an adequate test of bench alignment, and they fell short of predicting what the equipment would do when the lens under test had several wavelengths of aberration. These deficiencies showed up in a test program conducted under the auspices of the Avionics Laboratory at the Wright Air Development Center. A 12-in. $f/4$ Covogon aerial lens was circulated around eight laboratories, which were equipped to make OTF measurements. The spread of the resulting MTF curves was significantly in excess of the likely errors predicted for the various instruments. The obvious conclusion: Measurements of MTF had not yet become accurate enough to justify its use in standard specifications.

During the 1960s, lenses for low-light level television, microfiche storage, and microcircuit applications called for the development of a reproducible, objective, and reliable method of specifying and measuring image quality. To bring the measurement state of the art up to the standard required, it was suggested at a meeting of the International Society of Photogrammetry (London, 1966) that a number of special reference standard lenses be designed and constructed for in-house measurement at the various laboratories. To serve this purpose, the standard lenses should produce a typical range of OTF curves and, yet, be of the simplest possible geometrical construction so that they could be accurately made at a reasonable cost.

As one might expect, improving the state of the art required more than just issuing a series of reference standard lenses. Intercomparisons of lenses had already shown that the cause of disagreement could be found only by a thorough study of all aspects of OTF calculation and measurement. The close collaboration between OTF laboratories around the world required to this end was appropriately provided by the SIRA (Scientific Instrument Research Association) Institute in England, which administered a collaborative group project entitled *Assessment and Specification of Image Quality* that was initiated on April 1, 1967. Participating group members included various commercial and noncommercial institutions from many parts of the world. The expressed aims of the group were:

1. To assess the accuracy of instruments for measuring OTF and veiling glare, to establish the procedures and conditions necessary for accurate measurement, and to provide simple means for checking and maintaining this accuracy.
2. To determine the relation between OTF and veiling glare and the performance of various optical systems when the influence of psychophysical or other considerations are taken into account.
3. To establish standard procedures for specifying performance.

To achieve these objectives, the group started by developing a range of reference standard lenses whose measured performance closely matched their theoretically predicted performance. These were tested in different laboratories and on different instruments to determine sources of error and disagreement. The results also indicated the level of accuracy that could be expected under controlled conditions. It is expected that series of reference standard lenses of this type will be generally available to laboratories for routine checking of the OTF instruments.

Adjusting the OTF art to practice could hardly have been achieved without the modern electronic calculator. An engineer can accomplish a large part of the data processing with a hand-held programmable calculator. However, the amount of data to be processed in typical OTF calculations is still great enough to require computer time; in fact, in some instances, compromises have to be made between time and accuracy. Thus, the appearance of the Cooley–Tukey algorithm and subsequent *fast* and *fast–fast* procedures for calculating the Fourier transform and autocorrelation [46–50] were particularly timely for extensive application of the OTF.

As already suggested by our emphasis on calculator and computer efficiency, the OTF is seldom measured directly but is actually derived from some other measured optical parameter. Among the primary measurements is the determining of wavefront distortion by interferometric experiments. The OTF can also be calculated from data obtained by scanning the appropriate image in test setups for measuring point spread function, the line spread function, or the edge trace. Hence, development of both the hardware and the software of data processing systems, as well as of optical techniques and instruments, in the 1950s and 1960s contributed directly to practical OTF measurement.

During the 1960s, Hopkins continued his contributions relating to OTF with the introduction of what he has called the *canonical pupil coordinates* [51].

A second seminar on OTF was held at the University of Rochester under the sponsorship of the SPIE [52].

ACCEPTANCE–THE 1970s

Though still too close for good perspective, the 1970s appear to have been a decade in which OTF received general acceptance for optical performance evaluation. Acceptance was helped by the vacuum that was left when limiting resolution was discredited as a comprehensive indicator of optical imaging properties at the beginning of the decade. Along with acceptance came a better perception of where OTF techniques fell short and where further development work was required.

The National Bureau of Standards has established an optical measurement service, which is available to commercial firms and government agencies that want NBS authentication of their products and services. The new service also fills a gap for organizations that find it impractical to operate their own measurement equipment [53].

An indication of the breadth of interest in OTF was that most of the papers dealt with some aspect of that subject at the May 1974 seminar on *Image Assessment and Specification* in Rochester, New York [54].

In spite of the many refinements in OTF techniques, no one has been able to derive the OTF of a cascaded system from the OTFs of the parts. Nor has anyone done very well at diagnosing specific defects, except for gross aberrations, from the OTF. In other words, *how* an optical system fails can be described in detail, but the *why* remains obscure. As the OTF art matures, one can expect that its limitations, as well as its capabilities, will be usefully defined.

The advances in calculating techniques, which were important in advancing OTF measurements in the 1960s, continued through the 1970s. As small computers became increasingly available, routine use of the OTF in specifications and design became even more feasible. Comprehensive assessments that would require a series of limiting computer runs could now be accomplished quickly and inexpensively by using these small dedicated computers.

During the 1970s, the SIRA group, which was set up in the 1960s, began to produce results. Sources of errors in OTF measurements were found to be lack of control precision and stability in lens test benches, poor spectral data on light filters, poor-quality slits, and insufficient control of incoherence of the illumination. The use of standard reference lenses with known performance characteristics turned out to be a great help in tracking down sources of error and in training of laboratory personnel. A satisfactory level of agreement between different OTF measurements has been generally achieved over small field angles, but more work is required to get similar agreement for large angles off axis.

Although difficult to document, increased performance-to-cost ratios appear to have been achieved in the optical industry by firms that depend on routine OTF testing.

THE 1980s

By 1980, the momentum of OTF development was substantial, and it has remained so on into the decade. Already on the horizon is the extension of present successful practice in conventional designs to wide-angle systems, that is, systems with high numerical aperture and wide field, and also to even more unusual and exotic systems. Evidence that the OTF is being widely used is that advances in the science and art of the OTF now generally accompany advancements in other aspects of the science and art of lens design and of optical system measurement and evaluation. At an international conference on lens design in 1980 [55], 78 papers were presented, and the theme running throughout the program was the developments achieved in both the hardware and software for designing lenses. Every paper, it seems, discussed a lens design program, and every program had at least an option for calculating the OTF from design data. The conference stressed greater computer power for lens design. In the 1980s the microcomputer, especially the small computer dedicated to optical design calculations, is coming into widespread use.

As testimony to the explosion of activity in optical design, specification, and assessment, five other conferences were held during the first five years of the 1980s, almost on an annual basis. The international interest in optics continued to be evident in that one conference was in Oxford, England, and another in Geneva. Papers were presented by recognized authorities that discussed familiar topics, which indicated that continuing investigations are being rewarded by significant progress. Among the topics were international intercomparisons of MTF measurements on a 50-mm, $f/2$, camera lens, further activities of the SIRA group, optical quality and assessment, and the practical comparison of aberration polynominals for optical design. The Zernike polynomials seem more evident in the papers than the traditional Seidel polynomial.

Algorithms for calculating the Fourier transform are being steadily improved, and algorithms faster than the widely used Cooley–Tukey are now available. The Winograd Fast Fourier Transform (FFT) seems to be the most promising [56]. This algorithm can be easily and efficiently adapted to computation of the OTF, and results obtained with it come faster than with other methods without loss in accuracy.

A pattern that pervades the recent optical literature is that a well-developed lens design program grows with the user's experience in optics. A modern comprehensive optimization procedure has been typically fine-tuned by years of revision and use, each advance having been carefully examined so that it is a clear benefit to the user. A modern program is no longer merely a ray-tracing routine but includes facility with complex optical problems such as tilted and decentered elements, ''heads-up'' displays, and multistage infrared designs for airborne or laser telescopes in which packaging needs call for off-axis or oddly shaped components. This means that the computer handles details like the task of proper selection of damping factors, derivative increments, and constraint monitoring and control, thus allowing the optical designer to concentrate on the solution of more subtle optical problems.

In addition to the familiar damped-least-squares method of optimization, newer methods are coming into practice; one example is the pseudo-second-derivative (PSD) in a program called SYNOPSYS described by Dilworth [57].

Several research students at the Imperial College, London, have been studying improved optical design software with particular emphasis on programs for OTF calculation.

A second international lens design conference was held in June 1985 at Cherry Hill, New Jersey (see Ref. [58]).

PERSPECTIVE

Expansion of OTF practice into areas so far not exposed to its benefits will probably require deliberate communication effort by it users. Included among the people who might profit by knowing the fundamentals and subtleties of the OTF are practicing optical engineers and technicians. Buyers and users of optical lenses and systems can benefit by digging under the buzz words to learn how the OTF may be applied to their specifications and testing.

Now, after many hours of searching the literature, scanning and studying published papers, writing about the OTF, and contemplating the practical aspects of the OTF and its use, we have a growing feeling that the future of the OTF depends not on the professionals, the scholars in optics, and the designers of optical systems, but on those who should perhaps be called the ''paraprofessionals''—the users–purchasers, the preparers of specifications, and the evaluators. The latter group understands the OTF, but generally not quite to the point of visualizing, for instance, just what and how specific MTF changes modify the appearance of a particular object. A fundamental concept necessary to the intelligent and powerful use of the OTF is an understanding of how the low-pass filtering of spatial frequencies alters the distribution of radiant flux density in the object. A simple example of this phenomenon is the rounding off of the edges of a square-wave distribution by the attenuation of high spatial frequencies in the spectrum of a square wave.

An evaluator studying a lens under test cannot directly see the spatial frequency spectrum nor the spatial frequency distribution in the image whereas the flux density distribution is manifest. Hence, when using a criterion based on the OTF, the evaluator is always one step removed from a direct observation. In spite of this complication, the OTF is here to stay, and it is being used to advantage in a number of practical ways. Therefore, we expect the art, at least in its application, will mature; and we anticipate not slowly. This is why we believe continued work on the assessment criteria and their measurement is of great importance.

A salable feature of the OTF is that its evaluation applications are often cost-saving replacements of other procedures, but greater gains from the OTF are likely where it assumes some role that cannot be handled in any other way. This occurs, for instance, when sensitive techniques are needed for aberrations of just a few wavelengths or less. Here, where maximum wave-front discrepancies are on the order of 10 μm, spot diagrams and other geometrical optics techniques become impotent, whereas OTF methods perform well.

Among the characteristics that give the OTF an advantage over other means of evaluating image quality are the following:

1. The validity of an OTF depends primarily upon linear relations between object and image characteristics. Under these conditions, the user can postulate the legitimacy of superposition and shift invariance.
2. Application of the OTF does not require any specific theory of light, any particular spectral composition, any assumptions about the shape of the aperture stop, or any limitation on the type or magnitude of aberrations (provided the shift-invariance postulate is not violated).
3. The OTF for any given region of the image provides a complete analytical description of the object–image relation for that region without any restriction on the specific form of the test object.
4. The OTF is a direct application of the highly developed Fourier transform theory in which a two-dimensional variation of intensity over the object plane is analyzed into a two-dimensional spectrum of spatial frequencies. The OTF then describes how each of these Fourier components is attenuated in amplitude and shifted in phase as it appears in the image. Alternatively, any two-dimensional object may be treated as the superposition of a set of one-dimensional objects in different azimuths, each azimuth requiring only a one-dimensional transfer function.
5. The OTF can be both calculated directly from the design data of any system and measured for that system after fabrication. Thus, for the first time, exactly the same index of quality can be calculated and measured for evaluation of the fabrication process and the precision of measurement.

These characteristics are discussed in the appropriate chapters of this book.

In retrospect, it seems reasonable to suspect that wide application of the OTF for assessing image quality in industry might have been set back as much as a decade by the flood of publications on techniques and instruments that failed to give adequate attention to precision of instrumentation. When these methods failed in production, OTF received an ill-deserved black eye that took some time to heal.

REFERENCES

1. J. W. Krutch, *Human Nature and the Human Condition.* Random House, New York, 1959, p. 51.
2. H. H. Hopkins, The Development of Image Evaluation Methods. *SPIE Proc.* **46,** 2 (1974).

   Note: Numerous references to papers in various issues of *SPIE Proc.*, which are proceedings of technical papers presented at conferences or seminars and which are published by the Society of Photo-Optical Instrumentation Engineers (SPIE), are made throughout this book. To obtain copies of any volume or to obtain information about proceedings that are available, write to the Society of Photo-Optical Instrumentation Engineers, P.O. Box 10, Bellingham, WA 98225.

3. L. R. Baker, Status of OTF in 1970. *Opt. Acta* **18,** 81 (1971).
4. F. D. Smith, Optical Image Evaluation and the Transfer Function. *Appl. Opt.* **2,** 335 (1963).
5. H. Friesser, Photographic Resolution of Lenticular Films. *Zeits. f. Wiss. Photogr.* **37,** 261 (1938).
6. G. B. Airy, On the Diffraction of an Object Glass with Circular Aperture. *Trans. Cambridge Philos. Soc.* **5,** 283 (1835).
7. L. Foucault, Mémoir sur la Construction des Télescopes en Verre Argenté. *Ann. de L'Observatoire Imp. de Paris* **5,** 197 (1859).
8. J. W. Strutt (Lord Rayleigh), On the Theory of Optical Images, with Special Reference to the Microscope. *Philos. Mag.* **42,** 167 (1896). (Reprinted in *Scientific Papers of Lord Rayleigh*, Vol. 4. Cambridge Univ. Press, London, 1899, p. 235.
9. E. Abbe, Beiträge zur Theorie des Mikroskops und der Mikroskopischen Wahrnehmung. *Arch. Mikrosk. Anat.* **9,** 413 (1873).
10. R. W. Ditchburn, *Light.* Academic, New York, 1963, pp. 291–296.
11. A. W. Conway and J. L. Synge (Eds.), *The Mathematical Papers of Sir William Hamilton*, Vol. 1. Cambridge Univ. Press, London, 1931.
12. O. N. Stavroudis, *The Optics of Rays, Wavefronts, and Caustics.* Academic, New York, 1972.
13. M. Born and E. Wolf, *Principles of Optics.* Pergamon, New York, 1965.
14. H. Bruns, Das Eikonal. *Leipsig. Abh. Kgl. Sachs. Ges. Wiss.* (*Math.-Phys. Kl.*) **21,** 370 (1895).
15. von E. Lommel, Die Beugungserscheinungen einer kreisrunden Oeffnung und eines kreisrunden Schirmchens, *Abh. K. Bayerischen Akad. der Wiss.* (*Math.-Phys. Cl.*) **15**(2), 233 (1885); **15**(3), 531 (1886).
16. J. W. Strut (Lord Rayleigh), On the Theory of Optical Images, with Special Reference to the Spectroscope. *Philos. Mag.*, **8,** 403 (1879). (Reprinted in *Scientific Papers of Lord Rayleigh*, Vol. 1. Cambridge Univ. Press, London, 1899, pp. 432–435.)
17. K. Strehl published several papers during the years between 1898 and 1905 (See for example K. Strehl, *Z. f. Instrumkde.*, **22,** 213 (1902).); however, it was in his book, *Die Beugungstheorie des Femrohrs*, published near the end of the nineteenth century that he first discussed the loss of light power from the central disk to the ring system when aberrations are present.
18. A. E. Conrady, *Applied Optics and Optical Design.* (Please see Ref. 3 of Chapter 6 for more information.)
19. A. Maréchal, Study of the Combined Effect of Diffraction and Geometrical Aberrations on the Image of a Luminous Point. *Rev. d'Optique* **26,** 257 (1947).

20. W. D. Wright, Television Optics. In *Reports on Progress in Physics*, Vol. 5. Institute of Physics and the Physical Society, London, 1938, p. 203.

21. O. H. Schade, Electro-Optical Characteristics of Television Systems. *R. C. A. Rev.* **9,** 5, 245, 490, 653 (1948).

22. E. W. H. Selwyn, The Photographic and Visual Resolving Power of Lenses. *Photogr. J. B* **88,** 6, 46 (1948).

23. R. K. Luneberg, *Mathematical Theory of Optics.* Wiley-Interscience, New York, 1965.

24. P. M. Duiffieu, *L'Intégral de Fourier et ses Applications à L'Optique.* Besancon, 1946. (Printed privately but now available as a second edition from Masson et Cie, Paris, 1970.) (Also available from Wiley, New York, as a book in the series on Pure and Applied Optics, 1983.)

25. K. Nienhuis and B. R. A. Nijboer, Diffraction Patterns in the Presence of Small Aberrations. *Physica* **10,** 679 (1947); **13,** 605 (1947); **14,** 590 (1948).

26. N. G. Van Kampen, On Asymptotic Treatment of Diffraction Problems. *Physica* **14,** 575 (1949); **16,** 817 (1950).

27. H. H. Hopkins, *Wave Theory of Aberrations.* Oxford Univ. Press, Oxford, 1950.

28. P. Elias, D. S. Grey, and D. Z. Robinson, Fourier Treatment of Optical Processes. *J. Opt. Soc. Am.* **42,** 127 (1952).

29. *Symposium on the Evaluation of Optical Imagery, 1951.* NBS Circular 526 (1954).

30. H. H. Hopkins, The Frequency Response of a Defocused Optical System. *Proc. R. Soc. London Ser. A* **231,** 91 (1955).

31. H. H. Hopkins, The Frequency Response of Optical Systems. *Proc. Phys. Soc. (London) Ser. B* **69,** 562 (1956).

32. E. L. O'Neill, *Introduction to Statistical Optics.* Addison-Wesley, Reading, MA, 1963.

33. K. Rosenhauer and K.-J. Rosenbruch, The Measurement of the Optical Transfer Functions of Lenses. In *Reports on Progress in Physics*, Vol. 30, Part 1, A. C. Stickland (Ed.). Institute of Physics and the Physical Society, London, 1967.

34. H. H. Hopkins, Geometrical-Optical Treatment of Frequency Response. *Proc. Phys. Soc. (London) Ser. B* **70,** 1162 (1957).

35. M. De, The Influence of Astigmatism on the Response Function of an Optical System. *Proc. R. Soc. London Ser. A* **233,** 91 (1955).

36. N. S. Bromilow, Geometrical-Optical Calculation of Frequency Response for Systems with Spherical Aberration. *Proc. Phys. Soc. (London) Ser. B* **71,** 231 (1958).

37. A. S. Marathay, Geometrical Optical Calculation of Frequency Response for Systems with Coma. *Proc. Phys. Soc. (London)* **74,** 721 (1959).

38. K. Miyamoto, On a Comparison between Wave Optics and Geometrical Optics by Using Fourier Analysis, 1. General Theory, *J. Opt. Soc. Am.* **48,** 57 (1958).

39. E. H. Linfoot, Convoluted Spot Diagrams and the Quality Evaluation of Photographic Images. *Opt. Acta* **9,** 81 (1962).

40. K. Miyamoto, Wave Optics and Geometrical Optics in Optical Design. In *Progress in Optics*, Vol. 1, E. Wolf (Ed.). North Holland, Amsterdam, 1961, p. 31.

41. G. Black and E. H. Linfoot, Spherical Aberration and the Information Content of Optical Images. *Proc. R. Soc. London Ser. A* **239,** 522 (1957).

42. W. H. Steel, A Study of the Combined Effects of Aberration and a Central Obscuration of the Pupil on Contrast of an Optical Image. Application to Microscope Objectives Using Mirrors. *Rev. d'Optique* **32,** 4, 143, 269 (1953).

43. R. Barakat, Computation of the Transfer Function of an Optical System from the Design Data for Rotationally Symmetric Aberrations. Part I. Theory. *J. Opt. Soc. Am.* **52,** 985 (1962); Part II. Programming and Numerical Results, with M. V. Morello, *J. Opt. Soc. Am.* **52,** 992 (1962).

44. A. M. Goodbody, The Influence of Spherical Aberration on the Response Function of an Optical System. *Proc. Phys. Soc. (London)* **72,** 411 (1958).

45. A. M. Goodbody, The Influence of Coma on the Response Function of an Optical System. *Proc. Phys. Soc. (London)* **75,** 677 (1960).

46. J. W. Cooley and J. W. Tukey, An Algorithm for Machine Calculation of Complex Fourier Series. *Math. Comput.* **19,** 296 (1965).

47. M. L. Forman, Fast Fourier-Transform Technique and Its Application to Fourier Spectroscopy. *J. Opt. Soc. Am.* **56,** 978 (1966).

48. S. H. Lerman, Application of the Fast Fourier Transform to the Calculation of the Optical Transfer Function. *SPIE Proc.* **13,** 51 (1969). (Please see the note following Ref. 2.)

49. E. O. Brigham, *The Fast Fourier Transform.* Prentice-Hall, Englewood Cliffs, NJ, 1974.

50. B. Liu (Ed.), *Digital Filters and the Fast Fourier Transform.* Academic, New York, 1975. (*Benchmark Papers in Optics*, published by Dowden, Hutchinson, and Ross.)

51. H. H. Hopkins, Canonical Pupil Coordinates in Geometrical and Diffraction Image Theory. *Japan. J. Appl. Phys.* **4,** Suppl. 1, 31 (1965).

52. R. Wollensak and R. R. Shannon (Eds.), *Modulation Transfer Function. SPIE Proc.* **13** (1969). (Please see the note following Ref. 2.)

53. R. E. Swing, The Case for the Pupil Function. *SPIE Proc.* **46,** 104 (1974). (Please see the note following Ref. 2.)

54. D. Dutton (Ed.), *Image Assessment and Specification. SPIE Proc.* **46** (1974). (Please see the note following Ref. 2.)

55. R. E. Fischer (Ed.), *1980 International Lens Design Conference. SPIE Proc.* **237** (1980). (Please see the note following Ref. 2.)

56. D. Heshmaty-Manesh and S. C. Tam, Optical Transfer Function Calculations by Winograd Fast Fourier Transform, *Appl. Opt.* **21,** 3273 (1981).

57. D. C. Dilworth, Pseudo-Second-Derivative Matrix and Its Application to Automatic Lens Design. *Appl. Opt.* **17,** 3372 (1978).

58. W. H. Taylor and D. T. Moore (Eds.), *1985 International Lens Design Conference. SPIE Proc.* **554** (1986). (Please see the note following Ref. 2.)

2

Concepts

INTRODUCTION

For the OTF to become a part of our everyday thinking, we need to understand the full meaning and the essence of certain basic concepts; we also need to acquire a working knowledge of related principles and to attain a facility with certain mathematical procedures. Pertinent concepts are discussed in this chapter; prominent among them are:

1. *Spatial frequency*, *spatial frequency spectrum*, and *distribution*.
2. *Contrast* and *contrast transfer* in an optical system.
3. *Point spread function*, *line spread function*, and *edge trace*.
4. *Isoplanatism.*
5. *Linear superposition*.

We direct attention to three specific distributions: *point spread function*, *line spread function*, and *edge trace*. Two especially important mathematical concepts are those of *convolution* and *autocorrelation*, which are discussed in Appendix B. To convert a distribution in a space domain into a spatial frequency spectrum, which is an equivalent distribution in the frequency domain, the Fourier transform is applied in a number of examples. The significance of these concepts and procedures becomes apparent in later chapters when, for example, an optical system is found to function as a low-pass filter of spatial frequencies. In the language of *systems analysis*, the optical system is a two-dimensional, space-invariant, fixed-parameter linear system; and the point spread function is its impulse response.

Two further basic concepts are *wave aberration* and *wave aberration function*, which are treated at length in later chapters.

SPATIAL FREQUENCY

Figure 2.1 is a picture of a white picket fence of regular spacing. Measurements on the actual fence indicated that each picket and each space are the same width,

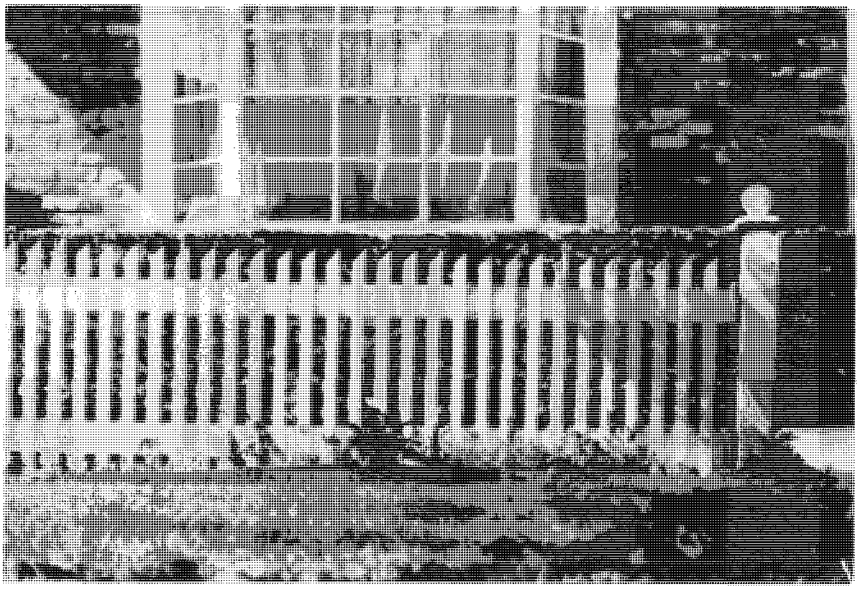


Figure 2.1. The white picket fence has 8.06 pickets per meter.

6.2 cm, which means that a grasshopper progressing along the horizontal rail would encounter a new picket each time the distance of 12.4 cm is traveled. Further, he would note (if sufficiently astute) after moving a measured distance of one meter that he had gone 8.06 times through the repeating cycle of a picket followed by a space. The fundamental horizontal *spatial frequency* of the repeating cycle along the fence could be described then as 8.06 cycles per meter. Measurements on the picture of the fence in Fig. 2.1 would lead to a corresponding spatial frequency of the white and dark cycles; but, of course, the dimensions and frequency would be different because of the change of scale. Also, the description would no longer be of physical pickets and the spaces between them but rather of the changing color from white to black in cycles as our eyes scan along the picture of the fence. The physical quantity in the picture is the changing capability to reflect light, which gives the pattern of black, gray, and white regions. In this connection, *reflectance* is defined as the fraction of the incident light that is reflected.

To simplify the mathematical analysis of the picket-fence image, the picture of Fig. 2.1 has been reduced to the bar pattern shown in Fig. 2.2. Here the reflectance of the pickets is assumed unity and that of the spaces, zero (complete reflectance alternating with complete absorption). A graph of the reflectance as a function of distance along the fence is shown in Fig. 2.3. Because the ordinate

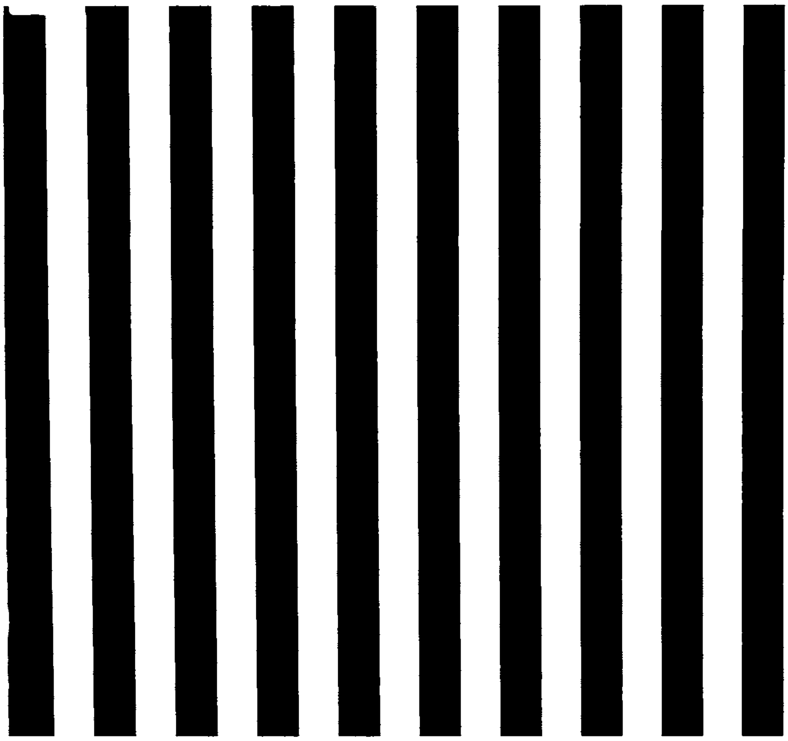


Figure 2.2. A "perfect" picket fence constituting a bar chart.

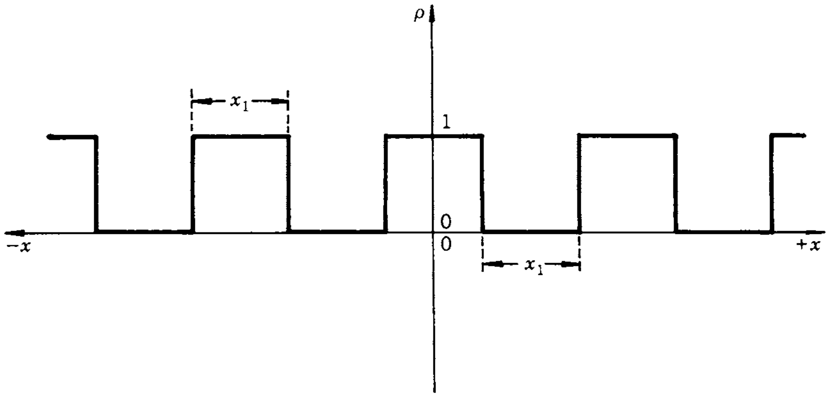


Figure 2.3. Reflectance as a function of distance along the fence.

in the figure represents reflectance ρ, which is a function of x alone, values of reflectance at points in the picture are given by the mathematical expression for $\rho = f(x)$, and the expression is

$$f(x) = \rho(x) = 1 \quad \text{when} \quad (-x_1/2) \leq x \leq (+x_1/2)$$

and

$$(2n + \tfrac{3}{2})x_1 \leq |x| \leq (2n + \tfrac{5}{2})x_1,$$

and

$$f(x) = \rho(x) = 0 \quad \text{when} \quad (2n + \tfrac{1}{2})x_1 \leq |x| \leq (2n + \tfrac{3}{2})x_1,$$

for

$$n = 0, 1, 2, 3, \cdots . \tag{2-1}$$

In these expressions we have assumed that the fence extends very far in both directions, which corresponds to allowing n to become arbitrarily large. Equation (2-1) describes mathematically the *distribution* of reflectance in the bar chart of Fig. 2.2 and the graph of Fig. 2.3.

FLUX DENSITY AND DISTRIBUTIONS

For the purpose of our discussion, *distribution* can be defined as a point function of position representing values of some physical quantity that varies in a prescribed way over the surface (for instance, over the image plane). *Distribution* refers to the manner in which the physical quantity varies. For example, in the picture of Fig. 2.2, the pattern of bars shows how reflectance is distributed. Other quantities represented in distributions are *intensity*, *flux density*, and even the *complex amplitude* over a phase front (wave front) in a coherent light beam. In discussions of *flux* the term may apply to the amount of radiant energy in joules per second, or it may be the amount of luminous energy in lumens, which is visible light that has been evaluated for its efficacy to stimulate the human eye. *Flux density*, a common term, depends on the context to tell whether incident, reflected, or emitted flux is involved. Also, whether radiant energy or luminous energy is meant ordinarily is evident from the topic under consideration. Intensity and flux density are defined as follows:

Intensity, I, is the amount of flux leaving a point source per unit of solid angle.

Flux density, $\mathcal{W}$, is the amount of flux incident upon or leaving a surface per unit of surface area.

Complex amplitude is discussed in Appendix C and in Chapter 4 in connection with diffraction.

Besides the example given in Eq. (2-1), where $f(x)$ represents reflectance at points x, a distribution could be $\mathcal{W}(\xi, \eta)$, representing flux density at points (ξ, η).

In the bar chart of Fig. 2.2, the combination of a white bar and an adjacent black bar constitutes a cycle. As in the primitive example involving the grasshopper, the number of such cycles occurring in a unit distance constitutes the spatial frequency of the pattern. The length of one cycle or *period* is the width of a white bar plus the width of a black bar. In Eq. (2-1) and in Fig. 2.3, a period is equal to $2x_1$. The frequency ω and the period are reciprocals:

$$\omega = 1/(2x_1). \tag{2-2}$$

In the example, the units of ω are m^{-1}; however, cm^{-1} and mm^{-1} are also commonly used. In the bar chart, we have measured the frequency and period perpendicular to the bars (which is generally done in simple patterns of this type).

The graph of Fig. 2.3 is usually referred to as a *square wave* of reflectance; but because, strictly speaking, it does not represent a wave like the moving swell on the surface of water, terms like *crenalate distribution* and *top hat distribution* of reflectance are sometimes preferred.

The sinusoidal distribution shown in Fig. 2.4 is especially convenient in the analysis of complex patterns. It can be mathematically represented by

$$f(x) = a + b \cos 2\pi\omega_1 x$$

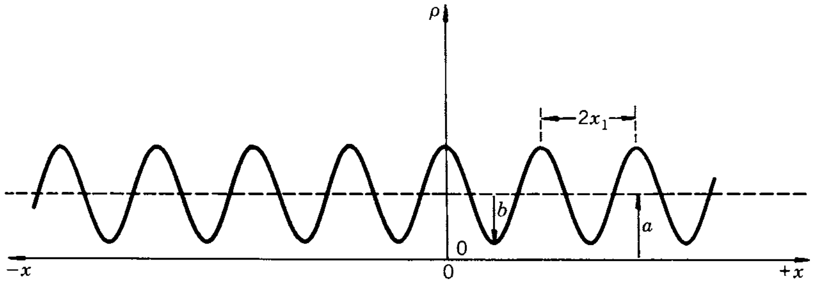


Figure 2.4. A sinusoidal distribution of reflectance.

or by

$$f(x) = a + b \cos \pi x/x_1, \qquad (2\text{-}3)$$

where

$$a + b \leq 1, \qquad b \leq a,$$

and

$$\omega_1 = 1/(2x_1).$$

In Eq. (2-3), a is the average reflectance, sometimes referred to as the "dc level." The amplitude of the cosine distribution (or "wave") is b.

In practice the application of the spatial frequency definition (number of cycles per unit distance) to sinusoidal distributions like the one in Fig. 2.4 is straightforward; but square wave or bar distributions (Fig. 2.2) are often specified by the number of "lines" per unit distance, and this practice has led to some confusion. The question is whether the combination of a black and an adjacent white bar constitutes a line or each bar is counted separately. This ambiguity can be avoided by always referring to line *pairs*. The number of these in a unit distance will be equal to the frequency. Unless some other unit is clearly specified, the practice in this book is to give the frequency in cycles per millimeter (units: mm^{-1}).

FREQUENCY SPECTRUM

The square wave distribution represented by Eq. (2-1) can also be expressed as the sum of cosine terms as

$$f(x) = 0.5 + \sum_{n=1,5,9,\cdots} \frac{4}{n\pi} \cos \frac{2n\pi x}{2x_1} - \sum_{n=3,7,11,\cdots} \frac{4}{n\pi} \cos \frac{2n\pi x}{2x_1}. \qquad (2\text{-}4)$$

(See [1, p. 26].) Each of the cosine terms is a sinusoidal *component* of the total distribution. We speak of analyzing, or breaking down, a distribution such as the square wave into its components. The sum, implying the physical superposition, of all components represented in Eq. (2-4) is the square wave distribution of Eq. (2-1) and Fig. 2.3. In fact, any realizable periodic distribution of reflectance or flux density can be represented by a Fourier series of sine and cosine functions of distance. Each of the terms is recognized as a sinusoidal frequency, the lowest one of which is equal to the frequency of the original periodic function that is represented by the series. Strictly speaking, an infinite number of terms would be required to make their sum exactly duplicate the original function; but in many applications of Fourier series, the partial sum converges so rapidly to the true value with the addition of terms that the sum of just a few terms is a good approximation of the function.

The series representation of any periodic function as illustrated by Eq. (2-4) has profound analytical significance. The fact that all arbitrary periodic distributions are composed of but one kind of building block, the sine wave, allows an investigator to explore the response of a given system to one sinusoidal distribution at a time. (See [3].) Then the system response may be generalized by displaying (for example, by plotting a curve of response versus frequency) the individual response to each of the several components of an arbitrary distribution. When it is useful for the analysis, the individual responses can be summed (with attention to phase) to arrive at the complete response distribution, which then can be compared with the applied arbitrary periodic distribution that produced it.

An obvious objection to a Fourier series analysis is that the reflectance and flux density distributions encountered in optics are not generally periodic. To meet this objection, the method of analysis has to be developed one step further.

When the terms of Eq. (2-4) are examined as representative of Fourier series terms in general, it is apparent that as the period $2x_1$ of the function is lengthened, the interval between successive frequencies ($1/(2x_1)$, $3/(2x_1)$, $5/(2x_1)$, etc.) is shortened. By allowing the period to expand without limit, the interval between frequencies can be made arbitrarily small, to any degree desired; and, finally, a continuous spectrum of frequencies results. This limit corresponds to a period of unbounded length; that is, a single cycle of the function extends over all values of x, negative and positive. In other words, the function analyzed over a finite path, however long, need no longer be periodic. Mathematically, at the limit, the Fourier series becomes a Fourier integral. (It should be noted, however, that the actual mathematical derivation of the Fourier integral can be more conveniently handled by not involving the Fourier series. The discussion here is intended to establish a conceptual relationship.) The Fourier transform to convert a distribution as a function of distance $f(x)$ to the same distribution as a function of frequency $F(\omega)$ is (see [2] and also Appendix B)

$$F(\omega) = \int_{-\infty}^{+\infty} f(x) \exp(-i2\pi\omega x)\, dx. \tag{2-5}$$

The reverse transformation, to convert a distribution as a function of frequency to the same distribution as a function of distance, is

$$f(x) = \int_{-\infty}^{+\infty} F(\omega) \exp(i2\pi\omega x)\, d\omega. \tag{2-6}$$

The function $F(\omega)$ as defined in Eq. (2-5) is called the *frequency spectrum* or the *Fourier transform* of $f(x)$. Because the exponent in the Fourier transform is imaginary, the expression for $F(\omega)$ is generally complex, giving both the amplitude and the phase of each sinusoidal frequency component making up the spectrum. However, the particular spectrum for a three-bar pattern, which is discussed with its spectrum (Eq. (2-7)) in the next section, is real (the phase is zero or π radians), meaning that all component waves are either in phase or π radians out of phase at the origin of x. Application of Fourier analysis is illustrated in a number of examples that follow.

THREE-BAR PATTERN SPECTRUM

In the picket fence example illustrated in Figs. 2.1–2.3, and formulated in Eq. (2-1), an unbounded number of pickets (or bars) has been assumed. The Fourier transform of $f(x)$ in Eq. (2-1) can be worked out as a problem for bounded n (finite number of pickets). When this transform is applied for three pickets or bars as shown in Fig. 2.5, the spectrum is given by

$$F_3(\omega) = \frac{\sin 6\pi x_1 \omega}{2\pi\omega \cos \pi x_1 \omega}. \tag{2-7}$$

A portion of this spectrum is shown in Fig. 2.6 where F_3 is plotted as a function of ω. Relatively great local maxima in the amplitude are noted in the vicinity of frequency values $1/(2x_1)$, $3/(2x_1)$, and $5/(2x_1)$. These can be related to

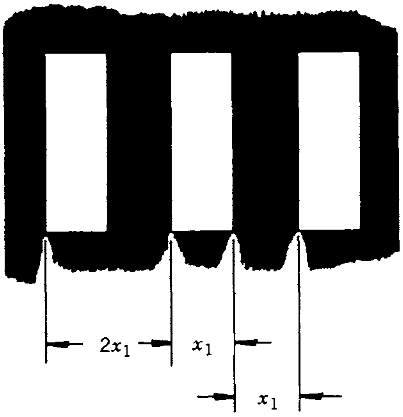


Figure 2.5. A three-bar pattern.

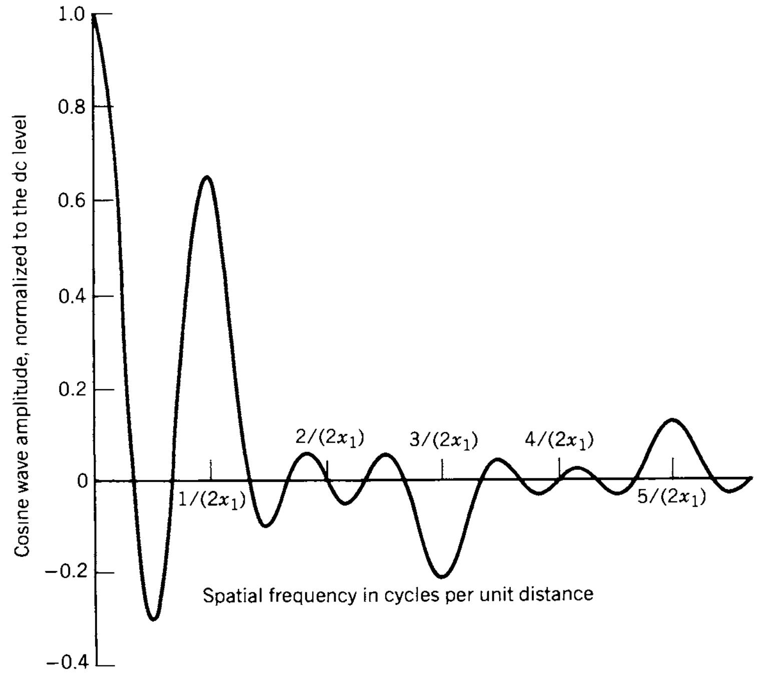


Figure 2.6. Frequency spectrum of the three-bar pattern shown in Fig. 2.5.

the frequency of the bar pattern ("pickets") in Fig. 2.5 where $2x_1$ is shown to be the period of a cycle and x_1 the width of a single bar (black or white). Therefore, from our discussion of spatial frequency, the first local maximum occurs near the frequency of the bar pattern, the second near three times the pattern frequency, and the third near five times the pattern frequency. These three frequencies are called the *fundamental*, *third harmonic*, and *fifth harmonic*, respectively. (In this sequence, the fundamental could also be called the *first harmonic*, but this term is rarely used.)

Further observations based on the example illustrated in Fig. 2.5 and Fig. 2.6 suggest general characteristics of Fourier analysis. Some of these will be emphasized in additional examples.

If the three-bar "pulse" of Fig. 2.5 were increased to more bars at the same spatial frequency, the plot corresponding to Fig. 2.6 would show narrower and more pronounced peaks at the fundamental and the odd harmonics. As the number of bars approaches infinity, Fourier theory indicates that all frequencies except the fundamenal and the odd harmonics vanish; and the amplitude plot becomes a series of vertical lines at the fundamental and odd harmonic frequencies. As indicated earlier, the spectrum would be represented by an infinite sinusoidal series rather than an integral. In all instances, because the average of the bar function is positive, the plot that is a function of frequency would have a positive, nonzero value at the origin ("dc" or zero frequency).

EVEN AND ODD HARMONICS AND FUNCTIONS

In the discussion of harmonics, one could well ask why there are no even (second, fourth, etc.) harmonics in the multiple-bar spectrum. The absence of these frequency components results from choosing an $f(x)$ that has the property

$$f(x + x_1) = -f(x) + \text{constant}, \tag{2-8}$$

where $2x_1$ is the period and the constant can have any real value, positive or negative. Any function having this kind of symmetry is represented by a Fourier series with only odd (fundamental, third, fifth, etc.) harmonics. If, instead of the function shown in Fig. 2.3, the flat top were retained but the corners at the axis were rounded, even (as well as odd) harmonics would appear in the series. (See [1, p. 28].)

Although the discussion of Fourier series following Eq. (2-4) refers to both sine and cosine functions, only cosine functions appear in Eq. (2-4). This observation suggests that some property of $f(x)$ in Eq. (2-1) eliminates the sine terms that might be expected in the corresponding Fourier series. To discuss this property, the definitions of even and odd functions have to be used. If $f(x)$ is an even function, where $f(x) = f(-x)$, a series of cosines will result. If $f(x)$ is an odd function, where $f(x) = -f(-x)$, a series of sines will result. (Note that a cosine is an even function and a sine is an odd function.) When the even–odd property is explored in Fourier integral theory (see [2, p. 11]), one finds that the transform of a real and even function is another real and even function and that the transform of a real and odd function is another real and odd function. Also, the frequency components of even functions are cosine functions, and the frequency components of odd functions are sine functions.

The even–odd relations in Fourier integral theory suggest that Fourier mathematics can be considerably simplified by proper selection of the origin on the x-axis. By such a selection, the relatively simple expression of Eq. (2-7) resulted from the analysis of the three-bar configuration in Fig. 2.5. The origin of x was arbitrarily placed at the midpoint of the center bar, thus making $f(x)$ an even function as well as a real function. The Fourier transform of $f(x)$ plotted in Fig. 2.6 must, therefore, be a real and even function mathematically, although the negative frequencies have no physical meaning and are not shown on the graph. As indicated on the ordinate axis, the relative amplitudes plotted are of *cosine* waves.

The spectrum of Fig. 2.6 exhibits a general reduction of amplitude as the spatial frequency increases, which can be easily traced to the $2\pi\omega$ term in the denominator of Eq. (2-7). This property of practical spectra allows frequencies above some arbitrarily chosen limit to be ignored with little loss in accuracy.

A STEPLADDER BAR PATTERN

The bar pattern of Fig. 2.7 is made up of five shades of gray ranging from white to black with three intermediate steps. If reflectances are plotted along a horizontal path with the origin at the center of a white bar, Fig. 2.8 results, which shows one of an assumed infinite number of cycles. The expression for the corresponding spectrum is

$$F(\omega) = \rho_0 + \sum_{j=1}^{4} \sum_{n=1}^{\infty} \left[\frac{2j-1}{8} + \frac{2}{n\pi} \sin \frac{(2j-1)2n\pi}{16} \cos \frac{2n\pi x}{16x_1} \right]. \quad (2\text{-}9)$$

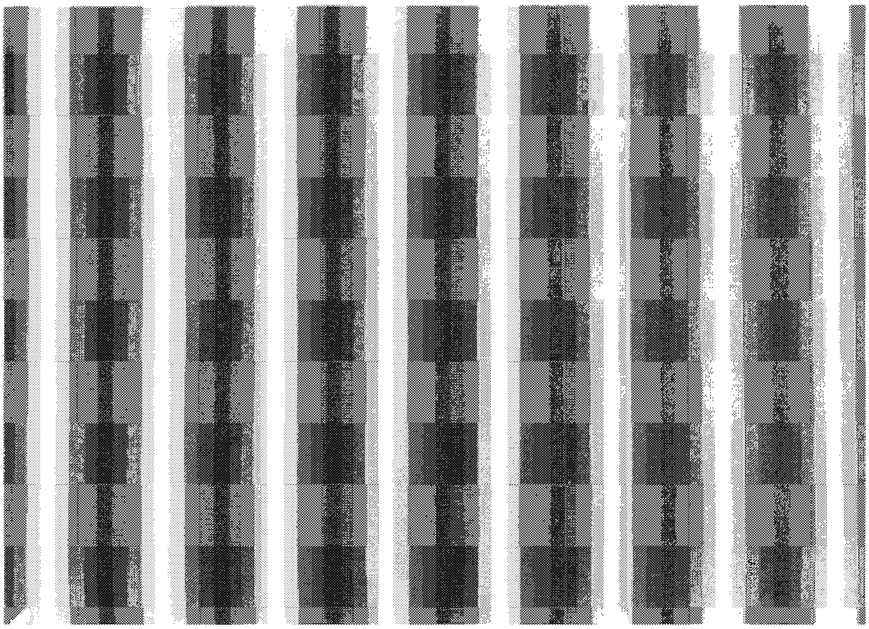


Figure 2.7. An involved bar chart having a more complicated frequency spectrum than the spectrum for the three-bar chart.

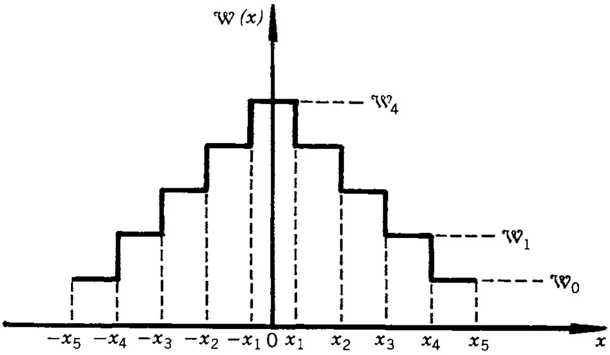


Figure 2.8. A plot of the reflectance as a function of distance across one cycle of the bar chart of Fig. 2.7.

In this expression, $16x_1$ is the period, and the fundamental ω_1 is $1/(16x_1)$. The constant (or dc) terms are

$$\rho_0 + \sum_{j=1}^{4} \frac{2j-1}{8} = \rho_0 + 2. \tag{2-10}$$

When the summation over j is completed, Eq. (2-9) becomes

$$F(\omega) = \rho_0 + 2 + \sum_{n=1}^{\infty} \left[\frac{2}{n\pi} \left(\sin\frac{n\pi}{8} + \sin\frac{3n\pi}{8} + \sin\frac{5n\pi}{8} + \sin\frac{7n\pi}{8} \right) \right] \times \cos\frac{2n\pi x}{16x_1}. \tag{2-11}$$

Again, the assumed function represented in Figs. 2.7 and 2.8 satisfies the condition of Eq. (2-8), so only odd harmonics appear in the spectrum. All harmonics (including the fundamental), expressed as $n/(16x_1)$ or $n\omega_1$ where $n = 1, 2, 3, \ldots$, at first seem to be present in Eq. (2-11); but substitution of any even integer for n in the coefficient of the cosine function results in a value of zero, whereas odd integers produce nonzero coefficients. Calculated amplitudes for the first few harmonics are given in Table 2.I and are shown in the spectrum plot of Fig. 2.9. The envelope of the amplitudes has to go to zero at multiples of $2/(16x_1) = 2\omega_1$ because even harmonics are not present.

Table 2.I Amplitudes of a Few Harmonics in the Spectrum for the Bar Chart in Fig. 2.7

Harmonic	Frequency	Amplitude
Fundamental	ω_1	1.66357
Third	$3\omega_1$	0.22969
Fifth	$5\omega_1$	0.13781
Seventh	$7\omega_1$	0.23765
Ninth	$9\omega_1$	−0.18484
Eleventh	$11\omega_1$	−0.06264
Thirteenth	$13\omega_1$	−0.05301
Fifteenth	$15\omega_1$	−0.11090
Seventeenth	$17\omega_1$	0.09786
Nineteenth	$19\omega_1$	0.03627
Twenty-first	$21\omega_1$	0.03281
Twenty-third	$23\omega_1$	0.07233
Twenty-fifth	$25\omega_1$	−0.06654
Twenty-seventh	$27\omega_1$	−0.02552
Twenty-ninth	$29\omega_1$	−0.02376
Thirty-first	$31\omega_1$	−0.05366
Thirty-third	$33\omega_1$	0.05041
Thirty-fifth	$35\omega_1$	0.01969
etc.		

SPECTRUM FOR A GENERAL DISTRIBUTION

In contrast to the bar charts illustrated earlier in this chapter, a picture like Fig. 2.10 does not suggest any recurring cycles of reflectance. Nevertheless, along a line in any direction, such as from A to B in the picture, recurring cycles at certain spatial frequencies actually do exist. In general, the spectrum of frequencies is continuous; but, as already demonstrated with the bar charts, some frequency components can have zero amplitudes. In fact, if one chooses a path parallel to the boundary of a bar, all frequency components, except the dc or constant level, have zero amplitudes.

The plot of Fig. 2.11 approximates the reflectance along the line AB in Fig. 2.10. If the coordinate along AB is ξ, the reflectance can be designated $f(\xi)$; it also meets the mathematical requirements for having a Fourier transform. The frequency spectrum, then, along AB is

$$F(\omega_\xi) = \int_{-\infty}^{+\infty} f(\xi) \exp(-i2\pi\omega_\xi \xi)\, d\xi. \tag{2-12}$$

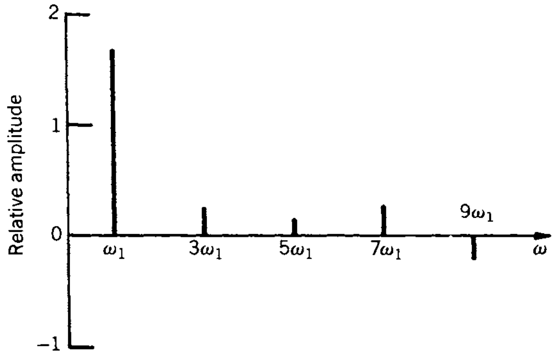


Figure 2.9. The beginning of the frequency spectrum for an infinite sequence of bars like those in Fig. 2.7.

Functions similar to the one shown in Fig. 2.11 are obtained by measuring the reflectance of prints with a scanning microreflectometer. This procedure does not provide a mathematical expression for $f(\xi)$; in fact, such expressions are seldom available. So experimental procedures are designed to provide values of $f(\xi)$ at specified intervals along the ξ-coordinate, and appropriate numerical

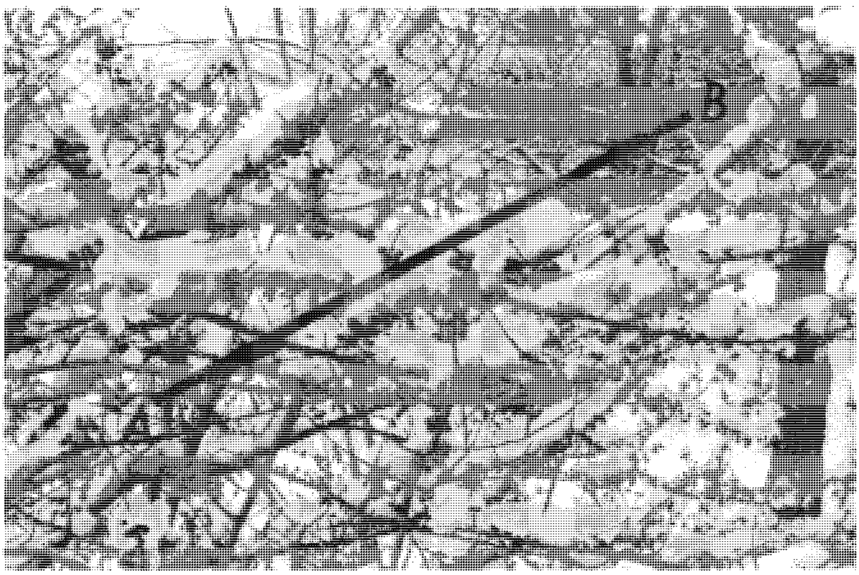


Figure 2.10. A nonrepetitive distribution of reflectance.

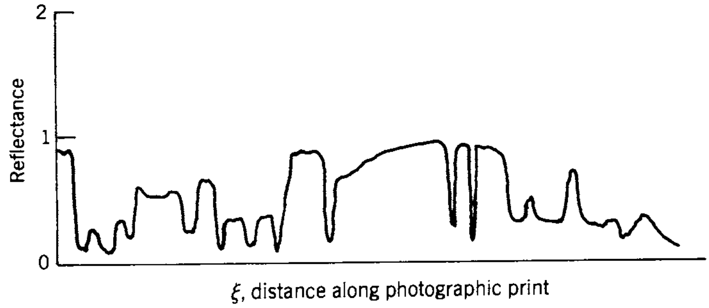


Figure 2.11. A plot approximating the reflectance along the line *AB* in Fig. 2.10.

methods are available to apply the Fourier transform to a sequence of such values.

A number of "fast" Fourier transform programs for computers and handheld calculators are now available. Some of these programs and the related methods for calculating spectra are discussed in later chapters on the measurement and computation of the optical transfer function.

If an image of a sinusoidal distribution, $f(x)$ of Eq. (2-3), is formed by an optical system, we say that the sinusoidal pattern of the object is *transferred* from object space to image space. In this sense, we define (in a following section) an *optical transfer function*, which indicates the efficiency of an optical system to transfer sinusoidal distributions at various spatial frequencies.

Application of the optical transfer function requires the assumption for the present that each sinusoidal component, usually of a flux density distribution, is transferred through the optical element or system of interest without distortion of the waveform, that is, the corresonding output component remains sinusoidal. In general, however, the distribution of flux density in the image and the relative amplitudes and phases of the components will be affected by the optical element or system. In fact, it is the relative amplitude and phase effects, as functions of spatial frequency, that give useful characterization of the optics in the application of the optical transfer function. Study of this characterization is the purpose of this book.

EXTENSION TO TWO DIMENSIONS

In the examples already given, care was taken to reduce the two-dimensional bar charts and picture to one dimension by specifying a path along which observations of reflection would be taken. This was done to simplify the discussion of concepts and mathematical operations. The present purpose is to restore the second dimension of the general problem in Fourier analysis. Extension to three dimensions will not be undertaken because all effects that we will treat with the optical transfer function will be in lateral directions; no effort will be made to discuss spatial frequencies in the direction of the optical axis.

When the Fourier transforms of Eqs. (2-5) and (2-6) are written in two dimensions rather than one, the following equations result:

$$F(\omega_x, \omega_y) = \iint\limits_{-\infty}^{+\infty} f(x, y) \exp[-i2\pi(\omega_x x + \omega_y y)]\, dx\, dy; \qquad (2\text{-}13)$$

$$f(x, y) = \iint\limits_{-\infty}^{+\infty} F(\omega_x, \omega_y) \exp[i2\pi(\omega_x x + \omega_y y)]\, d\omega_x\, d\omega_y. \qquad (2\text{-}14)$$

CONTRAST AND CONTRAST TRANSFER

The term *constrast* is found in discussions of optics, photography, vision, and photogrammetry, among other subjects, and usually refers to the range of recognizable shades of gray between the dimmest and brightest parts of an object or image. Definitions leading to numerical values are structured to be most helpful in the different contexts of the various fields. In this book, contrast C is defined by

$$C = (\mathcal{W}_{\mathrm{max}} - \mathcal{W}_{\mathrm{min}})/(\mathcal{W}_{\mathrm{max}} + \mathcal{W}_{\mathrm{min}}), \qquad (2\text{-}15)$$

where $\mathcal{W}_{\mathrm{max}}$ is the maximum flux density and $\mathcal{W}_{\mathrm{min}}$ is the minimum flux density in the "picture" or field. In the field described by Eq. (2-3), $\mathcal{W}_{\mathrm{max}}$ is $(a + b)$ and $\mathcal{W}_{\mathrm{min}}$ is $(a - b)$. When these values are substituted in Eq. (2-15),

$$C = b/a. \qquad (2\text{-}16)$$

Because of the resemblance of this definition to that of amplitude modulation in communication theory, C is sometimes referred to as *modulation contrast.*

If the sinusoidal pattern described by Eq. (2-3) is transferred through an optical system, our earlier assumption about such patterns indicates that the equation describing the image will have the same form as the object equation but with different values for the constants, which we will designate by primes (see Eq. (B-61)):

$$f(x') = a' + b' \cos(2\pi\omega_1' x_1' + \phi). \tag{2-17}$$

The added term, ϕ, is the phase difference between object and image caused by transmission through the optical system. For a given sinusoidal component, then, we can define contrast transfer T_C as

$$T_C = C'/C = (b'/a')/(b/a). \tag{2-18}$$

In this context, the unprimed values are said to denote *object space* and the primed values *image space.*

In a given optical system, T_C will be a function of the spatial frequency and is usually presented as a plot against a normalized frequency on the abscissa; this function is called the *modulation transfer function*, commonly shortened to *MTF.* When this magnitude function is abetted by a corresponding phase function, specifying the relative phase angle as a function of frequency, the combination is called the *optical transfer function*, shortened to *OTF.* The phase part of the combination is the *phase transfer function* (PTF). To express magnitude and phase simultaneously, the OTF is put in complex form:

$$\text{OTF}(\omega) = T(\omega) \exp\left[i\phi(\omega)\right], \tag{2-19}$$

in which T is the MTF and ϕ, the phase, is the PTF.

Figure 2.12 shows the MTF for a "perfect" lens, that is, a lens or lens

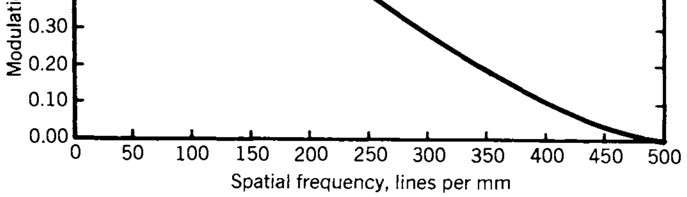


Figure 2.12. The MTF for a diffraction limited lens. The phase transfer function (not shown) is zero for all frequencies.

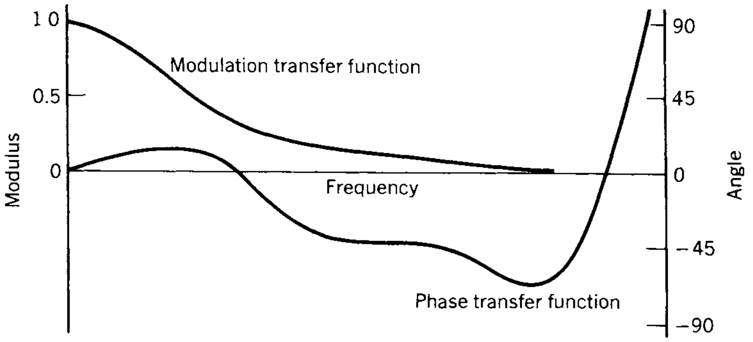


Figure 2.13. A measured OTF showing plots of both the MTF and the PTF [13].

system limited only by diffraction effects. The PTF, since it is zero at all frequencies, is not shown. Figure 2.13 shows the measured OTF of an actual lens. Both the MTF and the PTF are plotted as functions of spatial frequency. The phase transfer function is discussed again in Chapter 7 in connection with image quality.

DISTRIBUTIONS OF PHYSICAL QUANTITIES

Earlier in this chapter when functions $f(x)$ and $F(\omega)$ were used in connection with illustrated patterns, they were reflectance functions and could have been more explicitly designated $\rho(x)$ and $\rho(\omega)$. The reflected flux density could have been represented using the symbol M, and the incident flux density by the symbol H. The quantity represented by H is called incidance; and the quantity represented by M is called exitance. (Note: *incidance* and *exitance* as used here are correctly spelled with an a; both terms are often defined in treatments of photometry. We tacitly assume here that negligible radiant flux is emitted by the surface because the term exitance would otherwise correctly include any flux leaving the surface whether emitted or reflected.) Under a uniform distribution of flux density H_c we could have

$$M(x) = H_c\rho(x), \qquad M(\omega) = H_c\rho(\omega). \tag{2-20}$$

In the examples given earlier in this chapter, $f(x)$ and $F(\omega)$ could have represented these flux-density functions as $M(x)$ and $M(\omega)$, respectively, instead of the corresponding reflectances. By further adjustments of the physical circumstances of the examples, yet a third alternative significance could be given to the $f(x)$ and $F(\omega)$ functions. Assume that the various patterns had been displayed by optically projecting each pattern from a slide projector onto a screen of uniform reflectance ρ_c. Then the following relations would apply:

$$M(x) = H(x)\,\rho_c, \qquad M(\omega) = H(\omega)\,\rho_c. \tag{2-21}$$

Now $M(x)$ and $M(\omega)$ could represent exitance in lieu of the flux-density functions $f(x)$ and $F(\omega)$, respectively.

We will not use the terms (or their symbols) *incidance* and *exitance* further in this book, except where the context requires their use for clarity. When a symbol for flux density is desired and we need not differentiate between incidance and exitance, $\mathcal{W}$ is used.

POINT SOURCES

Point sources of light, which are easy to visualize and convenient in mathematical analyses, cannot quite be realized in practice. However, for practical purposes, a source of finite dimensions is a point source if its dimensions are negligibly small compared to other significant dimensions in the optical configuration. Some examples of small sources are the miniature, grain-of-wheat incandescent bulb; the low-power (2-W) concentrated zirconium arc, approximately 0.13 mm in diameter; and the 100-W, high-pressure, mercury arc (actually with a little argon), approximately 0.3 mm arc length. A source of appreciable size can be made to function as a point source by forming an image of the source on a small hole in a metallic screen. The illuminated hole then functions as the point source. Some sources that are used in this way are high-power, high-pressure gas arcs and the positive pole of a carbon arc.

Specifying the "strength" of the source in different directions requires the concept of *radiant intensity*, which, as defined earlier, is the radiant flux leaving a point source per unit solid angle given in watts per steradian. (Units: W-sr^{-1} and lm-sr^{-1}, respectively.) An ideal point source of light would radiate uniformly over 4π steradians—a complete sphere—but in most optical applications, a considerably smaller solid angle of near uniform flux distribution—constant intensity—meets requirements.

Point sources, either individually or as building blocks in continuous sources, are frequently considered as objects in optical image-forming systems, and certain tacit assumptions are usually made concerning the flux from such sources as it passes through a system. If no information is given concerning the loss properties of the system, it is usually assumed that all the flux entering the entrance pupil (defined in the next section) from the object source will leave the system through the exit pupil (100% efficiency). Actually, of course, the flux is diminished in transmission by absorption, by scattering, and sometimes by vignetting for off-axis sources.

For convenience, the total flux originating at the source and passing through the system is usually assigned a unit value in the appropriate units.

STOPS AND PUPILS

When light passes through an optical system from object to image, the bundle of rays from each point on the object successively goes through (or is reflected from) whatever optical elements that have been placed in the path. The elements may be refracting lenses, mirrors, apertures in opaque screens, mounts holding cross hairs, etc. Each element has a boundary beyond which the element no longer passes rays; thus, the boundary defines an effective size. These boundaries can be any shape, but in many optical designs they are circular with the optic axis passing through their centers. The first element encountered by the rays from a point on the object defines a bundle of rays, which passes on to the next element in the series. Whether the second element further limits the bundle depends on the relative sizes of the bundle and the second element. As this process is repeated through the whole optical system, it will usually turn out that the outer edge of one particular element determines the size of the bundle. This element is called the *aperture stop.* Although the aperture stop may be any kind of element, it is commonly an opening in an opaque screen when the optical system has been designed for optimum performance. Certain aberrations can be appreciably reduced by proper placement of this element along the optic axis. Also, different trade-offs in system performance can be realized by making the size of the aperture stop adjustable.

Two *pupils* are defined as particular optical images of the aperture stop. Generally other elements will both precede and follow the stop. The image formed in object space by the optical elements preceding the aperture stop is the *entrance pupil.* Similarly, the image formed in image space by the elements following the aperture stop is the *exit pupil.*

Most multielement optical systems can in principle be reduced to the simple configuration shown in Fig. 2.14. To attain the simple configuration for a system, all the elements preceding the aperture stop are replaced by a single element having all the attributes of the original combination; and similarly all the elements following the stop are replaced by their equivalent element. The resulting simplified system is called the *reduced optical system.* If a particular system is designed so that no elements precede the aperture stop, it is called the *front stop* and also functions as the entrance pupil. If, on the other hand,

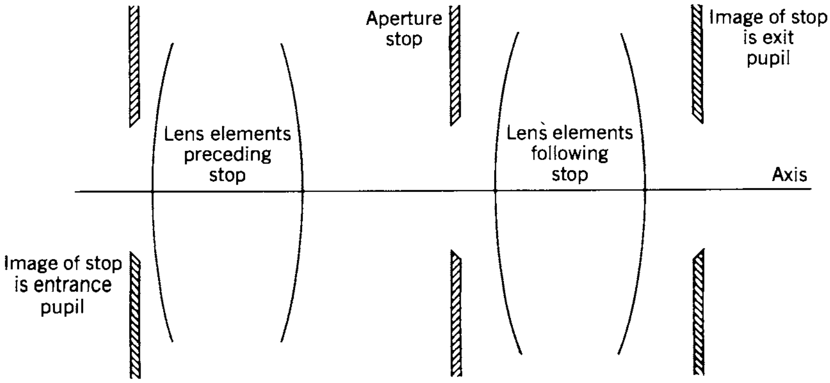

Aperture stop
Image of stop is exit pupil
Lens elements preceding stop
Lens elements following stop
Axis
Image of stop is entrance pupil

Figure 2.14. Reduced optical system.

the system has no elements following the aperture stop, it functions as the exit pupil.

When one's eye is positioned on axis at the object location looking into the optical system, the limiting opening seen is the entrance pupil. Looking the other way from on axis at the image position, one sees the exit pupil. Except for the two special cases already described where no optical elements intervene between the eye and the physical aperture stop, the observer is looking at an image, real or virtual, whose location in space is quite different from the location of the stop itself.

In the frequent reference to entrance and exit pupils, they are treated as if they were physical openings rather than the images that they usually are. This should cause no difficulty because they do indeed function as if they were physical openings limiting the light fluxes at their respective positions.

POINT SPREAD FUNCTIONS

The image of a point source can never be as precise as the point source itself. Several factors cause a spreading of the radiant energy reaching the image plane of the optical system. Dust particles on optical surfaces and scratches in these surfaces, foreign particles (air bubbles, for example) within lens material, irregularities on the edge of the aperture stop, diffraction of the light beam by the aperture stop, and aberrations (including defocusing) all cause the light to scatter and spread out about the point where the image would otherwise be formed.

A physical analogy useful in visualizing the distribution of flux density in the image of the point can be set up by making a contour map in which the isopleths of constant altitude represent lines of constant flux density. The typical hills-and-valleys plot would show a large hill in the center. Discussion of a variety of these plots will be deferred to later sections; our present discussion will confine itself to the plot shown in Fig. 2.15, which consists of a single hill, boss, or mound representing the flux distribution in the image plane. The mathematical function representing this boss, $\mathcal{W}(x, y)$, gives the flux density as a function of rectangular coordinates on the image plane, the usual origin being the location of the ideal image point. The function $\mathcal{W}(x, y)$ is called the *point spread function*.

Our references to *the image plane* in the preceding paragraphs and throughout the book usually imply the Gaussian image plane, which is defined in the next chapter. In actual optical systems, the choice of image plane tends to be arbitrary: Its optimum position usually depends upon what trade-offs we accept

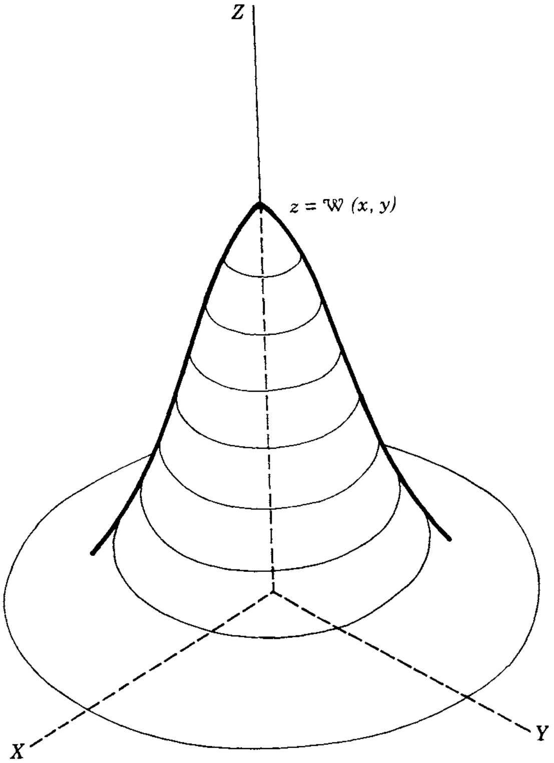


Figure 2.15. Light mound or boss showing isopleth lines.

in image quality. The rays emanating from a point source and intercepted by the entrance pupil of the optical system have a conical envelope (tangent to the outer rays everywhere), called the *caustic*, with the source at the vertex and the entrance pupil defining the base. As these rays leave the exit pupil, they converge so that the bundle again is the shape of a cone, in the ideal situation, with the image at the vertex. A plane perpendicular to the optic axis and passing through the vertex of the cone is the image plane. However, for the reasons already given for the spreading of the point image, the rays converge imperfectly and never do meet at a point before diverging again. The envelope rather necks down to an interval of small cross section and then expands indefinitely thereafter. The actual image plane would be located perpendicular to the optic axis and somewhere along the interval of small cross section. Once a position of optimum image quality has been found, "images" formed on other planes, parallel to but not coincident with the plane at the optimum position, are said to be *defocused*.

For purposes of discussing the flux-density distribution in the cross section of the image ray bundle, a hypothetical image plane serves as well as an actual one to define the cross section. An image formed on a hypothetical image plane is called an *aerial image*.

If we assume that the total light flux passing through the optical system from a point source is unity, the following integral applies to the point spread function:

$$\iint_{-\infty}^{+\infty} \mathcal{W}(x, y)\, dx\, dy = 1. \tag{2-22}$$

Finite limits of the order of the optic system dimensions can be applied to the integral of Eq. (2-22) with little loss of accuracy because typical spread functions decline to negligible values at relatively short distances from the ideal point image position.

A practical way of measuring the point spread function is first to photograph the point image by exposing a photographic film in the image plane. The density produced on the developed film then corresponds to the image flux density in phots, but usually not linearly. After photographic processing, the results are measured point by point with either a microdensitometer (for a transparency) or a microreflectometer (for a print). Conversion of the observations is then accomplished by referring to the appropriate characteristic curve, which gives log *exposure* versus photographic density. Exposure is defined as the product of incident flux density and the exposure time. From the field of flux-density data so obtained, contour lines of constant flux density can be plotted to show

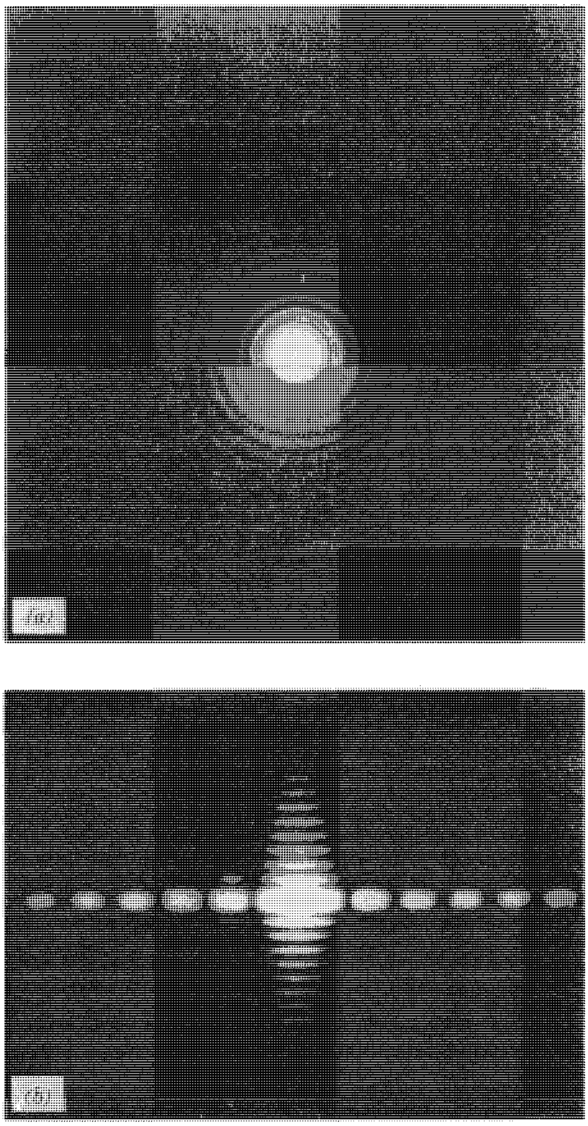


Figure 2.16. (*a*) Photograph of the point spread function produced by a diffraction-limited optical system of circular symmetry [14]. (*b*) Photograph of the point spread function produced by a diffraction-limited optical system having circular symmetry but a rectangular aperture stop. The direction of a short side of the stop is parallel to the long direction of the spread function pattern.

graphically the point spread function. Figure 2.16*a* is a photograph of the point spread function produced by a diffraction-limited optical system of circular symmetry. This spread function is known as the *Airy pattern*. Figure 2.16*b* is a photograph of the point spread function produced by a diffraction-limited optical system having a rectangular-shaped, rather than circular, aperture stop. The long dimension of the spread function is parallel to the short sides of the stop.

To accomplish a satisfactory measurement of the point spread function by the photographic method described, certain capabilities are required of the materials and instruments employed. The photographic film must provide continuously varying shades of gray corresponding to the range from the maximum flux density in the image to a level that is accepted as negligible. The resolution of the film must be adequate to record independently the flux density at adjacent points spaced at a distance corresponding to the minimum distance between contour lines in the final plot. The instrument probe that explores the photographic record must also be small enough to measure the record at the adjacent points independently.

When the graininess of the photographic film, the size of the probe, and other limitations in the measurement process compromise the resolution, the plotted spread function not only includes the effects of the optical system under measurement but incorporates the shortcomings of the measurement process as well. To get the true spread function of the optical system, one must "back out" the effects of the measuring process. To do this, it is convenient to assume that the measuring process has a point spread function of its own, which combines with the point spread function of the optical system to produce the plotted spread function. Theoretically, this kind of combination of two spread functions involves the convolution integral (see Appendix B, Eq. (B-31)),

$$\mathcal{W}(x', y') = \iint_{-\infty}^{+\infty} \mathcal{W}_0(x, y)\mathcal{W}_f(x' - x, y' - y)\, dx\, dy, \qquad (2\text{-}23)$$

where $\mathcal{W}_0(x, y)$ is the optical system spread function, and $\mathcal{W}_f(x, y)$ is the spread function of the measuring process.

Because grains in emulsions are oriented and distributed randomly, point spread functions for films are circularly symmetrical, which allows them to be written

$$\mathcal{W} = \mathcal{W}(r) = \mathcal{W}[(\mathrm{x}^2 + \mathrm{y}^2)^{1/2}]. \qquad (2\text{-}24)$$

On the other hand, optical point spread functions are generally not circularly symmetrical. This is illustrated in Figs. 2.17 and 2.18, which are asymmetric point spread functions for systems having circularly symmetric optics.

Departures from symmetry that cause the characteristics of point spread functions to be angularly dependent can often be identified with specific features and characteristics of the corresponding optical systems. For instance, the struts holding the secondary mirror in a reflecting telescope produce a characteristic star effect superposed on the diffraction pattern, the number of "points" corresponding to the number of struts (see Fig. 2.17*c*). Coma, astigmatism, and distortion each produces a characteristic asymmetry of the optical system point spread function.

Probing either an aerial image of a point or the photographic record of such an image to get numerical data about the point spread function can usually be considerably simplified by taking advantage of symmetry and other general characteristics of the flux-density distribution. For instance, a single scan along a straight line through the center of the symmetrical boss shown in Fig. 2.15 would produce a profile like the one in Fig. 2.19, which gives complete information about the distribution. For asymmetrical distributions, however, several scans in different directions are required to get sufficient data. A minimum for

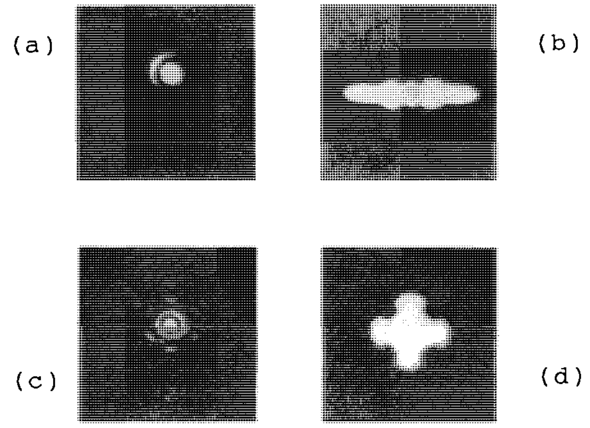


Figure 2.17. (*a*) Photograph of the point spread function produced by an optical system having circular symmetry and a small amount of coma [15]. (*b*) Photograph of the point spread function in a plane containing a focal line, produced by an optical system having circular symmetry and a small amount of astigmatism [15]. (*c*) Photograph of the point spread function produced by a diffraction-limited optical system having circular symmetry except for an obstruction of the incident beam by a secondary mirror held in place by three struts [16]. (*d*) Photograph of the point spread function produced by an optical system having a small amount of astigmatism. This is the spread function in an image plane midway between the two line images [15].

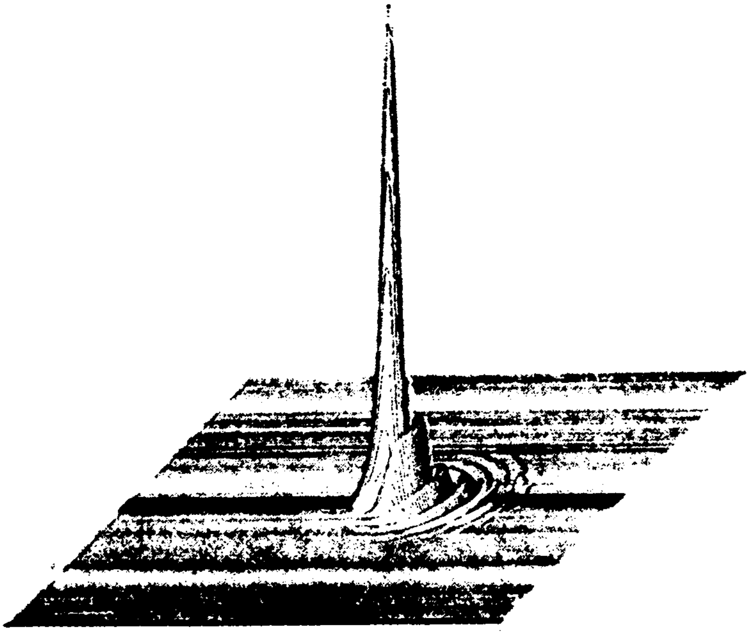


Figure 2.18. A computer-generated plot showing the point spread function produced by an optical system having circular symmetry and having a small amount of coma [17–19]. From "How Images Are Formed," F. D. Smith. Copyright © 1968 by Scientific American, Inc. All rights reserved.

such distributions would ordinarily be a scan parallel to the optical tangential plane and a second scan parallel to the sagittal plane.

As indicated in our earlier discussion about the microdensitometer or microreflectometer probe, any instrument probe to sense a point-by-point characteristic of a distribution must, to get good resolution, be as small as possible.

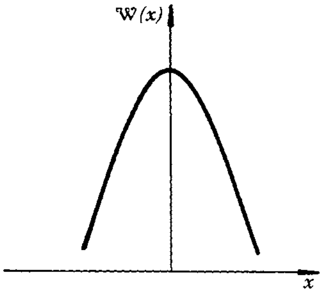


Figure 2.19. Profile through the boss of Fig. 2.15.

Because the sensitivity of a detector commonly declines as it is made smaller, a practical limit is reached because of noise in the sensing system. Also, the dimensions of a probe in an actual beam of light flux would have to be kept small to avoid disturbing the distribution being measured. To solve the sensitivity versus size problem in the probes for point image distributions, a line sensor rather than a point sensor is typically used, that is, the sensor area is elongated in one dimension to get sensitivity but is kept narrow in the dimension at right angles to the first to get resolution. This change, point to line sensor, requires, of course, a corresponding change in interpreting the data.

For point image applications, the long dimension of the probe sensor is made greater than the maximum image width; and the probe is swept across the image in a direction perpendicular to the long dimension. Simultaneously responding to the flux densities (or equivalent) at all points along the long dimension is in effect integrating the point spread function along a coordinate in the direction of this dimension. If the probe is swept in the y direction, the observed response would be related to the point spread function $\mathcal{W}(x, y)$:

$$\mathcal{W}'(y) = \int_{-\infty}^{+\infty} \mathcal{W}(x, y)\, dx. \tag{2-25}$$

$\mathcal{W}'(y)$ is called the *line spread function* and is discussed in a later section.

It is at once apparent that the same line spread function would result from a sweep in any direction across a symmetrical point image. However, the line spread function for an asymmetrical point image would depend on the direction of the sweep or scan. In general, several scans in different directions would be required to get sufficient information about an asymmetrical image.

SPREAD FUNCTIONS FOR SMALL ABERRATIONS

The point spread function has a particular significance because the optical transfer function, which is the central concern of this book, is derivable from it. (The derivation is discussed in Chapter 5.) As indicated earlier, the paragon of point spread functions, called the *Airy pattern*, would be produced by an ideal optical system having no aberrations and having a perfectly circular aperture stop. This pattern is the limit a designer strives for in successive corrections to an optical system. An example of a point spread function recorded for a highly corrected system is shown in Fig. 2.16*a*. Twenty diffraction rings appear on the original photographic film; but to show the outer rings, detail in the central disk had to be compromised by overexposure.

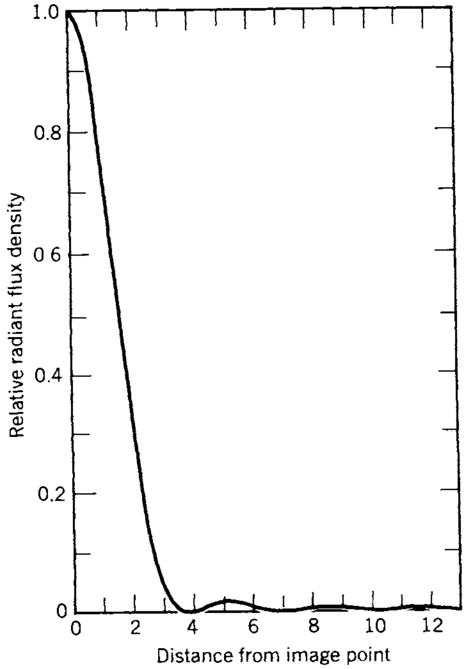


Figure 2.20. Profile through the spread function shown in Fig. 2.16*a*. This is also a plot of $\mathcal{W}_c(p')$ of Eq. (2-26) plotted as a function of r and normalized to $\mathcal{W}_{c0}$.

The profile along a radius of an Airy pattern is plotted in Fig. 2.20. Figure 2.21 is a computer-generated picture of the light mound corresponding to the photograph in Fig. 2-16*a* and the profile of Fig. 2.20. In the derivation of the expression for this curve, the only mechanism assumed for spreading the light is the diffraction at the circular aperture stop; hence, the Airy pattern is often referred to as the diffraction-limited spread function. The expression for the profile is (image space index of refraction assumed unity)

$$\mathcal{W}_c(p') = \mathcal{W}_{c0}[2\,J_1(u)]^2/u^2, \qquad (2\text{-}26)$$

where $\mathcal{W}_c$ is the flux density, and $\mathcal{W}_{c0}$ is its value at the center of the pattern; J_1 is the first order Bessel function of the first kind; and $u = 2\pi r \sin \alpha'/\lambda \cong 2\pi r \rho_m/(\lambda s')$. The symbols in the expressions for u are defined in Fig. 2.22, where $s' = R$ so that the numerical aperture is equal to $\sin \alpha' \cong \rho_m/s'$, a good approximation except for systems having high numerical apertures. The point source is assumed on the optic axis, so the Gaussian image of the source would also be on the axis. The general point p' in the pattern is at a distance r from

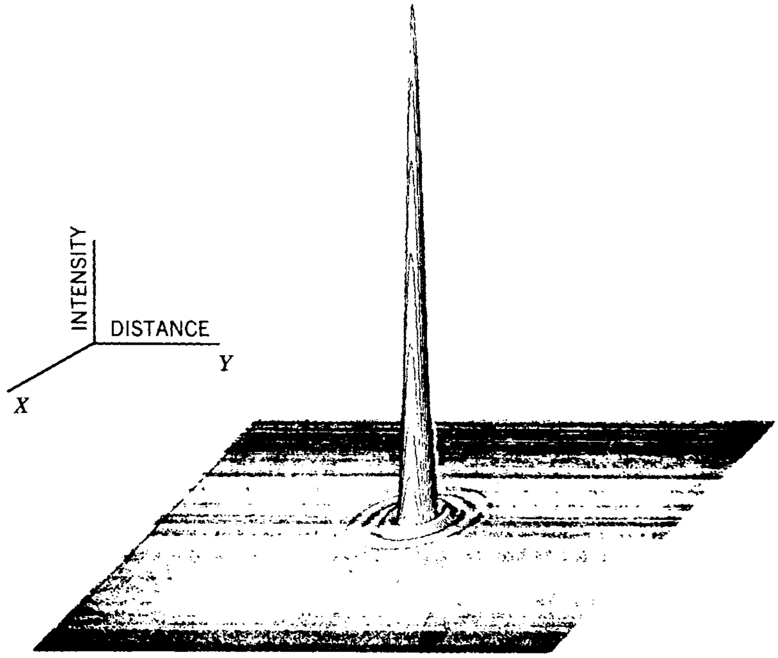


Figure 2.21. A computer-generated picture of the light mound or boss corresponding to the spread function shown in Fig. 2.16*a*, to the profile of Fig. 2.20, and to the distribution expressed by Eq. (2-26) [17-19]. From "How Images Are Formed," F. D. Smith. Copyright © 1968 by Scientific American, Inc. All rights reserved.

the intersection of the axis with the image plane. The source is assumed to radiate light in a narrow band at wavelength λ. Actual images of point sources depart from the Airy pattern represented by the expression in Eq. (2-24) for a number of reasons. For instance, the rays from an off-axis point object sometimes encounter a vignetted aperture, which is not circular; as a result, the spread function departs from the perfect Airy pattern. Departures can also be caused by slight defocusing or small aberrations, either for on-axis or off-axis point sources.

When the actual performance of an optical system differs only slightly from the ideal, the corresponding departure of the image of a point source from the Airy pattern is almost imperceptible. One might expect a very small aberration to cause a slight increase in the size of the spread function with little change in brightness. Actually, in most instances, the reverse is true. The apparent size of the central disk and the positions of the luminous rings surrounding it remain almost unaltered. The significant change is rather that the central disk declines

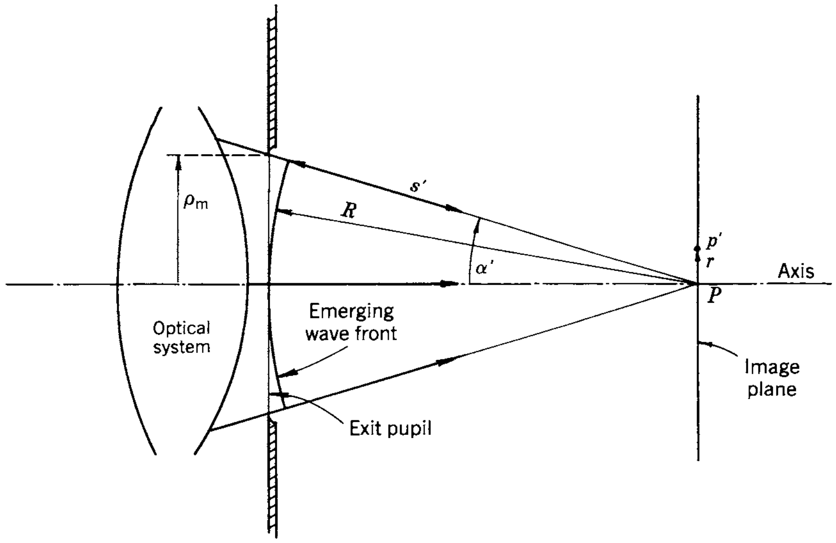


Figure 2.22. Schematic of the image space of a diffraction-limited optical system having circular symmetry and defining the parameters used in Eq. (2-26).

in brightness, and the lost light is distributed mostly to the innermost diffraction rings. The central disk is so bright compared with the rings that the redistribution is typically undetected by the human eye and barely recordable on photographic film. The flux density at the center of the perfect Airy disk is about 57 times the maximum flux density in the first ring. The central disk contains 84% of the total flux transmitted by the optical system from the point source. The remaining 16% is distributed throughout the ring pattern. If the system is defocused just enough to produce a quarter wavelength distortion in the wave front at the exit pupil (see Chapter 4 for further discussion)—the popular "Rayleigh criterion"—an additional 17% of the total flux is moved from the central disk to the ring system, but no appreciable change is made in the disk size or in the positions of the rings. The reduction of flux density in the disk by the factor $(84 - 17)/84 = 0.80$ is barely perceptible to the eye and would even be difficult to measure if a perfect and a defocused disk were compared in the laboratory. However, the flux in the ring pattern is doubled (from 16 to 33%). The extra flux may be distributed uniformly in the ring pattern, but is more often distributed asymmetrically to the first few rings as shown in Figures 2.17 and 2.18.

In the discussions of the optical transfer function in later chapters of this book, it will be shown that wave-front distortions equivalent to the defocusing

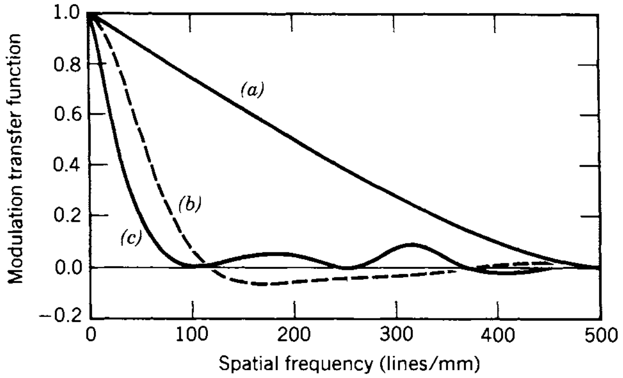


Figure 2.23. Modulation Transfer Function for an optical system having circular symmetry: (*a*) MTF for a diffraction-limited system with sin $\alpha' = 0.125$ and $\lambda = 500$ nm. (*b*) MTF for the same optical system but with a slight defocusing. (*c*) MTF for the same system but with 4 wavelengths maximum wave-front distortion corresponding to 4 wavelengths of spherical aberration [20].

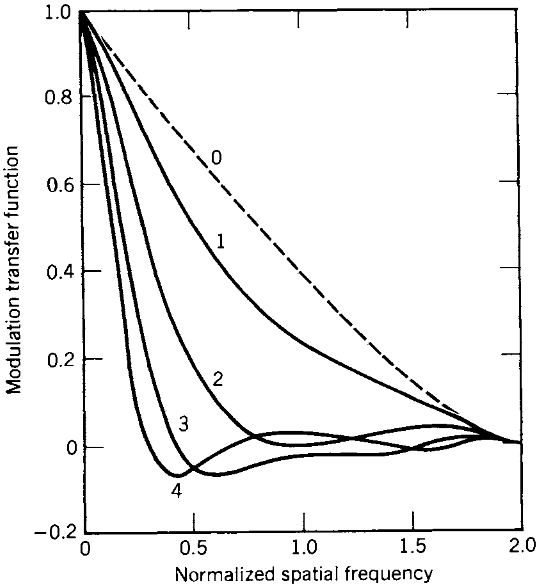


Figure 2.24. MTF for a system free of aberrations but having a defect of focus. The maximum wave-front distortion is n/π wavelengths where n is the number shown on each curve [21].

described produce highly perceptible changes in the OTF as can be seen in Fig. 2.23. These show up as loss of contrast at the high spatial frequencies, which corresponds to loss of contrast in the fine detail in the images of extended objects. Thus, one of the most valuable characteristics of the OTF is that the modulation transfer function (MTF) is significantly changed for small degradations in highly corrected optical systems whereas the corresponding changes in the spread function are hardly detectable.

When a system of rays originating at a single object point is constructed so that the rays are uniformly distributed over the entire entrance pupil, the plot of their consequent intersections with the image plane is called a *spot diagram*. Spot diagrams are sometimes studied by experienced optical designers to find what kinds of aberrations are present in a system. It is interesting to compare this method with the inspection of MTF curves to identify aberrations. From the comparisons in Figs. 2.23 and 2.24, as well as the discussion in previous paragraphs, one observes that as aberrations are introduced into a system, changes in the MTF curves will indicate their presence though the wave-front distortion remains quite small. After the aberrations are increased beyond the detection level, changes in the MTF curve begin to characterize the kind of aberration.

LINE SPREAD FUNCTIONS

The point spread function, which we have discussed at length, is the mathematical expression for the flux-density distribution in the image of a point source. Similarly, the *line spread function* is the mathematical expression for the flux-density distribution in the image of a line source.

A convenient way to approximate an ideal line source in the laboratory is to focus the image of a high-pressure mercury arc, or other intense source, onto a long, narrow aperture; the illuminated aperture then serves as the line source. A practical way to make the aperture or slit is to etch an engraved line through a thin sheet of metal. Another technique is to engrave a line through an opaque coating of aluminum on a glass substrate.

In our treatments of the line spread function, unless something is said to the contrary, we will assume that the source is completely incoherent over its surface. As with the corresponding plot of the point image, an isopleth plot of the line-source image results in a hill, boss, or light mound; but this time it is cylindrical as shown in Fig. 2.25. To acquire data for such a plot, a long, narrow probe may be used to scan the image. The axes of the probe and image are maintained parallel, and the probe is scanned perpendicular to these axes.

Because the line spread function is easier to measure, it is usually preferred over the point spread function in optical analysis. Several methods are used to

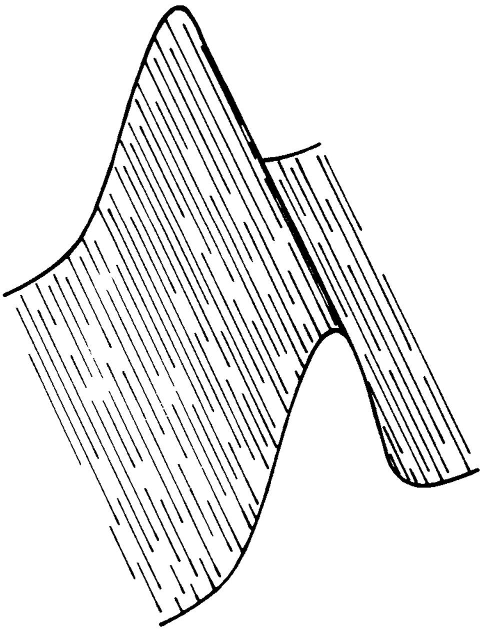


Figure 2.25. Isopleth plot of a line spread function.

obtain a line spread function. One has already been described in the discussion of Eq. (2-26). As suggested there, the indicated integration may be accomplished mathematically when the point spread function is known or experimentally when an aerial or a recorded image of a point source is available. Another method of obtaining a line spread function is to differentiate the edge trace, which is discussed in the next section.

The line spread function corresponding to the Airy pattern may be derived by first making the following substitutions in Eq. (2-26): $u = (u_x^2 + u_y^2)^{1/2}$, where $u_x = 2\pi\rho_m r_x/\lambda s'$ and $u_y = 2\pi\rho_m r_y/\lambda s'$. Comparison of the above expressions with the definition of u following Eq. (2-26) shows that r_x and r_y are the rectangular coordinates of the general point p' in the image plane. With these substitutions, Eq. (2-26) becomes

$$f_1(u_x, u_y) = \mathcal{W}_{c0}\, 2\left\{J_1\left[u_x^2 + u_y^2)^{1/2}\right]\right\}^2 / (u_x^2 + u_y^2). \qquad (2\text{-}27)$$

Conversion of this expression for the Airy pattern to the corresponding line spread function is done by integrating with respect to u_y. The result is known

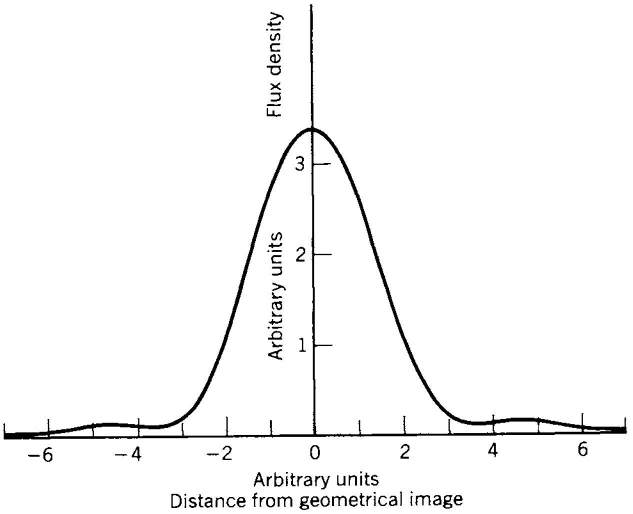


Figure 2.26. Line spread function for a diffraction-limited optical system corresponding to Eq. (2-28).

as a *Struve function of the first order*, $S_1(2u_x)$. The relations are

$$f_2(u_x) = \int_{-a}^{+a} f_1(u_x,\, u_y)\, du_y = 4\mathcal{W}_{c0}S_1(2u_x)/u_x^2, \qquad \text{(2-28)}$$

where $f_2(u_x)$ is the diffraction-limited line spread function (Fig. 2.26). The limit a is the value of $u_y = 2\pi\rho_m r_{ya}/\lambda s'$, where r_{ya} is an arbitrary distance from the Gaussian image point chosen so that $f_1(u_x,\, u_y)$ has negligible values at greater distances. In the mathematical literature, Struve functions are associated with Bessel functions; and tabulated values of Struve functions are usually found in tables of Bessel functions.

THE EDGE TRACE

Figure 2.27 shows schematically an optical configuration where a knife edge is placed in the object plane so that it intersects the optic axis. The resulting image is a plane half "white," half "black" separated by a straight line, which is the image of the knife edge. Close examination of the line, however, reveals that

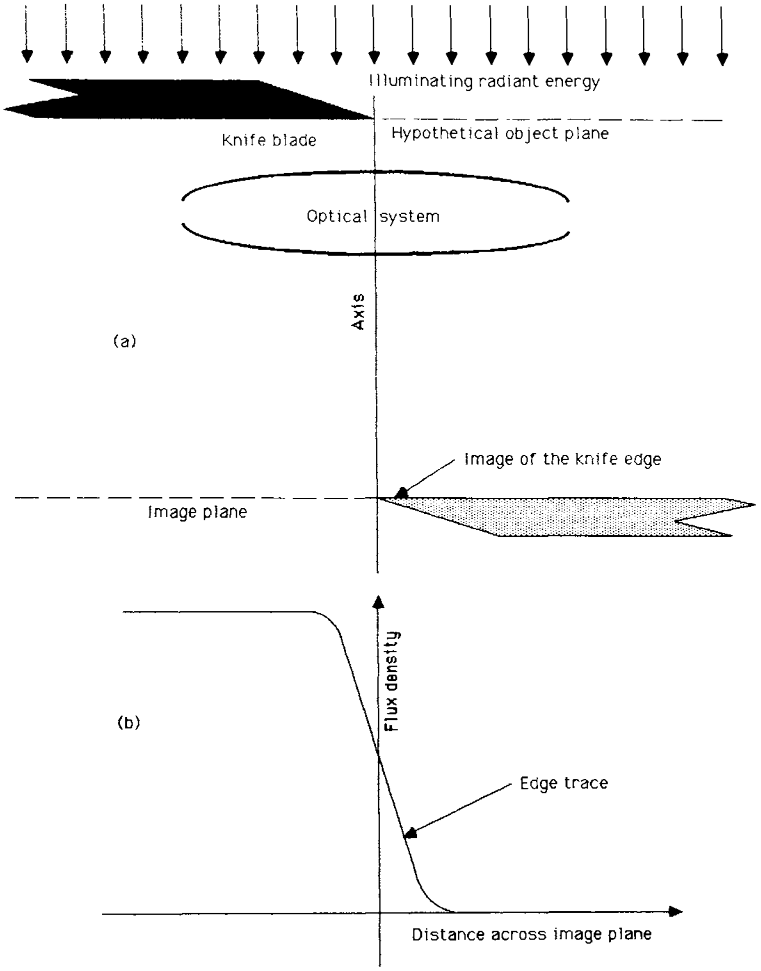


Figure 2.27. Schematic showing the formation of an edge trace.

the transition from white to black is not abrupt but occurs gradually because of diffraction, aberrations, and defocusing. When flux density is plotted against distance along a path perpendicular to the dividing "line," the *edge trace* shown in Fig. 2.28 results.

Inasmuch as both the edge trace and the line spread function are dependent upon the same characteristics of an optical system, it is not surprising that one can be transformed mathematically into the other. In fact, the edge trace is the

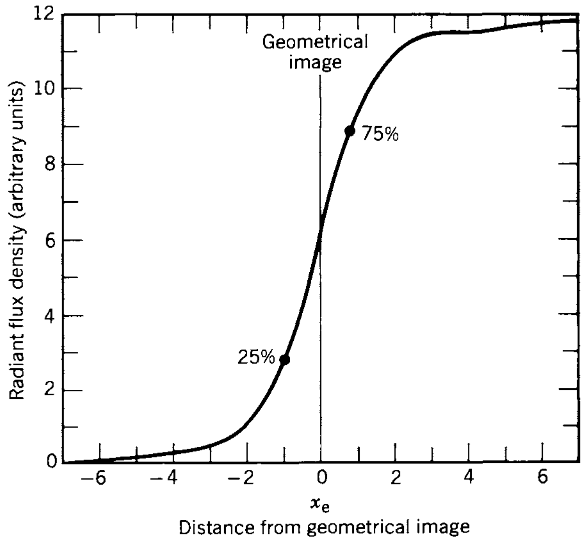


Figure 2.28. Edge trace for a diffraction-limited optical system with circular symmetry.

integral of the line spread function:

$$\mathcal{W}(x_e) = \int_{-\infty}^{x_e} \mathcal{W}(x')\, dx', \qquad (2\text{-}29)$$

where $\mathcal{W}(x_e)$ is the edge trace, and $\mathcal{W}(x')$ is the line spread function. Because of this relation, the edge trace is sometimes called the *accumulated line spread function*.

The edge trace is also the *convolution* (see Appendix B) of the line spread function and a step function:

$$\mathcal{W}(x_e) = \int_{-\infty}^{+\infty} U(x') f_2(x_e - x')\, dx', \qquad (2\text{-}30)$$

where $f_2(x')$ is the spread function and $U(x')$ is the step function, which is defined as

$$\begin{aligned} U(x') &= 1 \quad \text{when} \quad 0 \le x' < \infty, \\ U(x') &= 0 \quad \text{when} \quad -\infty < x' < 0. \end{aligned} \qquad (2\text{-}31)$$

ISOPLANATISM

An optical system is said to be *isoplanatic*, in the context of OTF discussions, if a translation of the object point in the object plane produces just a proportional translation of the entire point spread function in the image plane. That is, if an object point source at (x_1, y_1) produces the point spread function $f_1(x', y')$ in the image plane, a translation of the object point to $(x_1 + \Delta x, y_1 + \Delta y)$ produces the spread function $f_1(x' + m\,\Delta x, y' + m\,\Delta y)$ where m, a constant, is the magnification of the system and $m\,\Delta x = \Delta x'$ and $m\,\Delta y = \Delta y'$. Under these conditions, the translation of the object point causes no change in the spread function (except its location), and the shapes of the wave fronts approaching the image point are unchanged. As a result, the wave-front aberration (defined in the next chapter) also remains unchanged. The region that includes all isoplanatic points in the object plane (or the corresponding region in the image plane) is called an *isoplanatism patch*. The greatest isoplanatism patch containing a particular point *P* is called the *isoplanatism patch belonging to P*.

In practical optical systems, aberrations are known to change appreciably as an object point is moved considerable distances in the object plane. In the terminology just defined, this is to say that optical systems are not generally isoplanatic. Unfortunately the mathematical requirements for applying the Fourier transform to calculate the OTF reduce in part to requiring isoplanatism; in fact, the concept of a unique OTF for a given system requires that the system be isoplanatic. This problem is overcome by deciding first what variation of wave-front shape (aberrations) can be regarded as insignificant and then subdividing the object plane into regions, isoplanatism patches, so that the accepted variation is not exceeded within a given patch. Each patch will have its peculiar OTF. (Since wave-front shapes cannot be observed directly, some related optical characteristic (MTF, image distortion, etc.) will have to be observed to establish the boundaries of isoplanatism patches.)

LINEAR SUPERPOSITION

In addition to isoplanatism, the mathematical conditions for applying the Fourier transform to calculate the OTF include *linear superposition*. Mathematically, linear superposition can be illustrated by assuming the "strengths" of two sources to be $I_1(x_1, y_1)$ and $I_2(x_1 + \Delta x, y_1 + \Delta y)$, which produce $f_1(x', y')$ and $f_2(x' + \Delta x', y' + \Delta y')$ in the image plane, respectively, when each source is applied by itself. If the two sources are applied simultaneously to produce $f_3(x', y')$ in the image plane, linear superposition means that

$$f_3(x', y') = f_1(x', y') + f_2(x' + \Delta x', y' + \Delta y'). \tag{2-32}$$

Furthermore, if the two source strengths are modified by the two arbitrary factors A and B so that they become $AI_1(x_1, y_1)$ and $BI_1(x_1 + \Delta x, y_1 + \Delta y)$, they will individually produce $Af_1(x', y')$ and $Bf_2(x' + \Delta x', y' + \Delta y')$ in the image plane. When the modified sources are applied simultaneously, the total effect in the image plane will be $Af_1(x', y') + Bf_2(x' + \Delta x', y' + \Delta y')$.

The natures of the source "strengths" and the resulting effects (f_1, f_2, and f_3) in the image plane were purposely left ambiguous in the discussion above because both the sources and the images could be characterized in different ways—all legitimate as long as linear superposition holds.

As indicated by the mathematical discussion, linear superposition means that the effect in the image plane will be directly proportional to the strength of the source. When two or more sources are applied simultaneously, the combined effect at each point in the image plane will be the sum of the individual effects observed when the various sources are applied one at a time.

Linear superposition breaks down when the phase in the beam from one source depends on the phase in the beam from a second source. Such dependence is called *coherence*.

COHERENCE

Electromagnetic energy in a beam of light exists in discrete packets called photons. Each photon is thought to be accompanied by a short electromagnetic wave train of increasing and then decreasing amplitude. Unless the photon-generating mechanism imposes a special discipline on the photons with respect to their placement in both space and time, as in a laser, the phase relations between photons are random; so no fixed-phase dependence usually exists between any two places in an ordinary beam. Without a stable phase relation, the beam is said to be *incoherent*. However, if the beam from a source, especially a point source, is divided and the two separate beams are then led over different paths and recombined appropriately, the phase in one beam may have a constant difference with the phase of the other. This dependency, which is called *coherence*, cannot be observed directly by comparing the alternations in the two beams; no detector exists with a short enough time constant. When the beams are superposed correctly and there is a sufficient degree of coherence, in-phase regions of the composite beam will be characterized by above average flux density, out-of-phase regions by below average flux density. On the image plane, these in-phase and out-of-phase regions show up as light and dark fringes, which are manifestations of constructive and destructive interference, respectively, in the superposed beams. The existence of observable fringes implies a partial fixed-phase dependence, and also implies at least a degree of coherence. The process of imposing certain characteristics on a beam of light—such as collimating, narrow-band filtering, and passing through small apertures, even through microscope objectives—often introduces a degree of coherence into an otherwise incoherent beam. Many of these creations of coherence are unintentional. Whatever the reason that coherence is produced in the light reaching the image plane, coherence will tend to violate the conditions necessary for applying the Fourier transfer function to calculate the OTF.

Coherence is discussed from a number of points of view in Refs. 4–12. The accuracy of the various procedures must be evaluated by the OTF practitioner from a general comprehension of coherence principles and a knowledge of the particular laboratory setup. In certain procedures for measuring the OTF, a degree of partial coherence can cause the results of an experiment to be inaccurate. These problems with coherence are discussed with the appropriate context in later sections of this book. (See Ref. 23.)

REFERENCES

1. R. D. Stuart, *An introduction to Fourier Analysis*. Methuen, London, 1961.
2. A. Papoulis, *The Fourier Integral and Its Applications*. McGraw-Hill, New York, 1962.
3. M. V. Berry and D. A. Greenwood, On the Ubiquity of the Sine Wave. *Am. J. Phys.* **43,** 91 (1975).
4. M. J. Beran and G. B. Parent, Jr., *Theory of Partial Coherence*. Society of Photo-Optical Instrumentation Engineers, P.O. 10, Bellingham, WA. (First published by Prentice-Hall, 1964.)
5. L. Mandel and E. Wolf, Coherence Properties of Optical Fields. *Rev. Mod. Phys.* **37,** 231 (1965).
6. M. Born and E. Wolf, *Principles of Optics*. Pergamon, Oxford, 1965, Chapter 10.
7. E. L. O'Neill, *Introduction to Statistical Optics*. Addison-Wesley, Reading, MA, 1963, Chapter 8.
8. M. Françon, *Optical Interferometry*. Academic, New York, 1966, pp. 1–32, 162–182.
9. C. S. Williams and O. A. Becklund, *Optics: A Short Course for Engineers and Scientists*. Wiley-Interscience, New York, 1972, pp. 18–21, 261–265.
10. S. G. Lipson and H. Lipson, *Optical Physics*. Cambridge Univ. Press, Cambridge, 1969, Chapter 8.
11. W. T. Cathey, *Optical Information Processing and Holography*. Wiley, New York, 1974, pp. 74–93, 250–252.
12. H. J. Caulfield (Ed.), *Coherent Optical Processing. SPIE Proc.* **49** (1974). (Please see the note following Ref. 2 of Chapter 1.)

13. H. Lavin and M. Quick, The OTF in Electro-Optical Imaging Systems. *SPIE Proc.* **46,** 279 (1974). (Please see the note following Ref. 2 of Chapter 1.)
14. F. S. Harris, Jr., ''Light Diffraction Patterns,'' *Appl. Optics,* **3,** 909 (1964).
15. K. Nienhuis and B. R. A. Nijboer, The Diffraction Theory of Optical Aberrations. *Physica* **14,** 590 (1948).
16. E. Everhart and J. W. Kantorski, Diffraction Patterns Produced by Obstructions in Reflecting Telescopes of Modest Size. *Astron. J.* **64,** 455 (1959).
17. F. D. Smith, How Images Are Formed. *Sci. Am.* (Sept. 1968, 97).
18. R. R. Shannon, Some Recent Advances in the Specification and Assessment of Optical Images. In *Optical Instruments and Techniques*, J. Home Dickson (Ed.). Oriel Press, Stocksfield, Northumberland, UK. (Proceedings of a conference held at the University of Reading, 4–9 July 1969.)
19. S. H. Lerman, Application of the Fast Fourier Transform to the Calculation of the Optical Transfer Function, *SPIE Proc.* **13,** 51 (1969). (Please see the note following Ref. 2 of Chapter 1.)
20. E. H. Linfoot, *Fourier Methods in Optical Image Evaluation*. Focal Press, London, 1964.
21. H. H. Hopkins, The Frequency Response of a Defocused Optical System. *Proc. R. Soc. London Ser. A* **331,** 91 (1955).
22. S. A. Rodionov, Isoplanatism in Arbitrary Optical Systems. *Opt. Spectrosc.* **46,** 315 (1979).
23. H. H. Hopkins, Image Formation with Coherent and Partially Coherent Light. *Photo. Sci. Eng.* **21,** 114 (1977).

3

Notation and Coordinates

INTRODUCTION

There seems to be no clear consensus on the choice of symbols and terms in the field of optics. Even the choice of which direction of a directed quantity such as a ray is to be positive or negative is not standard. Many attempts have been made to find a common language that all would accept, but it seems that each textbook and each lens computer program still has at least a few unique elements in its language. Because there is no generally accepted standard in the optics literature, we often made an independent choice as to what is used in this book. Since more significant papers about the OTF have probably been published in England than in our own country, the selection of symbols and terms becomes especially difficult when the notation accepted in England, for example, is not commonly found in the United States.

Because of the existing muddle, the sometimes irksome and often difficult to visualize subjects of notation, sign convention, and coordinate systems, which constitute the special language of the optics of the optical transfer function, are treated in some detail in this chapter. Our main purpose is to introduce the reader to the usage in this book and, we hope, quite generally in the field of optics.

For historical and tutorial reasons, we begin with the conventional way of presenting cross sections of optical elements and optical systems, including ray paths through the elements. We briefly explain cardinal points and paraxial notation. We then take up the reduced and canonical coordinates of Hopkins, which, when they are changed back to real-space coordinates, become the more familiar conventional notation. Finally, while we are discussing coordinates, we take up the relations involved in a shift of image plane and with magnification in anamorphic imaging.

Our choices begin with the selection of terms and symbols. As already indicated, we try to use terms and symbols that have been used most often and most effectively by other writers. We pay particular attention to the American National Standard, ANSI PH3.57-1978, which is mentioned in Chapter 8, Ref. 5.

An optical system, in our discussions, is a sequence of transparent media, usually isotropic, each separated from the next by a smooth, polished surface. It is often useful to think of two successive surfaces and the optical medium between them as a unit called a *simple lens* or *element*.

Like other three-dimensional geometric systems, the optical system is rarely treated mathematically until certain simplifying conditions are imposed. Unless something is said to the contrary, all surfaces are surfaces of revolution on a common axis; in fact, the further assumption that the surfaces are spherical can often be made. Of course, a plane surface can be regarded as a spherical surface of infinite radius. The *vertex* of a surface is its intersection with its axis.

If only rays that are both close to the axis and almost parallel to the axis are to be considered, the system is said to be *paraxial*, *Gaussian*, or *ideal*. Under this assumption, an angle α made by the ray with either the axis or a perpendicular to an optical surface will be small so that the following approximations can be made (α in radians):

$$\sin\alpha = \alpha, \qquad \tan\alpha = \alpha, \qquad \cos\alpha = 1. \tag{3-1}$$

The approximation for each trigonometric function is recognized as the first term in the series expansion for the function, the accuracy of the approximation depending on how close the value of the first term is to the sum of the infinite series.

When the paraxial assumption gives acceptable results, nonspherical surfaces can usually be adequately represented by spherical approximations [1, p. 17]. The wave front approaching an image point in paraxial calculations is spherical, so the paraxial system is free of aberrations. Thus, only one optical transfer function applies (Fig. 2.12) under this simplifying assumption.

Treating an optical system as a Gaussian (paraxial) system is useful for defining focal length and power (optical sense) and for locating the *cardinal points*. Included in the cardinal points are the *focal points*, *principal points*, *object point*, *image point*, and *nodal points*. A plane through any one of these points and perpendicular to the axis has the same name as the point, as, for example, the Gaussian *image plane*, which is perpendicular to the axis at the Gaussian image point.

Discussions of *relative aperture*, *entrance pupil*, *exit pupil*, *aperture stop*, *vignetting*, and *field* (of view) *coverage* involve rays that are relatively remote from the axis and, therefore, cannot be conducted in terms of the Gaussian system. Instead, a more comprehensive system (fewer simplifying assumptions) is needed. This is also true when the optical transfer function is to be used in optical analysis, evaluation, and design because moderately large fields of view and entrance pupils characterize optical systems in which questions of aberration residuals occur.

SIGN AND NOMENCLATURE CONVENTIONS

Because assignment of algebraic signs is largely arbitrary in setting up optical problems and because no common convention has been adopted regarding what is called positive and negative, explicit rules have to be set up and meticulously followed concerning signs. The rules we use are included in the following definitions and conventions:

1. Light proceeds from left to right in the optical diagrams unless otherwise stated.
2. A distance measured in the direction that light is proceeding (usually from left to right) is positive.
3. A distance is always measured from a refracting surface or from a principal plane (which will be defined later).
4. The vertex of a refracting surface is its intersection point with the axis of symmetry.
5. A radius of curvature is positive if the direction from the vertex of a surface to its center of curvature is from left to right.
6. The surfaces are numbered in the order in which light passes through them.
7. Subscript numbers indicate the surface at which refraction is taking place.
8. Whenever a distinction must be made, a prime indicates that the quantity applies after the ray has been refracted at a surface.
9. A reflecting surface requires the use of a negative index of refraction for the medium following the surface to account for the change in direction or "folding" of the optical system.
10. *Object space* is that region containing the rays before they enter the lens or optical system.
11. *Image space* is that region containing the rays after they have passed through the lens or system.

CARDINAL POINTS

Associated with each lens and with each combination of lenses are certain significant points, that are useful in analyzing optical systems, called the *cardinal points*. We postulate the existence of these points and planes related to them and define them as follows:

1. Two unit *conjugate planes* (conjugate here meaning that there is an object–image relation between them), called the *first* and *second principal planes*, *unit planes*, or *Gauss planes*, are perpendicular to the optic axis; their points of intersection with the optic axis are called the *first* and *second principal points*, respectively.
2. The *first* and *second focal planes* (which can be called *infinite conjugates*) are perpendicular to the optic axis; their intersections with the optic axis are called the *first* and *second focal points*, respectively.
3. Parallel rays in object space that are incident upon the first principal plane and that pass through the system will reach a common point in the second focal plane; parallel rays in image space, having passed through the optical system, will have passed through a common point in the first focal plane.
4. A ray that is incident on the first principal plane at a distance h from the optic axis and that passes through the system will leave the second principal plane at a distance $h' = h$ from the axis.
5. There are two points on the optic axis called *nodal points* such that a ray passing through the first nodal point will also pass through the second nodal point and its direction in image space will be parallel to its direction in object space.

The principal points, focal points, and nodal points together are called the cardinal points for the system. When the positions of these points are known, the location of the image of an object and its magnification can be determined by a simple ray-tracking procedure. Since a centered system is postulated, only a two-dimensional plot is needed.

If the refractive indices of the media in object space and image space are the same, the nodal points will coincide with the principal points. In the following discussions we assume this to be true. In most practical optical systems, the medium is air, which has a refractive index of approximately unity. An important exception is the human eye.

PARAXIAL NOTATION

Although the paraxial system is too elementary for significant OTF developments, its notation is a convenient stepping stone for a more complex coordinate system and notation appropriate for OTF analysis.

Figure 3.1 shows the symbols and the coordinates for a paraxial system; for illustration the angles (α, α', and θ) have been drawn larger than would be permissible for the paraxial approximations discussed earlier. As shown in the

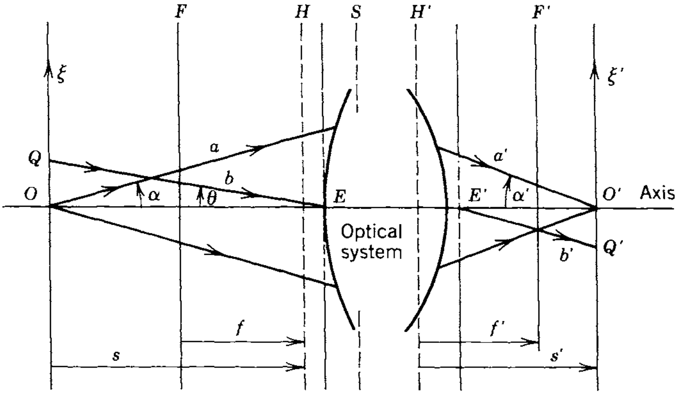


Figure 3.1. Symbols and parameters for the paraxial optical system. The size of each angle is exaggerated for clarity.

figure, unprimed symbols are usually reserved for object space and primed symbols for image space. The diagram, which is not meant to represent any actual system, shows the two spaces nicely separated so that unprimed symbols appear on the left and primed on the right. For some systems the object and image spaces overlap so that F' and H' could occur to the left of F and H. The paraxial symbols on the diagram are defined as follows:

O, O'	On-axis object and image points
Q, Q'	Off-axis object and image points
F, F'	First and second focal planes
H, H'	First and second principal planes
S,	Aperture stop
E, E'	Intersections of entrance and exit pupil planes with the axis
a, a'	A rim or edge ray
n, n'	Indices of refraction in media of object and image spaces
b, b'	Chief or principal ray (from an off-axis point)
θ	Field of view half angle
$n \sin \alpha$, $n' \sin \alpha'$	Object and image space numerical apertures
(ξ, η), (ξ', η')	Rectangular, real space coordinates in object plane and image plane with origins at O and O' (η and η' not shown because they are measured perpendicular to the plane of the diagram)

(x, y), (x', y') Rectangular, real space coordinates in entrance pupil and exit pupil with origins at E and E' (not shown in diagrams)

f, f' First and second focal distances, which are equal when the media of object space and image space are the same or have the same index of refraction

s, s' Object and image distances measured from the principal planes, first and second respectively

As indicated earlier, O and O' determine the positions of the object and image planes.

NEED FOR SPECIAL COORDINATES

In the early work on the optical transfer function, real-space coordinates, such as the system described for paraxial optics, led to a number of distracting difficulties. One particular improvement was to define points relative to coordinates on a wave front rather than relative to pupil plane coordinates. H. H. Hopkins was the pioneer in this work, and he developed what are known as *reduced and canonical coordinates* [2–6]. Before these coordinates are defined, we will discuss some of the difficulties that led to their use.

The word *reduced* means here that the effects of magnification are removed from analysis of the image. Without this adjustment, a spatial frequency ω' in image space is related to the corresponding spatial frequency ω in object space by the magnification m of the optical system according to

$$\omega' = \omega/m. \tag{3-2}$$

For example, when the magnification is 10, 20 cycles/mm in object space becomes 20 cycles for each 10 mm or 2 cycles/mm in image space. However, in reduced coordinates, magnification is always unity, so the reduced frequency does not change from object to image space.

In real-space coordinates for nonparaxial systems, application of Eq. (3-2) is no longer simple when m is not constant. For instance, the aberration called distortion causes m to depend on the object point distance OQ (Fig. 3.1). Furthermore, at off-axis points many optical systems tend to be anamorphotic (having different magnification in each of two perpendicular meridians), so m then depends on the direction of the object point distances as well as on its magnitude. In a test situation, this means that the apparent magnification of an optical system depends on the azimuthal orientation of a line-pair test chart. Reduced and canonical coordinates remove the need for explicit attention to these and other effects of aberrations and vignetting (restrictive action of the edge of the aperture for rays that are not axial).

In real-space coordinates, certain diffraction integrals and certain equations for wave-front distortion due to focal plane shift become indeterminant when either the exit pupil or the image plane is at infinity. These indeterminancies do not occur in canonical coordinates.

Because of the relative characteristics illustrated in the previous paragraphs, computer programs for calculating the optical transfer function are simpler and more efficient when expressed in reduced and canonical coordinates rather than in the more conventional real-space coordinates.

WAVE-FRONT ABERRATION

Figure 3.2 shows a wave front diverging from the off-axis object point $\bar{Q}$ toward the entrance pupil of an optical system. Because of the physical nature of wave propagation, the wave front is spherical and centered on $\bar{Q}$. The rays related to the wave front are normal to the wave front and pointed radially away from $\bar{Q}$. Figure 3.3 shows the wave front after it has emerged from the exit pupil of the optical system. If the system is assumed free of aberrations, the wave front will be spherical and centered on an image point $\bar{Q}'$, upon which it converges. The related rays are again normal to the wave front but this time pointed radially inward toward $\bar{Q}'$.

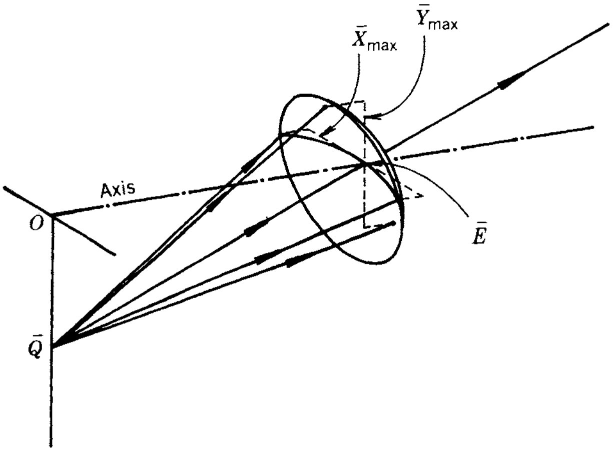


Figure 3.2. Symbols and parameters for an off-axis object point.

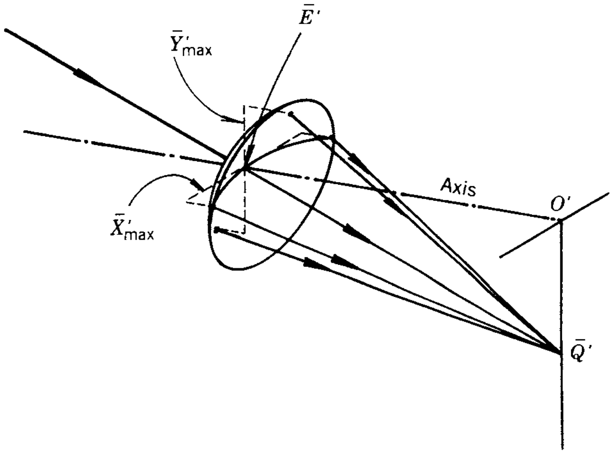


Figure 3.3. Image space symbols and parameters corresponding to object space designations of Fig. 3.2.

Because of aberrations, the emerging wave front is generally not spherical, nor does it necessarily tend to converge on the designated image point (defined later). To specify the nature of the actual wave front, it is compared with a spherical wave front of the aberration-free system; the spherical wave front, which may be hypothetical, is called either the *reference sphere* or the *pupil sphere*. The distance along a ray from the reference sphere to the actual wave front is the *wave-front aberration* or *wave-front distortion*; its numerical value is the optical path length obtained by multiplying the geometrical distance along the ray by the refractive index of the medium. In Fig. 3.4, n' times the distance between $\bar{B}'$ and D' is the amount of the wave-front aberration at $\bar{B}'$. The aberration is positive when measured from reference sphere to wave front in the direction of the ray; hence, the aberration depicted in Fig. 3.4 would be positive. (The terms wave-front aberration and wave-front distortion are often shortened to *wave aberration* and *wave distortion*.)

Since wave-front aberration, represented by the symbol W, varies with position in the wave front, it is a function of the coordinates on the reference sphere:

$$W = W(\bar{B}'). \tag{3-3}$$

Optical designers can make an estimate of the actual wave-front shape in the vicinity of the reference sphere when doing ray tracing. (See the section on

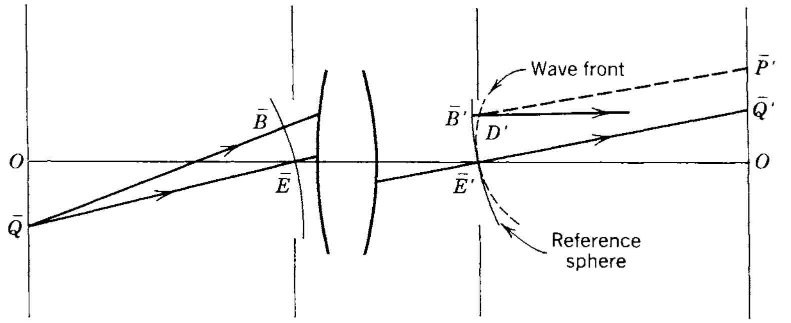


Figure 3.4. Object and image spaces showing tangential plane symbols and an image space reference sphere.

transfer equations for a discussion of ray tracing.) The approach is to make a number of exploratory optical path length measurements along rays from the off-axis point $\bar{Q}$ (Fig. 3.4) to the reference sphere, which is centered on $\bar{Q}'$ and passes through $\bar{E}'$. (The location of $\bar{E}'$ is discussed in the next section, which describes nonparaxial notation.) As indicated earlier, *optical* path length D is the product of the refractive index over each segment and the geometric length l of the segment:

$$D = \sum n(l)\,\Delta l \quad \text{or} \quad D = \int_{Q}^{\bar{B}'} n(l)\,dl, \tag{3-4}$$

which is the optical length of the ray $\overline{QBB}'$ where $n(l)$ is the index of refraction at the element dl of geometrical path length. The optical lengths for various points $\bar{B}$, resulting in various corresponding points $\bar{B}'$, should be equal if there are no aberrations. In general, however, the various optical path lengths so calculated are not equal; the difference between each length and the length for the ray $\overline{QEE}'$ is a measure of the aberrations present in the system. Since the quality of the point image is determined by the whole wave front, the optical designer has to make enough exploratory calculations of ray length to get a feel for the discrepancies between the actual wave front and the reference sphere.

A defective image of a point source is caused by a combination of interrelated wave-front and ray characteristics: failure of an emergent wave front to be spherical, failure of all rays (normals to the wave front) to point toward the image point, and failure of all points on the wave front to arrive at the image plane exactly in phase.

By definition, all points of a wave front originating from a point source are at the same optical distance from the source and are all in phase. So far, in our discussion, wave-front aberration has been indicated as the optical ray length increment between the reference sphere and the wave front. However, the effect, at the image point, of this discrepancy is more simply related to the phase increment than to the length increment. Phase increments can be given either in radians or in wavelengths in free space. A wavelength corresponds to 2π radians or one cycle of phase. To calculate wave-front aberration phase advance from the wave-front aberration $W(\bar{B}')$, $W(\bar{B}')$ is multiplied by k where $k = 2\pi/\lambda$ and λ is the wavelength in free space. At $\bar{B}'$ in Fig. 3.4, the phase advance is $2\pi n'(\bar{B}'\ D')/\lambda$ radians or $n'(\bar{B}'\ D')/\lambda$ wavelengths; of course, if D' were to the left of $\bar{B}'$ ($\bar{B}'\ D'$ negative), the aberration would be described as a phase lag.

In an aberration-free system where the emerging wave front coincides with the reference sphere, all rays arrive precisely in phase at the image point, and a maximum exchange of energy takes place from the wave front to the image point. Any aberration will disturb this condition; energy from all parts of the wave front will not arrive in phase at the image point (which is equivalent to saying that all rays will not be directed at the image point), and the exchange of energy from the wave front to the image point is diminished. It is conceivable that the wave front could be spherical and yet not be centered on the designated image point so that the maximum energy would be delivered to another point. However, if the wave front is nonspherical, something less than the maximum possible energy exchange will take place at any point because of the consequent spreading out of the energy.

NONPARAXIAL NOTATION

In Fig. 3.2, $\bar{Q}$ is an off-axis object point outside the paraxial region, and O is the axial object point in the same object plane as $\bar{Q}$, so $O\bar{Q}$ is the distance of $\bar{Q}$ from the axis. In nonparaxial notation, the bar on symbols like $\bar{Q}$ indicates that an off-axis object or image point is involved. As in notation discussed earlier, primes usually denote image space.

The plane that includes the axis and the line $O\bar{Q}$ is called the *tangential plane* in optics. Two-dimensional diagrams like Fig. 3.5, which include $\bar{Q}$ and the axis, are drawn in the tangential plane. A cone of rays originates at $\bar{Q}$ and passes through the entrance pupil of the optical system. Only a "fan" of rays originating at $\bar{Q}$ can be shown on the two-dimensional diagrams. In Fig. 3.5, the fan is represented by two edge rays and a central ray. The particular wave front shown was chosen so that the two edge rays intersect the wave front at equal distances from the axis ($|+\bar{Y}_{\max}| = |-\bar{Y}_{\max}|$). The chosen wave front,

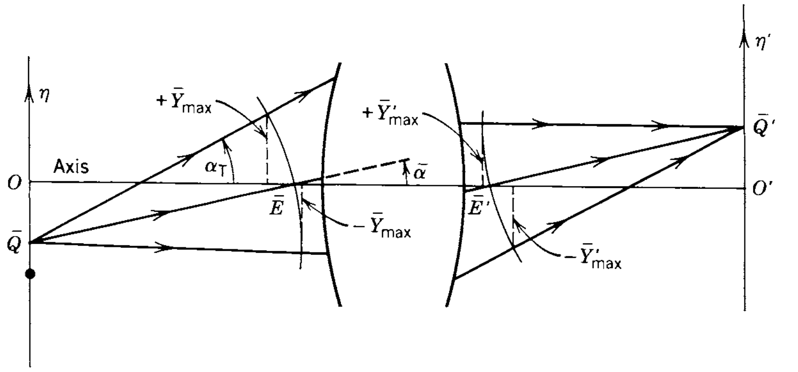


Figure 3.5. Tangential plane of an optical system.

in turn, intersects the axis at $\bar{E}$, and the central ray, drawn from $\bar{Q}$ through $\bar{E}$, is called the *pupil ray*. This ray is sometimes confused with the *principal* or *chief ray*, which passes through the intersection of the entrance pupil plane and the axis. In the coordinate system being defined, $\bar{E}$ locates a new off-axis entrance pupil whose "plane" is actually a portion of a sphere centered on $\bar{Q}$ with a radius $\bar{R}$ equal to the distance $\overline{QE}$. As indicated in Fig. 3.5, the pupil ray makes an angle $\bar{\alpha}$ and the upper edge ray an angle α_T with the axis.

The conjugate of $\bar{E}$ in image space is $\bar{E}'$, which locates the exit pupil in this new system of coordinates.

Real-space coordinates in the object plane are $\bar{\xi}$ and $\bar{\eta}$ with O as the origin. For points in the newly defined pupil "plane," the coordinates are $\bar{X}$ and $\bar{Y}$, where $\bar{X}$ is measured parallel to $\bar{\xi}$, $\bar{Y}$ is measured parallel to $\bar{\eta}$, and the *axis* is the origin. Because of the spherical shape of the new pupil surface, different coordinate points will not only have different $\bar{X}$ and $\bar{Y}$, but the distances from the axis that determine the values of $\bar{X}$ and $\bar{Y}$ will generally be measured from different points of origin on the axis.

A plane through the pupil ray and perpendicular to the tangential plane is defined as the *sagittal plane*. In the isometric diagram of Fig. 3.6, the sagittal plane passes through the pupil ray marked with an $\bar{R}$ and the x-axis, which passes through $\bar{E}$. As indicated earlier, a single tangent plane extends from the object point, through the optical components of the system, and to the image point; however, each time that the pupil ray is refracted at an optical surface, a new sagittal plane has to be defined.

In Fig. 3.6, the curved dashed line passing through $\bar{E}$ and $\bar{P}_S$ is the intersection of the pupil sphere and the xz coordinate plane; hence, it is the locus of points on the pupil sphere with the coordinates $(\bar{X}, 0)$. The *sagittal rays* are defined as those that originate at $\bar{Q}$ and pass through this intersection. It is to be noted that only the pupil ray in the sagittal fan is in the sagittal plane; the other rays in the fan lie in a conical surface with apex at $\bar{Q}$. In Fig. 3.6, the sagittal fan, except for the pupil ray, is above the sagittal plane.

Just as $\bar{Y}_{\max}$ has already been defined as the pupil sphere coordinate of the ray at $\bar{X} = 0$, which passes through the edge of the entrance pupil, we now define $\bar{X}_{\max}$ as the pupil sphere coordinate of the ray at $\bar{Y} = 0$, which passes through the edge of the entrance pupil. As in the earlier definition, there is a second edge ray at $-\bar{X}_{\max}$. In Fig. 3.6, $\bar{X}_{\max}$ is shown as a distance in the xz plane from the point $\bar{P}_S$ to the optic axis. As indicated in the figure, the *sagittal pupil angle* $\alpha_S = \text{arc sin}\,(\bar{X}_{\max}/\bar{R})$.

When one makes the simplifying paraxial assumptions, a paraxial ray is conventionally used from the axial object point O to define ray heights at the paraxial entrance and exit pupils. Similarly, in the canonical system now being defined, rays "close to" the pupil ray are used to define corresponding close-ray heights and close-ray angles in the tangential and sagittal sections and in the entrance and exit pupil spheres at $\bar{E}$ and $\bar{E}'$. A simplifying close-ray assumption is to measure ray heights on a plane tangent to the pupil sphere at $\bar{E}$. This plane is represented in Fig. 3.7 by a broken line perpendicular to the pupil ray at $\bar{E}$. Consistent with subscript and prime usage already shown, the notation for the various close-ray heights and angles is h_T, h_S, h'_T, h'_S, α_T, α_S, α'_T, and α'_S. When the object point is moved to O on the axis where no tangent–sagittal

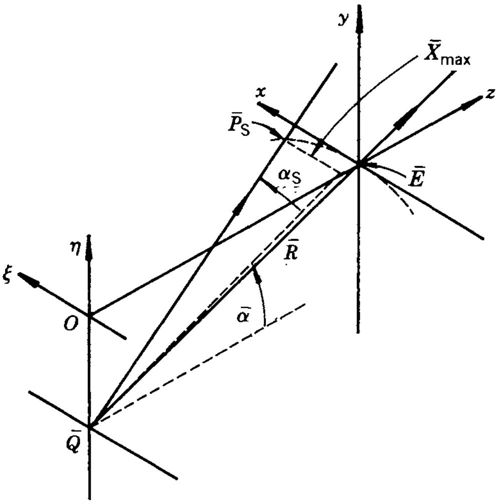


Figure 3.6. Sagittal plane in object space.

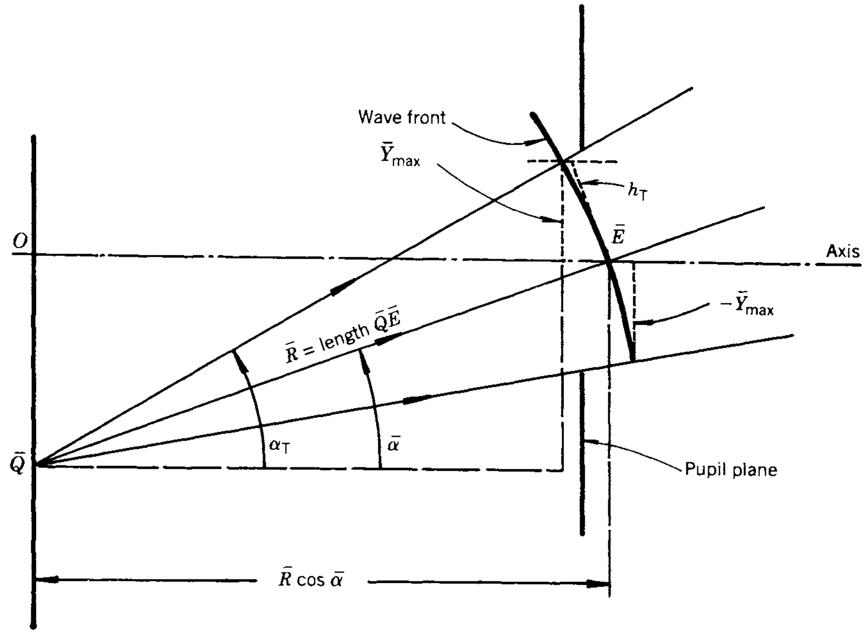


Figure 3.7. Reference wave front and edge-ray height in object space.

distinction can be made, this notation reduces to h, h', α, and α', which are the symbols for paraxial ray heights and ray angles.

As suggested by the notation, close-ray heights and angles are defined only for rays in the tangential and sagittal fans. A ray in the tangential fan has the coordinates $(0, \bar{Y})$ in object space and $(0, \bar{Y}')$ in image space. The corresponding ray heights are

$$h_T = \bar{Y}/\bar{N}, \qquad h_T' = \bar{Y}'/\bar{N}', \tag{3-5}$$

where $\bar{N}$ and $\bar{N}'$ are the direction cosines of the pupil ray with respect to the optic axis (z-axis). A close-ray in the sagittal fan has the coordinates $(\bar{X}, 0)$ in object space and $(\bar{X}', 0)$ in image space. The corresponding ray heights are

$$h_S = \bar{X}, \qquad h_S' = \bar{X}'. \tag{3-6}$$

Formulation of the ray height in object space (Fig. 3.7) in Eq. (3-5) results from the solution of a small right triangle in the vicinity of $\bar{E}$ (triangle not shown), which includes as one of its acute angles the pupil ray angle $\bar{\alpha}$. In the tangential plane, the pupil sphere and the tangent plane are coincident at $\overline{E}$. The direction cosine in this instance is $\overline{N} = \cos \overline{\alpha}$. The expression for h'_T is obtained by the same procedure as the one for h_T except that the geometry is in image space. By inspection, one can see in Fig. 3.6 that close-ray heights in the sagittal fan are simply their $\overline{X}$ coordinates (Eq. (3-6)).

Close-ray analysis for off-axis sources, like paraxial analysis for on-axis sources, tends to be just a preliminary procedure followed by an analysis that takes into account the peculiarities of the rays near the edge of the pupils. To prepare for this analysis for a specified system, the edge rays are first found by ray tracing, which is discussed in the next section. From the locations of the edge rays in the tangential plane, each pupil sphere is found by locating the spherical surface centered on the object or image point and having equal $\overline{Y}_{\text{max}}$ values for the edge rays. The intersection of the pupil sphere with the optic axis then establishes the location of the pupil ray.

In Fig. 3.7, the particular close-ray height h_T is that of the edge ray. From the geometry of the figure and the definition already given for $\overline{N}$,

$$h_T = \overline{Y}_{\text{max}}/\overline{N} = \overline{R}\,(\sin \alpha_T - \sin \overline{\alpha})/\overline{N}. \tag{3-7}$$

An angle variable u_T can be defined as

$$u_T \equiv h_T/\overline{R} = (\sin \alpha_T - \sin \overline{\alpha})/\overline{N}. \tag{3-8}$$

The corresponding sagittal plane (Fig. 3.6) ray variables are

$$h_S = \overline{X}_{\text{max}} = \overline{R} \sin \alpha_S, \tag{3-9}$$

$$u_S = \overline{X}_{\text{max}}/\overline{R} = \sin \alpha_S. \tag{3-10}$$

It is evident that each of the second ray variables, u_T or u_S, is simply the first variable normalized to the pupil sphere radius $\overline{R}$. The ray variables defined in the four equations above correspond to the traditional h and u found in ray-tracing problems. Unlike the limited application of paraxial theory mentioned earlier in this chapter, the theory involving h_T, u_T, h_S, and u_S ordinarily is useful over the full aperture; in fact, in what follows, these variables usually apply to edge rays.

The numerical apertures for the tangential and sagittal planes are defined respectively as

$$(\text{N.A.})_T = n\overline{Y}_{\text{max}}/\overline{R} = n(\sin \alpha_T - \sin \overline{\alpha}), \tag{3-11}$$

$$(\text{N.A.})_S = n\overline{X}_{\text{max}}/\overline{R} = n \sin \alpha_S. \tag{3-12}$$

When the object point is moved to O on the axis where no tangent–sagittal distinction can be made, $\overline{\alpha}$ becomes zero, the ray variables become

$$h = h_T = h_S = Y = X = R \sin \alpha, \tag{3-13}$$

$$u = u_T = u_S = h/R = \sin \alpha, \tag{3-14}$$

$$\text{N.A.} = (\text{N.A.})_T = (\text{N.A.})_S = nY/R = nX/R = n \sin \alpha. \tag{3-15}$$

TRANSFER EQUATIONS

The ray variables h and u defined in the preceding section completely define the edge rays at the entrance pupil reference sphere. The process of evaluating the ray variables at successive surfaces of an optical system is called *ray tracing*. For paraxial rays, standard optics texts give *transfer equations* that calculate values of h and u at a given surface in terms of the values of these variables at the preceding surface. In these equations, the characteristics of the optical system taken into account are the distance along the axis between the two surface vertices, the radius of curvature of the second surface, and the index of refraction of the intervening medium. (See [1, pp. 41–45; 7, pp. 190–194; 8, pp. 135–140; 9, pp. 24–33; 10, pp. 81–95].) Ray variables in the exit pupil and the image plane are found, therefore, by starting with the variables in the object plane and successively calculating them at the succeeding surfaces until the exit pupil and the image plane are reached.

Ray tracing for nonparaxial rays, conventionally called *finite ray tracing*, is somewhat more involved than the procedure just described. In addition to tracing in the tangential, also known as the *meridional*, plane, one must trace rays in the sagittal plane and also those that are oriented so that they are in neither the tangential nor sagittal plane. The nonmeridional rays are called *skew rays*. Being the most general of ray classifications, skew rays require the most involved procedure for tracing. Again, the reader is referred to the standard texts for details. In proceeding from the object plane to the image plane, each ray intersection with an optical surface has to be located by three coordinate values, and the direction of the ray has to be described by three directions cosines. Application of Snell's law to get the refraction of a skew ray at an optical surface becomes an exercise in three dimensions to establish the new set of three direction cosines.

Even the very general ray tracing formulas that accommodate skew rays are usually applicable only to *spherical* optical surfaces. When aspherics are involved, a suitable expression describing the curvature at each ray–surface intersection must be available.

By whichever technique is appropriate for each lens design problem, ray tracing is the means by which image space parameters are determined.

The on-axis paraxial image point O' corresponding to the on-axis object point O is defined as the *on-axis image point*, and a plane perpendicular to the axis at this point is the *image plane*. The *pupil ray* $\overline{QE}$ in object space is found, by ray tracing, to cross the axis again in image space near the *exit pupil*, the paraxial exit pupil position E' having been determined by paraxial analysis. The intersection of the pupil ray with the axis in image space is labeled $\overline{E}'$ and is called the *off-axis exit pupil point*; the intersection of the pupil ray with the image plane is the *image point* $\overline{Q}'$. The image point $\overline{Q}'$ is the center of the *image space reference sphere*, which has a radius $\overline{R}'$ equal to $\overline{E}'\overline{Q}'$. The edge rays, traced from object space, intersect the reference sphere at the edge ray heights $\overline{Y}'_{\text{max}}$ and $\overline{X}'_{\text{max}}$, defined analogously to the corresponding edge ray heights in object space.

In object space, the four edge rays were arbitrarily originated at the point Q; however, because these rays then pass through an optical system that will generally have aberrations, they will not necessarily pass through the image point $\overline{Q}'$. Therefore, to set up a geometry by which we can define image space ray variables similar to the corresponding object space variables in Eqs. (3-7)–(3-12), we must assume an ideal optical system that does indeed bring all four edge rays together at $\overline{Q}'$. Then, by reasoning similar to that used for object space,

$$h'_{\text{T}} = \overline{Y}'_{\text{max}}/\overline{N}' = \overline{R}'(\sin\alpha'_{\text{T}} - \sin\overline{\alpha}')/\overline{N}', \tag{3-16}$$

$$u'_{\text{T}} = h'_{\text{T}}/\overline{R}' = (\sin\alpha'_{\text{T}} - \sin\overline{\alpha}')/\overline{N}', \tag{3-17}$$

$$(\text{N.A.})'_{\text{T}} = n'\overline{Y}'_{\text{max}}/\overline{R}' = n'(\sin\alpha'_{\text{T}} - \sin\overline{\alpha}'), \tag{3-18}$$

$$h'_{\text{S}} = \overline{X}'_{\text{max}} = \overline{R}'\sin\alpha'_{\text{S}}, \tag{3-19}$$

$$u'_{\text{S}} = \overline{X}'_{\text{max}}/\overline{R}' = \sin\alpha'_{\text{S}}, \tag{3-20}$$

$$(\text{N.A.})'_{\text{S}} = n'\overline{X}'_{\text{max}}/\overline{R}' = n'\sin\alpha'_{\text{S}}. \tag{3-21}$$

To find $\overline{Y}'_{\text{max}}$ and $\overline{X}'_{\text{max}}$ in the above expressions, edge rays are traced through the actual (not ideal) optical system to find their intersections with the image space reference sphere. The distances from these intersections to the optic axis are the respective maximum values. As the object point $\overline{Q}$ is moved toward the axis causing the image point also to approach the axis, the six equations above approach the following three as limiting equations:

$$h' = \overline{R}'\sin\alpha', \tag{3-22}$$

$$u' = \sin \alpha', \tag{3-23}$$

$$(\text{N.A.})' = n' \sin \alpha'. \tag{3-24}$$

Because certain assumptions were made to set up a workable ray geometry in image space, the image space ray variables and some of the angles in the expressions for them involve approximations not found in the corresponding object space quantities. However, when accurate calculations are to be based on image space parameters, attention is usually focused on the actual wave front and on the wave aberration function $W(\overline{B}')$ rather than on the rays; so we will not be concerned here with analyzing the errors introduced by the approximate image space ray geometry.

PUPIL VARIABLES

In Eqs. (3-5) and (3-6), h_T is the ray height of the ray whose coordinates on the pupil sphere are $(0, \overline{Y})$, and h_S is the ray height of the ray whose coordinates on the pupil sphere are $(\overline{X}, 0)$. Corresponding statements can be made about the ray heights in image space (primed values). As has been true in much of the previous discussion, we are now concerned with only the edge rays in the tangential and sagittal planes; so h_T and h_S are two constants equal to the indicated edge ray heights. If both sides of each equation in Eqs. (3-5) and (3-6) are divided by the ray height, the resulting ratios would all be unity. Starting with these particular values of these ratios in the tangential and sagittal planes, we allow $\overline{X}$ and $\overline{Y}$ to take on values off the two planes to satisfy the following equation for an ellipse:

$$(\overline{X}/h_S)^2 + (\overline{Y}/h_T\overline{N})^2 = 1. \tag{3-25}$$

Again, an identical equation except for the addition of primes can be set up for image space.

By assigning coordinate symbols to the ratios,

$$x_S = \overline{X}/h_S, \qquad y_T = \overline{Y}/h_T\overline{N}, \tag{3-26}$$

$$x'_S = \overline{X}'/h'_S, \qquad y'_T = \overline{Y}'/h'_T\overline{N}', \tag{3-27}$$

we have normalized coordinate systems for the entrance and exit pupils, respectively. When these symbols are substituted in Eq. (3-25) and the equivalent equation for image space,

$$x_S^2 + y_T^2 = 1 \quad \text{and} \quad x_S'^2 + y_T'^2 = 1, \tag{3-28}$$

the new coordinates describe the pupils as unit-radius circles. If $\overline{Q}\,'$ is in an isoplanatism patch, which is required for optical transfer function theory to be valid, it can be shown [3] that

$$x_S = x'_S, \qquad x_T = x'_T. \tag{3-29}$$

Equation (3-29) is equivalent to making aberrations stationary with image height and is also, according to Hopkins [3], a statement of the optical sine condition for freedom from coma near $\overline{Q}$. Insofar as the isoplanatism condition holds, Eq. (3-29) tells us that image space variables can be replaced by the corresponding object space variables without error. As a practical matter, the differences $x'_S - x_S$ and $x'_T - x_T$ are rarely found to exceed 0.01 in a normally corrected system, which means that errors are generally kept to not more than one percent of the pupil radius [6, p. 358].

REDUCED COORDINATES

Besides the coordinate systems already discussed, Hopkins and others have developed two other derived systems known as the *canonical* and *reduced* coordinates. A set of partial differential equations, the *canonical equations*, expressed in canonical coordinates, are significant in the analytical theory of aberrations and its application to automatic design. The reduced coordinates are particularly useful in diffraction theory.

We derive the reduced coordinates and formulas for reduced spatial frequencies here for the case of rotationally symmetric systems. Hopkins, in two important papers, has developed the reduced coordinates so that they can be applied to a completely general optical system [11].

The canonical and reduced coordinates are based on the object point $\overline{Q}$ and the image point $\overline{Q}\,'$ as origins in the object and image planes, respectively. If a point near $\overline{Q}$ has the coordinates (ξ_2, η_2) and the $\overline{Q}$ coordinates are $(\overline{\xi}_1, \overline{\eta}_1)$, the displacements in the two coordinate directions are

$$\Delta\xi = \xi_2 - \overline{\xi}_1 \quad \text{and} \quad \Delta\eta = \eta_2 - \overline{\eta}_1. \tag{3-30}$$

The object plane canonical coordinates (G_S, H_T) are defined in terms of these displacements as

$$G_S = n(\overline{X}_{\max}/\overline{R})\,\Delta\xi = n(\sin\alpha_S)\,\Delta\xi, \tag{3-31}$$

$$H_T = n(\overline{Y}_{\max}/\overline{R})\,\Delta\eta = n(\sin\alpha_T - \sin\overline{\alpha})\,\Delta\eta. \tag{3-32}$$

In the above definitions, the coefficients of $\Delta\xi$ and $\Delta\eta$ are recognized as the numerical apertures for the sagittal and tangential planes as defined in Eqs. (3-11) and (3-12). Image plane canonical coordinates are defined similarly:

$$G'_S = n'(\sin \alpha'_S)\,\Delta\xi', \tag{3-33}$$

$$H'_T = n'(\sin \alpha'_T - \sin \overline{\alpha}')\,\Delta\eta'. \tag{3-34}$$

If the condition expressed by Eq. (3-29) applies, it follows that

$$G_S = G'_S \quad \text{and} \quad H_T = H'_T. \tag{3-35}$$

This condition, as indicated earlier, means that the system is free of aberrations to the extent that a ray from $\overline{Q}$ (Fig. 3.4), which intersects the entrance pupil sphere at $\overline{B}$, would pass through the exit pupil sphere at $\overline{B}'$ and through the image plane at $\overline{Q}'$. Now if we introduce aberrations into the system so that the ray instead passes *near* $\overline{B}'$ and *near* $\overline{Q}'$, displaced from $\overline{B}'$ in the pupil sphere by the increments $\delta x'_S$ and $\delta y'_T$ and displaced from $\overline{Q}'$ in the image plane by the increments $\delta G'_S$ and $\delta H'_T$, the following relations can be derived [2, 3]:

$$\delta x'_S = \partial W/\partial G_S, \qquad \delta y'_T = \partial W/\partial H_T, \tag{3-36}$$

$$\delta G'_S = -\partial W/\partial x_S, \qquad \delta H'_T = -\partial W/\partial y_T, \tag{3-37}$$

which are the *canonical equations*—so called because of their similarity in form to Hamilton's canonical equations of motion. Because the coordinates G_S, H_T, G'_S, and H'_T are involved in these equations, they are known as *canonical coordinates* as we have already referred to them in setting up their definitions.

The canonical equations show that the aberration function W is a function of G_S and H_T, which locate the originating point of the ray on the object plane, as well as a function of x_S and y_T, which indicate where the ray passes through the pupil sphere:

$$W = W(x_S, y_T; G_S, H_T). \tag{3-38}$$

Although the scope of our discussion of the optical transfer function does not permit further interpretation of the canonical equations, the student of the analytical theory of aberrations and its application to automatic optical design could profitably pursue them in the technical literature.

Hopkins' *reduced coordinates* u_S, v_T, u'_S, and v'_T are derived from the canonical coordinates by normalizing with respect to wavelength:

$$u_S = G_S/\lambda, \qquad v_T = H_T/\lambda, \tag{3-39}$$

$$u'_S = G'_S/\lambda, \qquad v'_T = H'_T/\lambda. \tag{3-40}$$

Unfortunately the symbols u_S and u'_S used for two of the reduced coordinates are also used for two of the ray variables defined by Eqs. (3-10) and (3-20). Because this ambiguity tends to be the practice in the literature of optics, no attempt is made here to invent a new set of symbols for either the sagittal ray variables or the sagittal reduced coordinates. Usually the context will indicate which use of u_S and u'_S is meant.

Certain numerical aperture ratios, called *aperture scaling factors*, are closely related to significant optical system properties. The sagittal and tangential aperture scaling factors are defined as

$$\rho_S = (\text{N.A.})_S/(\text{N.A.}), \tag{3-41}$$

$$\rho_T = (\text{N.A.})_T/(\text{N.A.}). \tag{3-42}$$

An example of their significance is given in Chapter 5 where it is shown that the diffraction-based spread function is anamorphotic to the same degree as the ratio of the scaling factors. Having defined these factors, we sometimes find it convenient to express the reduced coordinates in terms of them:

$$u_S = (n\rho_S\, \Delta\xi/\lambda) \sin\alpha, \tag{3-43}$$

$$v_T = (n\rho_T\, \Delta\eta/\lambda) \sin\alpha. \tag{3-44}$$

Equation (3-43) results from combining the relations expressed in Eqs. (3-12), (3-15), (3-31), (3-39), and (3-41); Eq. (3-44) results from combining the corresponding equations for the tangential plane. Similar derivations for the image plane result in

$$u'_S = (n'\rho'_S\, \Delta\xi'/\lambda) \sin\alpha', \tag{3-45}$$

$$v'_T = (n'\rho'_T\, \Delta\eta'/\lambda) \sin\alpha'. \tag{3-46}$$

Reduced spatial frequencies in the u_S and v_T coordinates are defined by

$$s_S = 1/\delta u_S \quad \text{and} \quad s_T = 1/\delta v_T, \tag{3-47}$$

where δu_S and δv_T are the spatial periods in the sagittal and tangential planes, respectively, in object space. From the relations expressed in Eqs. (3-43) and (3-44), the following equations relating the periods in reduced coordinates to the periods in the object plane coordinates can be written:

$$\delta u_S = \left[(n\rho_S/\lambda) \sin\alpha\right] \delta(\Delta\xi), \tag{3-48}$$

$$\delta v_T = \left[(n\rho_T/\lambda) \sin\alpha\right] \delta(\Delta\eta). \tag{3-49}$$

The relations between frequencies, reciprocals of periods, are

$$s_S = [\lambda/(n\rho_S \sin \alpha)]\,\omega_S, \tag{3-50}$$

$$s_T = [\lambda/(n\rho_T \sin \alpha)]\,\omega_T, \tag{3-51}$$

where ω_S and ω_T are the spatial frequencies at the intersections of the object plane and the sagittal and tangent planes, respectively. On the image plane, the corresponding equations are

$$s'_S = [\lambda/(n'\rho'_S \sin \alpha')]\,\omega'_S, \tag{3-52}$$

$$s'_T = [\lambda/(n'\rho'_T \sin \alpha')]\,\omega'_T. \tag{3-53}$$

From the definitions of numerical aperture (Eqs. (3-11), (3-12), and (3-15)), it is evident that the dimensionless reduced spatial frequency that we have been discussing above is the real-space spatial frequency multiplied by the wavelength in free space and divided by the numerical aperture. Also from the definitions of numerical aperture and from the reciprocal relations between period and frequency, we can write

$$s_S = \lambda/[n\delta(\Delta\xi)\sin \alpha_S], \tag{3-54}$$

$$s'_S = \lambda/[n'\delta(\Delta\xi')\sin \alpha'_S]. \tag{3-55}$$

When

$$n'\delta(\Delta\xi')\sin \alpha'_S = n\delta(\Delta\xi)\sin \alpha_S, \tag{3-56}$$

it is commonly said that the *sine condition* holds (which we have already said is true over an isoplanatism patch); so, from Eqs. (3-54) and (3-55), the sine condition causes the reduced spatial frequencies in image and object spaces to be equal.

Equations corresponding to Eqs. (3-54) and (3-55) can be derived for the tangential plane.

SHIFTING THE IMAGE PLANE

Earlier in this chapter, the image plane was defined as a plane perpendicular to the optic axis at the paraxial image point. When the image point was moved off the optic axis, it was identified as the intersection of the pupil ray with the image plane. A reference sphere, centered on the image point, was set up so that its radius was the length of the pupil ray from the image point to the intersection of the pupil ray with the optic axis ($\bar{E}'\bar{Q}'$ in Fig. 3.5). Wave-front aberration was defined as the distance between this reference sphere and the actual converging wave front as it passed through $\bar{E}'$.

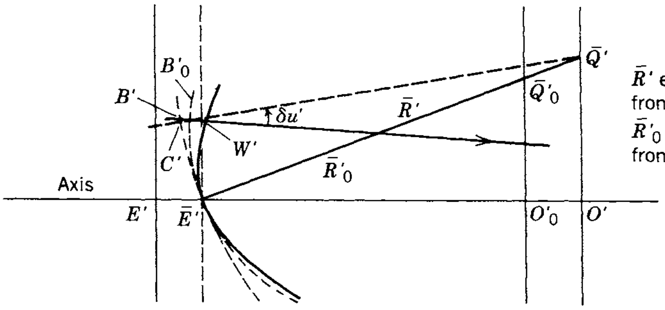
One way to improve the "fit" between the wave front and the reference sphere, that is, to reduce aberrations, is to alter the optical elements and the spaces between them to improve the shape of the wave front. An alternative approach is to change the reference sphere to reduce its separation from the wave front. In considering possible changes, it is convenient for the analysis to require that the sphere always pass through $\bar{E}'$ and that its center always lie along the pupil ray. This leaves one freedom, the selection of the sphere radius so that the reference sphere more closely conforms to the actual wave front. Since one end of the radius line, $\bar{E}'$, is fixed, changing the radius changes the position of the sphere center, which is on the image plane. So altering the reference sphere to reduce aberrations is equivalent to selecting a more desirable position for the image plane. Defocusing, a familiar bane among manipulators of optical equipment, can be said, therefore, to be the result of a poor choice of reference sphere caused by picking the wrong image plane.

From the foregoing discussion, it is apparent that an expression relating a change in wave-front distortion to a change in image position would be useful; in fact, applying such an expression would be equivalent to the oft repeated operation of focusing with actual optical equipment. The following analysis to get the desired expression follows a similar development first given by Hopkins and Yzuel [5].

In Fig. 3.8 the plane through the points O_0' and $\bar{Q}_0'$ is the paraxial image plane, and the line through O' and $\bar{Q}'$ represents the image plane in a slightly shifted position. (The shift displacement $O_0'O'$ has been exaggerated in the figure for clarity.) The pupil ray $\bar{E}'\bar{Q}_0'$ intersects the shifted plane at $\bar{Q}'$, which is the shifted image point or *shifted focus*. The paraxial exit pupil point is E'. The conventionally defined reference sphere centered on $\bar{Q}_0'$ with radius $\bar{R}_0' = \bar{E}'\bar{Q}_0'$ is represented by the arc passing through $\bar{E}'$ and $\bar{B}_0'$; the reference sphere centered on the shifted focus $\bar{Q}'$ with radius $\bar{R}' = \bar{E}'\bar{Q}'$ is represented by the arc passing through $\bar{E}'$ and B'. Besides the two radii just defined, the lengths of other line segments in Fig. 3.8 are also useful parameters and will be identified by symbols as

$$\tilde{R}_0' = \overline{\bar{E}'O_0'}, \qquad \tilde{R}' = \overline{\bar{E}'O'},$$
$$R_0' = \overline{E'O_0}, \qquad R' = \overline{E'O'}. \tag{3-57}$$

With the image plane through O_0' in Fig. 3.8, the wave aberration for the in-



$\bar{R}'$ extends from $\bar{E}'$ to $\bar{Q}'$.
$\bar{R}'_0$ extends from $\bar{E}'$ to $\bar{Q}'_0$.

Figure 3.8. A shift of the image plane. $\bar{R}'_0$ is the length of line segment $\overline{\bar{E}'\bar{Q}'_0}$. $\bar{R}'$ is the length of line segment $\overline{\bar{E}'\bar{Q}'}$.

dicated ray, according to the definition already given, is n', the index of refraction of image space, times the distance from B'_0 to W', the product being the optical path length between the reference sphere and the wave front, normal to the wave front at W'. If the image plane is shifted to O', the distance factor in the wave aberration is measured from B' on the new reference sphere to W'; so the change in the wave aberration δW_L is given by

$$\delta W_L = n'p, \tag{3-58}$$

where p is the distance from B' to B'_0.

Our purpose is to develop Eq. (3-58) so that δW_L is expressed in more convenient parameters. By mathematical experimentation, one finds that derived expressions are considerably simplified if the distance from B' to B'_0 is approximated by the distance from C' to B'_0. This approximation is seen to be good in a practical system where O' is close to O'_0, and $\delta u'$ is much smaller than the angle shown large for clarity in Fig. 3.8.

The coordinates of $\bar{Q}'$, with $\bar{E}'$ as the origin, are $(\bar{R}'\bar{L}', \bar{R}'\bar{M}', \bar{R}'\bar{N}')$ where $\bar{L}'$, $\bar{M}'$, and $\bar{N}'$ are the direction cosines for the pupil ray through the points $\bar{E}'$, $\bar{Q}'_0$, and $\bar{Q}'$. Since no initial condition has been placed on the coordinate system as to its orientation around the $\bar{Z}'$ axis (optic axis), we are free to rotate the coordinate system so that the pupil ray lies in the $\bar{Y}'\bar{Z}'$ plane, thus making the $\bar{L}'$ direction cosine zero and the $\bar{Q}'$ coordinates $(0, \bar{R}'\bar{M}', \bar{R}'\bar{N}')$. The coordinates of B'_0 will be designated $(\bar{X}', \bar{Y}', \bar{Z}')$; B'_0 is not necessarily in the $\bar{Y}'\bar{Z}'$ plane. A direction cosine identity in analytic geometry gives us the following relation:

$$\bar{M}'^2 + \bar{N}'^2 = 1. \tag{3-59}$$

The square of the distance between points B_0' and $\bar{Q}'$ is

$$\overline{B_0'\bar{Q}}'^2 = (\bar{R}' - p)^2 = \bar{X}'^2 + (\bar{R}'\bar{M}' - \bar{Y}')^2 + (\bar{R}'\bar{N}' - \bar{Z}')^2. \quad (3\text{-}60)$$

Similarly,

$$\overline{B_0'\bar{Q}_0'}^2 = \bar{R}_0'^2 = \bar{X}'^2 + (\bar{R}_0'\bar{M}' - \bar{Y}')^2 + (\bar{R}_0'\bar{N}' - \bar{Z}')^2. \quad (3\text{-}61)$$

Both Eq. (3-60) and Eq. (3-61) are expressions of the Pythagorean theorem. By expanding the squared binomials in these equations and substituting according to Eq. (3-59), the following can be written:

$$p^2 - 2\bar{R}'p = \bar{X}'^2 + \bar{Y}'^2 + \bar{Z}'^2 - 2\bar{Y}'\bar{R}'\bar{M}' - 2\bar{Z}'\bar{R}'\bar{N}', \quad (3\text{-}62)$$

$$0 = \bar{X}'^2 + \bar{Y}'^2 + \bar{Z}'^2 - 2\bar{Y}'\bar{R}_0'\bar{M}' - 2\bar{Z}'\bar{R}_0'\bar{N}'. \quad (3\text{-}63)$$

Combining these two equations gives

$$2\bar{R}'p[p/(2\bar{R}') - 1] = (2\bar{Y}'\bar{M}' + 2\bar{Z}'\bar{N}')(\bar{R}_0' - \bar{R}'), \quad (3\text{-}64)$$

which reduces to

$$p = (\bar{Y}'\bar{R}_0'\bar{M}' + \bar{Z}'\bar{R}_0'\bar{N}')(1/\bar{R}_0' - 1/\bar{R}')[1 - p/(2\bar{R}')]^{-1}. \quad (3\text{-}65)$$

From Eq. (3-63), it is evident that the first binomial on the right side of Eq. (3-65) is $(\bar{X}'^2 + \bar{Y}'^2 + \bar{Z}'^2)/2$. Then, by recognizing that p will always be extremely small relative to $\bar{R}'$ in practical optical systems, the following reasoning toward an approximation can be made:

$$\begin{aligned} p &\ll 2\bar{R}', \\ p/(2\bar{R}') &\ll 1, \\ [1 - p/(2\bar{R}')]^{-1} &\cong 1. \end{aligned} \quad (3\text{-}66)$$

With the indicated substitution and approximation, the expression for the change in wave aberration becomes

$$\delta W_{\mathrm{L}} = n'p = (n'/2)(\bar{X}'^2 + \bar{Y}'^2 + \bar{Z}'^2)(1/\bar{R}_0' - 1/\bar{R}'). \quad (3\text{-}67)$$

The expressions in the second and third sets of parentheses can be stated in more meaningful parameters. From the definitions in Eq. (3-57) and the geometry of Fig. 3.8

$$\bar{R}_0' = \tilde{R}_0'/\bar{N}', \qquad \bar{R}' = \tilde{R}'/\bar{N}'. \quad (3\text{-}68)$$

Thus, the third set of terms in parentheses can be written

$$(1/\bar{R}'_0 - 1/\bar{R}') = \bar{N}'(1/\tilde{R}'_0 - 1/\tilde{R}'). \tag{3-69}$$

Since the difference between $\tilde{R}'$ and $\tilde{R}'_0$ is the distance the image plane is shifted, the derived factor on the right side of Eq. (3-69) is simply related to the shift. Because the coordinates of the point B'_0 with $\bar{E}'$ as the origin are (X', Y', Z'), application of the Pythagorean theorem gives the relation

$$\overline{\bar{E}'B'_0}^2 = \bar{X}'^2 + \bar{Y}'^2 + \bar{Z}'^2. \tag{3-70}$$

The line segment $\overline{\bar{E}'B'_0}$, whose length is squared in Eq. (3-70), would reduce to the ray height h' in the exit pupil in paraxial coordinates. If reduction to paraxial coordinates is further extended to the third parenthetical expression, Eq. (3-67) becomes

$$\delta W_{20} = (n'h'^2/2)(1/R'_0 - 1/R'), \tag{3-71}$$

where the radius symbols without bars can be substituted according to the definitions in Eq. (3-57). (The subscript 20 is explained in the next chapter.) For paraxial rays, E' and $\bar{E}'$ may be considered the same point. With this approximation, Eq. (3-71) can be written as

$$\delta \tilde{W}_{20} = (n'h'^2/2)(1/\tilde{R}'_0 - 1/\tilde{R}'). \tag{3-72}$$

By combining the relations in Eqs. (3-72), (3-69), and (3-67), we obtain

$$\delta W_{\mathrm{L}} = (\delta \tilde{W}_{20}/h'^2)\,\bar{N}'(\bar{X}'^2 + \bar{Y}'^2 + \bar{Z}'^2), \tag{3-73}$$

which gives the change in wave-front distortion in terms of the corresponding change in paraxial wave-front distortion, the paraxial ray height, and the real-space coordinates of the intersection of the edge ray with the exit pupil reference sphere. MacDonald and Hopkins and Yzuel [4, 5] have pointed out that a commonly used expression for δW_{L} that omits the $\bar{Z}'^2$ given in Eq. (3-73) can lead to significant errors. In fact, in the development of the aberration function in the next chapter, dependence on the $\bar{Z}'$ coordinate indicates defocusing affecting the coefficients of certain aberration types such as spherical aberration, coma, and astigmatism.

MAGNIFICATION WITH DISTORTION

Certain optical systems are designed to be *anamorphotic*; that is, the magnification in one direction, say horizontally, is significantly greater than at right angles to that direction. A familiar example is the wide-screen moving picture system in which the camera lens predistorts the film image by greater vertical magnification and the projector lens compensates by applying greater horizontal magnification from film to screen. Our interest in this book, however, is principally in centered systems having circular symmetry about the axis in which anamorphosis is an undesirable aberration [4]. Even a slight residual anamorphism in an optical system has a significant effect on the optical transfer function.

In the optical literature reference is made to *local magnification* and to *finite magnification*. Local magnification can be defined by referring to a small cross as the object having arm lengths of $\delta\xi$ and $\delta\eta$. The sagittal and tangential magnifications at the intersection of the cross are defined as

$$m_S = \lim_{\delta\xi \to 0} \delta\xi'/\delta\xi, \tag{3-74}$$

$$m_T = \lim_{\delta\eta \to 0} \delta\eta'/\delta\eta. \tag{3-75}$$

A nonlimiting finite magnification is defined by referring to Fig. 3.9. A region in object space, shown as $RSQP$, is part of an annulus centered on the axial object point O. Because the optical system is circularly symmetrical, the image of this region is also part of an annulus centered on the axis; and the angles $\angle ROP$ and $\angle R'O'P'$ are equal. If $\overline{OR} = \overline{OP} = \eta$ and $\overline{O'R'} = \overline{O'P'} = \eta'$, the finite tangential magnification is

$$m_0 = \eta'/\eta. \tag{3-76}$$

The nonlimiting sagittal magnification at the annulus is

$$m_S = \overline{R'P'}/\overline{RP} = \eta' \angle R'O'P' / \eta \angle ROP = \eta'/\eta = m_0. \tag{3-77}$$

The finite magnification is a function of η whatever there is distortion or whenever the system is anamorphotic:

$$m_0 = f(\eta). \tag{3-78}$$

It is of interest to find relations between the local magnifications defined in Eqs. (3-74) and (3-75) and the more easily measured finite tangential magnification

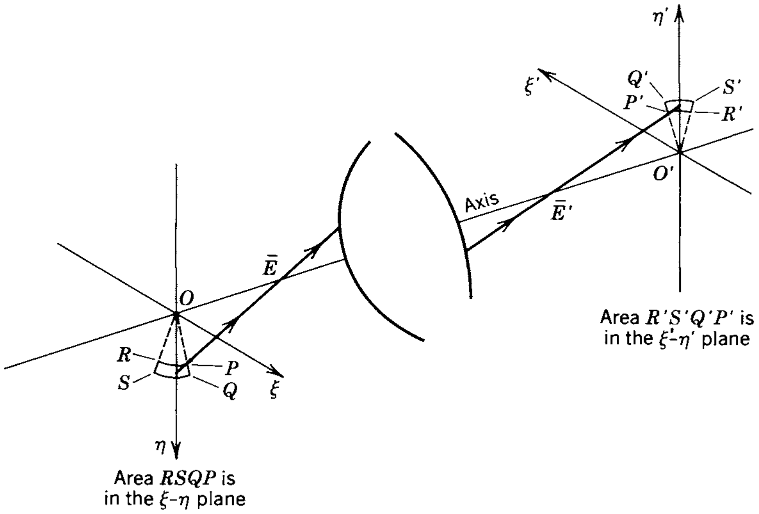

η'
ξ'
Q'
S'
P'
R'
O'
Axis
Ē'
Ē
O
R
S
P
Q
ξ
η
Area R'S'Q'P' is in the ξ'-η' plane
Area RSQP is in the ξ-η plane

Figure 3.9. Imaging a sector of an annulus.

defined in Eq. (3-76). As already suggested by the choice of notation for the two sagittal magnifications, they are equal to each other and to m_0, the finite tangential magnification. This results from symmetry of the system about the axis: Magnification along the arc $\overline{RP}$ is a constant; so, as this arc and its corresponding image are reduced in length in Eq. (3-77), m_S is not affected. Carried to the limit, this reduction causes the expression to be the same as Eq. (3-74), the definition of the local sagittal magnification.

The two tangential magnifications, however, are not in general equal. To find their relation, the expression for the finite tangential magnification m_0' at $\eta + \delta\eta$ is developed by Taylor's expansion:

$$\begin{aligned} m_0' &= f(\eta + \delta\eta) = f(\eta) + \delta\eta(\partial m_0/\partial\eta) + \mathcal{O}(\delta\eta)^2 \\ &= m_0 + \delta\eta(\partial m_0/\partial\eta) + \mathcal{O}(\delta\eta)^2, \end{aligned} \tag{3-79}$$

in which $\mathcal{O}(\delta\eta)^2$ represents the sum of all higher order terms. By the definition of Eq. (3-76), m_0' at $\eta + \delta\eta$ can also be expressed as

$$m_0' = (\eta' + \delta\eta')/(\eta + \delta\eta). \tag{3-80}$$

(Note that the prime in m_0' does not specify image space as has been the practice in most of the other primed symbol usage of this book.) By equating the two expressions for m_0' given in Eqs. (3-79) and (3-80), dropping the sum of higher order terms, dropping a term with $\delta\eta^2$ as a factor, and substituting $m_0\eta$ for η' according to Eq. (3-76), we obtain

$$\delta\eta'/\delta\eta = m_0 + \eta(\partial m_0/\partial\eta). \tag{3-81}$$

Because of the nature of the dropped terms, Eq. (3-81) improves in accuracy as $\delta\eta$ approaches zero. In the limit, according to Eq. (3-75), the derivative is equal to m_T, so

$$m_T = m_0 + \eta(\partial m_0/\partial\eta), \tag{3-82}$$

which is the desired relation between the local tangential magnification m_T and the finite tangential magnification m_0.

In a distortion-free system, $m_S = m_T = m_0$.

REFERENCES

1. W. T. Welford, *Aberrations of the Symmetrical Optical System.* Academic, London, 1974.
2. H. H. Hopkins, The Development of Image Evaluation Methods. *SPIE Proc.* **46,** 2 (1974). (Please see the note following Ref. 2 of Chapter 1.)
3. H. H. Hopkins, Canonical Pupil Coordinates in Geometrical and Diffraction Image Theory. *Japan J. Appl. Phys.* **4,** Suppl. 1, 31 (1965). (Proceedings of a conference on Photographic and Spectrographic Optics held in Tokyo, 1964.)
4. J. MacDonald, The Calculation of the Optical Transfer Function. *Opt. Acta* **18,** 269 (1971).
5. H. H. Hopkins and M. J. Yzuel, The Computation of Diffraction Patterns in the Presence of Aberrations. *Opt. Acta* **17,** 157 (1970).
6. H. H. Hopkins, The Use of Diffraction-Based Criteria of Image Quality in Automatic Optical Design. *Opt. Acta* **13,** 343 (1966).
7. M. Born and E. Wolf, *Principles of Optics*, 3d ed. Pergamon, Oxford, 1965.
8. C. S. Williams and O. A. Becklund, *Optics: A Short Course for Engineers and Scientists.* Wiley-Interscience, New York, 1972.
9. W. J. Smith, *Modern Optical Engineering: The Design of Optical Systems.* McGraw-Hill, New York, 1966.
10. O. N. Stavroudis, *The Optics of Rays, Wavefronts, and Caustics.* Academic, New York, 1972.
11. H. H. Hopkins, Image Formation by a General Optical System, 1. General Theory. *Appl. Opt.* **24,** 2491 (1985); Image Formation by a General Optical System, 2. Computing Methods. *Appl. Opt.* **24,** 2506 (1985).

4

Diffraction Integral and Wave-Front Aberration Function

INTRODUCTION

Discussions in earlier chapters have already indicated that the optical transfer function is mathematically related to the diffraction integral, the wave-front aberration function, and the spread function. This chapter is concerned with the diffraction integral and the wave-front aberration function (frequently shortened to *wave-front aberration, wave aberration, aberration function, wave aberration function, wave-front distortion,* or *wave distortion* depending on the emphasis and context).

To contribute to a complete mathematical description of an optical system output, the wave-front aberration function is made part of the *pupil function*, which will also be discussed in this chapter.

In the previous chapter, the concept of wave-front aberration in image space was developed by comparing the actual emerging wave front, which originated from a point source in object space, with an ideal spherical surface at the exit pupil. It was emphasized in that discussion that discrepancies between the wave front and the spherical surface accounted for aberration in the image. Even without defects of this kind, the desired point image actually becomes a spread-out diffraction pattern. In other words, even if the emerging wave front were perfectly spherical over its extent, the image of the point source would still be a bright central disk surrounded by unevenly spaced, concentric circles because the converging wave front producing the image is only a small part of a sphere. Although the image-forming optical system causes both the limiting of the wave front and the distorting of its shape from sphericity, the effect of the first, which is diffraction, is usually regarded as something to live with; but the effects of the second, aberrations, are targets for elimination. Thus, when this objective is achieved, the high quality of the optics is indicated by designating the design as a ''diffraction-limited system.''

To develop further the wave concepts in image space and to get some understanding of how diffraction comes about, the development of the diffraction integral is undertaken before the specific characteristics of various aberrations are studied.

WAVE-FRONT EXPRESSIONS AND THE DIFFRACTION INTEGRAL

Figure 4.1 schematically represents the imaging of an off-axis point object $\bar{Q}$ in the tangential plane. Two positions of a wave front, originating at $\bar{Q}$ and finally reaching the image plane in the vicinity of $\bar{Q}'$, are shown where it passes through the pupil points $\bar{E}$ and $\bar{E}'$. It is assumed that the wave front passes through the optical system undistorted (no aberrations) so that the wave front coincides with the pupil sphere. Furthermore, it is assumed that diffraction effects during passage of the wave front from the aperture within the system, where the wave front is actually truncated, to the exit pupil are negligible so that it is legitimate to investigate diffraction effects as if a limitation of the wave-front extent is fashioned at the exit pupil (near $\bar{E}'$).

A general point on the wave front is designated B in object space and B' in image space. Under the assumptions stated, the complex amplitude $\hat{U}_0(B')$ at B', in accordance with wave theory (see Appendix C), can be represented by

$$\hat{U}_0(B') = [G\exp(i\mathcal{k}n'\bar{R}')]/\bar{R}', \tag{4-1}$$

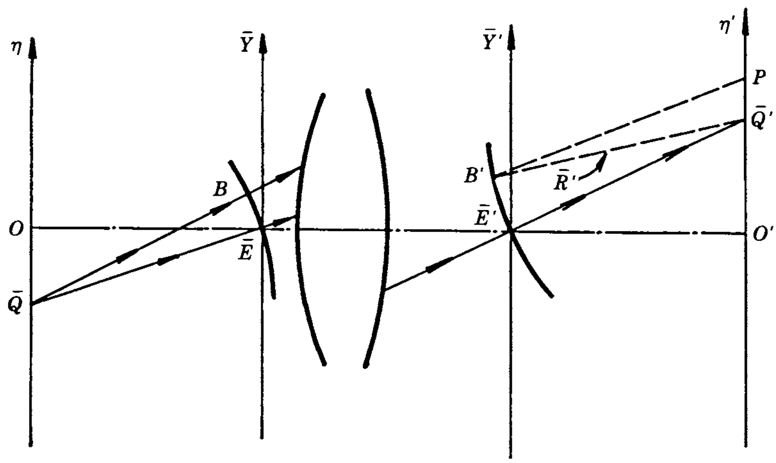


Figure 4.1. Schematic of a wave front diverging from the object point and converging toward the image point.

where G is a scalar constant, $\overline{R}'$ is the distance from B' to $\overline{Q}'$ and also the distance from $\overline{E}'$ to $\overline{Q}'$ because both distances are radii of the spherical wave front, n' is the refractive index in image space, k is the symbol for the ratio $2\pi/\lambda$, and λ is the wavelength in vacuum of the light propagating from $\overline{Q}$.

The coordinates at B' are (x'_S, y'_T) as defined by Eq. (3-27). Because the wave front has been assumed spherical, $\overline{R}'$ has a constant value for all (x'_S, y'_T) within the unit circle of the pupil. It follows that $\hat{U}_0(B')$ as represented by Eq. (4-1) is also constant over a wave front.

If the optical system has introduced aberrations and thus distorted the wave front now passing through $\overline{E}'$ so that it no longer coincides with the pupil sphere, $\overline{R}'$ and $\hat{U}_0(B')$ become functions of (x'_S, y'_T). In practical optical designs, the variation of $\overline{R}'$ over the whole wave front is just a few wavelengths of light. An inspection of the MTF curves given in Appendix A shows how much the spatial frequency response of an optical system deteriorates with wave-front distortion. It is obvious that with a maximum wave-front distortion of 10 wavelengths or more, the image is badly degraded; and, for most applications, a considerable amount of preliminary aberration correction would be required.

Ten wavelengths of near infrared at $\lambda = 1\ \mu\text{m}$ in air $(n' = 1)$ would measure $10\ \mu\text{m}$. On the other hand, 10 wavelengths of light at $\lambda = 0.5\ \mu\text{m}$ in a medium where $n' = 1.6$ would measure only a little over $3\ \mu\text{m}$. Yet, because the degradation of response is more directly related to the number of wavelengths of discrepancy in the distorted wave front than to the actual distance between pupil sphere and wave front, wave-front aberration is often expressed in terms of wavelengths.

A constant G in Eq. (4-1) for the amplitude of the numerator is usually an acceptable approximation; however, when the complex amplitude does vary significantly over the wave front, G has to be recognized as a function of the coordinates:

$$G = G(x'_S, y'_T). \tag{4-2}$$

Also, if $\overline{R}'$ varies significantly over the wave front, it is convenient to express it as a binomial:

$$\overline{R}' = \overline{R}'_0 - m\lambda, \tag{4-3}$$

where $\overline{R}'_0$ is the distance from $\overline{E}'$ to $\overline{Q}'$, the constant radius of the reference sphere; and m is the number of wavelengths that $\overline{R}'$ is shortened because the wave front at B' is closer to $\overline{Q}'$ than the reference sphere. (Of course, if the wave front is farther than $\overline{Q}'$ than the reference sphere, m is a negative number.)

If the expressions for G and $\overline{R}'$ from Eq. (4-2) and Eq. (4-3) are substituted in Eq. (4-1), then

$$\begin{aligned}\hat{U}_0(B') &= G(x'_S, y'_T)\left\{\exp[i k n'(\overline{R}'_0 - m\lambda)]\right\}/(\overline{R}'_0 - m\lambda)\\ &= G(x'_S, y'_T)\left[\exp(i k n'\overline{R}'_0)\right]\left[\exp(-i k n' m\lambda)\right]/(\overline{R}'_0 - m\lambda)\\ &= G(x'_S, y'_T)\left[\exp(i k n'\overline{R}'_0)\right]\left\{\exp[-i(2\pi n' m)]\right\}/(\overline{R}'_0 - m\lambda).\end{aligned} \tag{4-4}$$

The second term on the right side is a constant phase shift relative to the phase at $\overline{Q}'$. The ratio $\exp[-i(2\pi n'm)]/(\overline{R}'_0 - m\lambda)$ describes the significant phase variation of $\hat{U}_0(B')$. With m as a factor in the phase angle of the numerator, it is apparent that the variable phase shift of $\hat{U}_0(B')$ relative to the phase at the reference sphere is proportional to the number of wavelengths m that the wave front is displaced from the sphere. On the other hand, from our previous discussion of the magnitude of $m\lambda$, the denominator is negligibly affected by m:

$$\overline{R}'_0 - m\lambda = \overline{R}'_0(1 - m\lambda/\overline{R}'_0) \cong \overline{R}'_0. \tag{4-5}$$

From the definition of wave-front aberration as given in Chapter 3, we can assign the single symbol $W(x'_S, y'_T)$ to the product $m\lambda$. With the indicated changes in the terms of Eq. (4-4);

$$\begin{aligned}\hat{U}_0(B') &= G(x'_S, y'_T)\left(\exp\left\{i k n'[\overline{R}'_0 - W(x'_S, y'_T)]\right\}\right)/\overline{R}'_0\\ &= \hat{G}(x'_S, y'_T)\left\{\exp[i k n'\overline{R}'_0]\right\}/\overline{R}'_0,\end{aligned} \tag{4-6}$$

where $\hat{G}(x'_S, y'_T)$ represents the product $G(x'_S, y'_T)\ \exp[-i k n' W(x'_S, y'_T)]$. Thus $\hat{G}(x'_S, y'_T)$ is the *pupil function*, and $W(x'_S, y'_T)$ is the *wave aberration function*; both, in general, are functions of the coordinates (x'_S, y'_T).

So far, we have depended on an intuitive understanding of what is meant by a *wave front*. Actually this term could refer to a surface through points in a propagating wave of uniform instantaneous field value; however, unless otherwise stated, a wave front in this book refers to a surface through points of uniform phase. In Fig. 4.1, all points in each wave front are the same number of wavelengths from the object point $\overline{Q}$; however, in Eq. (4–6), distance to a general point on the wave front is measured from the *image* point $\overline{Q}'$, and according to the phase expression in braces, the phase (and, therefore, the number of wavelengths) from $\overline{Q}'$ to the wave front can vary from point to point on the wave front.

The ultimate objective, of course, in setting up the various expressions for the image space wave front at the exit pupil is to find what kind of image of a point object is formed on the image plane as a result of diffraction and aberrations of various kinds. Ideally, the point object at $\bar{Q}$ would form a point image at $\bar{Q}'$; however, to find what actual distribution of light occurs in the vicinity of $\bar{Q}'$ as the result of diffraction, a general point P (Figs. 4.1 and 4.2) is set up on the image plane near $\bar{Q}'$. Aberrations are eliminated from the diffraction study by assuming the wave front to be spherical, that is, that it coincides with the pupil sphere so that $W(x'_S, y'_T) = 0$. The distance from the general point B' to P is R'.

In accordance with the discussion associated with Eq. (3-30), the coordinates of P with reference to $\bar{Q}'$ as the origin would be $(\Delta\xi', \Delta\eta')$, but for convenience ξ_0 will be used for $\Delta\xi'$ and η_0 for $\Delta\eta'$ in the following development.

To find the complex amplitude at P due to the spherical wave passing through the exit pupil, each infinitesimal area on the wave front, functioning as a point source is assumed to radiate a spherical wave (Huygens' wavelet); the total effect at P is the integration of the complex amplitude of all such waves arriving at P but originating from points on the wave front. Appendix C on waves indicates how to set up the integral. The element of area on the wave front is $d\sigma$ and the general point on the wave front is B' in the following expression for the complex amplitude at P from the total area $\mathfrak{A}$ of the wavefront at the exit pupil:

$$\hat{U}_0(\xi_0, \eta_0) = (i/\lambda) \iint_{\mathfrak{A}} \left\{ [\hat{U}_0(B') \exp(-ikn'R')]/R' \right\} d\sigma. \quad (4\text{-}7)$$

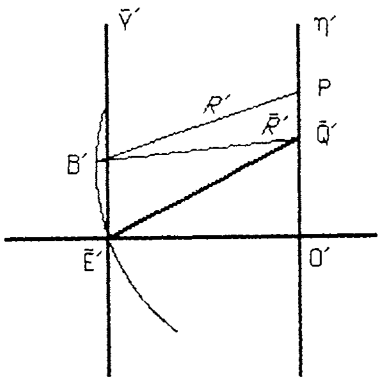


Figure 4.2. Geometry for calculating the electromagnetic disturbance at the image plane produced by a wave front at the exit pupil.

The element of area $d\sigma$ could be more explicitly written $dx'_S\, dy'_T$, or in real-space coordinates:

$$d\sigma = d\bar{X}'\, d\bar{Y}'/(h'_S h'_T \bar{N}'), \tag{4-8}$$

according to Eq. (3-27). By substituting the expression from Eq. (4-4), Eq. (4-7) becomes

$$\hat{U}_0(\xi_0, \eta_0) = (i/\lambda) \iint_a (1/\bar{R}'_0 R')\, \hat{G}(x'_S, y'_T) \exp\left[ikn'(\bar{R}'_0 - R')\right] dx'_S\, dy'_T. \tag{4-9}$$

Although R' is a variable depending on the location of B' on the wave front, the denominator product $R'_0\, R'$ actually varies negligibly compared with the exponential term in the numerator; so it is a good approximation to treat the product as a constant and to combine it with the other constants before the integral signs:

$$\hat{C} = i/(\lambda \bar{R}'_0 R'). \tag{4-10}$$

Even with this simplification, the remaining integral in Eq. (4-9) is still difficult to evaluate because of the form of the variables in the exponential term. This difficulty can be reduced by converting the terms $\bar{R}'_0$ and R' in $(\bar{R}'_0 - R')$ to coordinate expressions and then making further simplifying assumptions.

In the following development, the approximation of $\bar{R}'_0 \cong \bar{R}'$ will be dropped, and $\bar{R}'$ will be indicated throughout. The point B', which terminates the lengths $\bar{R}'$ and R' (Fig. 4.2), has real-space coordinates $(\bar{X}', \bar{Y}', \bar{Z}')$ with the origin of $\bar{Z}'$ at O'. In Fig. 4.2 if $\bar{E}'$ and $\bar{Q}'$ are regarded in the plane of the figure, this has to be the tangential plane in which $\xi' = 0$ and $\bar{X}' = 0$ by definition. However, since B' is a general point on the wave front, it may have a nonzero $\bar{X}'$ coordinate, and since P is a general point on the image plane, it may have a nonzero ξ_0 coordinate. Neither of these two general points need be in the plane of Fig. 4.2. The coordinates of P are $[\xi_0, (\eta' + \eta_0), 0]$; the coordinates of $\bar{Q}'$ are $(0, \eta', 0)$. Having thus set up the coordinates of the end points of $\bar{R}'$ and R', we can apply the Pythagorean theorem to express these two variables in terms of the end-point coordinates:

$$\begin{aligned}\bar{R}'^2 &= \bar{X}'^2 + (\bar{Y}' - \eta')^2 + \bar{Z}'^2 \\ &= \bar{X}'^2 + \bar{Y}'^2 + \eta'^2 - 2\bar{Y}'\eta' + \bar{Z}'^2,\end{aligned} \tag{4-11}$$

$$\begin{aligned} R'^2 &= (\bar{X}' - \xi_0)^2 + (\bar{Y}' - \eta' - \eta_0)^2 + \bar{Z}'^2 \\ &= \bar{X}'^2 + \xi_0^2 - 2\bar{X}'\xi_0 + \bar{Y}'^2 + \eta'^2 + \eta_0^2 \\ &\quad - 2\bar{Y}'\eta' - 2\bar{Y}'\eta_0 + 2\eta'\eta_0 + \bar{Z}'^2. \end{aligned} \tag{4-12}$$

When the terms in Eq. (4-12) are subtracted from those in Eq. (4-11),

$$\bar{R}'^2 - R'^2 = 2(\bar{X}'\xi_0 + \bar{Y}'\eta_0) - (\xi_0^2 + \eta_0^2 + 2\eta'\eta_0). \tag{4-13}$$

Then, if both sides of Eq. (4-13) are divided by the binomial $(\bar{R}' + R')$,

$$\bar{R}' - R' = \left(\frac{\bar{X}'\xi_0 + \bar{Y}'\eta_0}{\bar{R}'} - \frac{\xi_0^2 + \eta_0^2 + 2\eta'\eta_0}{2\bar{R}'}\right)\left(1 - \frac{\bar{R}' - R'}{2R'}\right)^{-1}. \tag{4-14}$$

Because $(\bar{R}' - R') << 2R'$, the value of the second brace on the right side of Eq. (4-14) is very close to unity and does not have to be shown as an explicit factor. With the simplifications indicated, Eq. (4-9) can be written

$$\hat{U}_0(\xi_0, \eta_0) = \hat{C} \iint_{\alpha} \hat{G}(x'_S, y'_T) \exp[\hat{f}(\xi_0, \eta_0)]\, dx'_S dy'_T, \tag{4-15}$$

where

$$\begin{aligned} \hat{f}(\xi_0, \eta_0) &= ikn'\left(\frac{\bar{X}'\xi_0 + \bar{Y}'\eta_0}{\bar{R}'} - \frac{\xi_0^2 + \eta_0^2 + 2\eta'\eta_0}{2\bar{R}'}\right) \\ &= i2\pi\left[\frac{n'\bar{X}'\xi_0}{\lambda\bar{R}'} + \frac{n'\bar{Y}'\eta_0}{\lambda\bar{R}'} - \frac{n'(\xi_0^2 + \eta_0^2 + 2\eta'\eta_0)}{2\lambda\bar{R}'}\right]. \end{aligned} \tag{4-16}$$

As indicated by where it occurs in Eq. (4-15), the expression for $\hat{f}(\xi_0, \eta_0)$ in Eq. (4-16) describes the phase at P. In the second line of Eq. (4-16), the terms in brackets have been so arranged that the third fractional term contains only constants as far as the integration process of Eq. (4-15) is concerned; so this part of the phase expression can be placed before the integration signs in company with $\hat{C}$. Inspection of this constant term indicates that its value is influenced strongly by the positions of the image point $\bar{Q}'$ and the general point P on the image plane. When this part of the phase term in Eq. (4-16) is separated from the rest of $\hat{f}(\xi_0, \eta_0)$, it is given a symbol and written as

$$\hat{f}_p(\xi_0, \eta_0, \eta') = \exp[-i\pi n'(\xi_0^2 + \eta_0^2 + 2\eta'\eta_0)/\lambda\overline{R}']. \tag{4-17}$$

The remaining two fractional terms in brackets in Eq. (4-16) can be written in terms of more convenient variables by referring to equations developed in the previous chapter. By applying Eqs. (3-19), (3-27), (3-33), and (3-40), one can write

$$n'\overline{X}'\xi_0/\lambda\overline{R}' = x'_S u'_S, \tag{4-18}$$

and by applying Eqs. (3-16), (3-27), (3-34), and (3-40), we have

$$n'\overline{Y}'\eta_0/\lambda\overline{R}' = y'_T v'_T, \tag{4-19}$$

where u'_S and v'_T can also be written in terms of certain optical angles as

$$u'_S = (\xi_0/\lambda)(n' \sin \alpha'_S), \tag{4-20}$$

$$v'_T = (\eta_0/\lambda)[n'(\sin \alpha'_T - \sin \overline{\alpha}')]. \tag{4-21}$$

For the relations expressed in Eqs. (4-20) and (4-21), see the discussion in the previous chapter related to Eqs. (3-17), (3-21), (3-41), (3-42), (3-45), (3-46). By incorporating a number of the simplifications discussed subsequent to the writing of Eq. (4-15), it becomes

$$\hat{U}_0(u'_S, v'_T) = \hat{C}\hat{f}_p(\xi_0, \eta_0, \eta') \iint_{\mathfrak{a}} \hat{G}(x'_S, y'_T)$$
$$\exp[i2\pi(x'_S u'_S + y'_T v'_T)]\, dx'_S\, dy'_T. \tag{4-22}$$

Equation (4-22) is the *diffraction integral*. Frequent references are made to it in following chapters.

The product of $\hat{U}_0(u'_S, v'_T)$ and its complex conjugate gives the radiant flux density at a point on the image plane corresponding to (u'_S, v'_T):

$$\mathcal{W}(u'_S, v'_T) = [\hat{U}_0(u'_S, v'_T)][\hat{U}_0^*(u'_S, v'_T)]. \tag{4-23}$$

In Eq. (4-23), flux density is represented by the symbol $\mathcal{W}$.

THE STREHL RATIO

The historical introduction of the Strehl ratio is discussed in the first chapter where the period from 1850 to 1940 is reviewed. This ratio was conceived as an arbitrary figure of merit for highly corrected optical systems. It compares the radiant flux density at the center of the diffraction pattern of a point object with (numerator) and without (denominator) aberrations. The useful range of the Strehl ratio is approximately 0.8 to 1.

The radiant flux density at a point (u'_S, v'_T) near $\overline{Q}'$ can be expressed by combining Eqs. (4-22) and (4-23):

$$\mathcal{W}(\mathrm{u}'_S, v'_T) = |\hat{C}|^2\,|\hat{f}_p|^2 \left| \int\!\!\int_{\mathfrak{a}} \hat{G}(x'_S, y'_T) \exp[i2\pi(x'_S u'_S + y'_T v'_T)]\, dx'_S\, dy'_T \right|^2 . \tag{4-24}$$

Because the Strehl ratio is concerned with only the center of the diffraction pattern, that is, at $\overline{Q}'$ where $u'_S = v'_T = 0$, both $\hat{f}_p$ and the exponential term in the integrand of Eq. (4-24) become unity, and the expression for the radiant flux density becomes

$$\begin{aligned} \mathcal{W}(0, 0) &= |\hat{C}|^2 \left| \int\!\!\int_{\mathfrak{a}} \hat{G}(x'_S, y'_T)\, dx'_S\, dy'_T \right|^2 \\ &= |\hat{C}|^2 \left| \int\!\!\int_{\mathfrak{a}} G(x'_S, y'_T) \exp[-ikn'W(x'_S, y'_T)]\, dx'_S\, dy'_T \right|^2 , \end{aligned} \tag{4-25}$$

where the complex $\hat{G}(x'_S, y'_T)$ has been expanded according to the definition following Eq. (4-6).

In the highly corrected systems for which the Strehl ratio is pertinent, G can be regarded as constant over the whole exit pupil; so when we set up the Strehl ratio, this factor can be written before the integral signs in both numerator and denominator. The only difference, then, between $\mathcal{W}(0, 0)$, the radiant flux density with aberrations, and $\mathcal{W}_1(0, 0)$, the radiant flux density without aberrations, is that the wave-front aberration function $W(x'_S, y'_T)$ occurs as a variable in the exponential term of the first and is identically zero in the second. The

Strehl ratio, therefore, can be written

$$\begin{aligned} R_S &= \mathcal{W}(0,0)/\mathcal{W}_1(0,0) \\ &= \frac{|\hat{C}|^2\, G^2 \left| \iint_{\mathcal{A}} \exp[-ikn'W(x'_S, y'_T)]\, dx'_S\, dy'_T \right|^2}{|\hat{C}|^2\, G^2 \left| \iint_{\mathcal{A}} dx'_S\, dy'_T \right|^2} \\ &= \frac{\left| \iint_{\mathcal{A}} \exp[-ikn'W(x'_S, y'_T)]\, dx'_S\, dy'_T \right|^2}{\mathcal{A}^2}. \end{aligned} \tag{4-26}$$

ANAMORPHOTIC STRETCHING

The reduced coordinates (u'_S, v'_T) in which the diffraction integral and the Strehl ratio have been discussed are convenient in optical analysis because of the normalizing involved in their definitions, but eventually one has to get back to real coordinates on the image plane to appreciate the optical results. The real coordinates, as previously defined, with $\overline{Q}'$ (Figs. 4.1 and 4.2) as the origin, are (ξ_0, η_0), which are identical to $(\Delta\xi', \Delta\eta')$ discussed in Chapter 3.

In going from reduced to real coordinates, it turns out that the two coordinates of a point do not generally convert proportionally; the discrepancy is referred to as *anamorphotic stretching*. For instance, a circular diffraction pattern in reduced coordinates tends to be a regular oval in real coordinates.

To find the relation between reduced and real coordinates, one can form the ratio of corresponding sides of Eqs. (4-20) and (4-21), and with some rearrangement obtain

$$\xi_0/\eta_0 = (u'_S/v'_T)\,[n'(\sin\alpha'_T - \sin\overline{\alpha}')]/(n'\sin\alpha'_S). \tag{4-27}$$

With reference to Eqs. (3-11) and (3-12) of Chapter 3, it is apparent that the trigonometric ratio on the right side of Eq. (4-27) is equivalent to a ratio of numerical apertures, which in turn is equivalent, according to Eqs. (3-41) and (3-42), to a ratio of scaling factors:

$$\xi_0/\eta_0 = (u'_S/v'_T)\,(\mathrm{N.A.})'_T/(\mathrm{N.A.})'_S = (u'_S/v'_T)\,(\rho_T/\rho_S). \tag{4-28}$$

Inasmuch as the aperture ratio ρ_S/ρ_T is normally greater than unity for off-axis points, the diffraction pattern, $\hat{U}_0(u'_S, v'_T)$ or $\mathcal{W}(u'_S, v'_T)$, is stretched—that is, longer—in the tangential (η) direction relative to the sagittal (ξ) direction when conversion is made to real coordinates.

THE PUPIL FUNCTION

The mathematical expression that has already been discussed following Eq. (4-6), which describes an actual wave front as it passes through the exit pupil, is called a *pupil function*. When this function is expressed in normalized coordinates, it is usually understood to have zero value outside the unit-radius circle:

$$\hat{G}(x'_S, y'_T) = G(x'_S, y'_T)\exp[-ikn'W(x'_S, y'_T)] \qquad \text{when} \quad (x'^2_S + y'^2_T) \le 1, \tag{4-29}$$

and

$$\hat{G}(x'_S, \mathrm{y}'_T) = 0 \qquad \text{when} \quad (x'^2_S + y'^2_T) > 1. \tag{4-30}$$

Assumptions connected with the coordinates x'_S and y'_T are reviewed in the discussion associated with Eq. (3-29).

Our earlier discussion of a "perfect" diffraction-limited optical system indicates that in such a system an incident diverging spherical wave front of uniform amplitude is transformed into a converging spherical wave front of uniform amplitude. (This concept can be extended to systems producing virtual images by describing the output as a diverging spherical wave front of uniform amplitude.) When the system is a diffraction-limited system, the complex amplitude $\hat{G}(x'_S, y'_T)$ is a real constant, and the wave aberration function $W(x'_S, y'_T)$ is equal to zero. In the form given in Eqs. (4-29) and (4-30), the pupil function incorporates complete information about the imaging properties of the optical system. The mathematical procedures for arriving at these properties from the pupil function are discussed in later chapters.

In our discussion of the pupil function, it has been tactily assumed that the object is a fixed point in the object plane and that only one frequency (or wavelength) of light is involved. As the subject matter requires, object plane coordinates and other independent variables will be explicitly expressed in addition to the coordinates (x'_S, y'_T) in the exit pupil.

The pupil function as expressed by Eqs. (4-29) and (4-30) remains valid whether the light beam illuminating the object is incoherent, partially coherent, or coherent.

The amplitude function $G(x'_S, y'_T)$ of the pupil function is extremely difficult to measure, but, fortunately, in the highly corrected systems with which we are most concerned, the typical variations in the amplitude have far less effect on the image than the typical variations in phase, which is contained in the wave aberration function $W(x'_S, y'_T)$. The phase characteristics can quite conveniently be measured interferometrically. So, in the absence of a statement to the contrary, the amplitude of the pupil function is assumed constant, and all aberration effects can be attributed to the wave aberration function. Optics conforming to this assumption are sometimes referred to as *Airy systems*. For those who require a pupil function with variable amplitude, Barakat [27] offers an approach for amplitudes having Gaussian-like radial tapers.

THE WAVE ABERRATION FUNCTION

The wave aberration function already discussed in this and preceding chapters has been defined as the optical path length along a ray between the pupil or reference sphere and the actual wave front. In the exponent of the pupil function discussed in the previous section, the function $W(x'_S, y'_T)$ is in terms of the actual physical distance between the wave front and the reference sphere; this distance is converted to the equivalent distance in a vacuum by multiplying by the refractive index n'; and, finally, application of another factor, $\mathscr{k}$, which is 2π divided by the vacuum wavelength, converts the equivalent distance in a vacuum to the number of radians phase shift between the wave front and the reference sphere. Not only does the wave aberration function go by a number of different names in the literature, as mentioned at the beginning of this chapter, but the reference may be to the actual distance, the equivalent distance in vacuum, or the phase shift between the wave front and the reference sphere. So far in this book, the wave aberration function has been indicated as dependent on its position B' at the pupil sphere or on the normalized rectangular coordinates (x'_S, y'_T) of B'. However, as we approach the task of an explicit mathematical expression for $W(x'_S, y'_T)$, the assumed circular aperture dictates polar coordinates (ρ, φ) or, better yet because of symmetry about the tangential plane, $(\rho, \cos\varphi)$.

The wave aberration function has been expressed in a number of different mathematical forms, two of which will be discussed in some detail here.

Early work in geometrical aberrations, especially by L. Seidel [7], was developed in terms of rays rather than wave fronts. However, as the mathematical description of the wave aberration function becomes the vehicle for aberration information, continuity with the older geometric practices is attempted by applying the Seidel classification names to the aberration groupings that emerge from the new mathematics. In a few instances the match is only fair. Seidel's aberration type names are *spherical aberration*, *coma*, *astigmatism*, *Petzval curvature*, and *distortion* (see [1, pp. 220–225; 8, Chapter 6]).

Besides expansion of the wave aberration function in a power series or in Zernike's circle polynomials, both of which are discussed in later sections of this chapter, the following are among the other wave aberration function expansion methods that deserve further study: Tatian [20] and Buchdahl [21] use an expansion of the mixed characteristic of Hamilton; Barakat [19] uses an expansion in Tschebyscheff polynomials; De, Hazra, and Purkait [22] use radial Walsh block functions. Among those who have written about Zernike's circle polynomials are Kintner and Sillitto [10] and Hawkes [23].

POWER SERIES EXPANSION OF THE WAVE ABERRATION FUNCTION

The generally useful and time-honored power series in applied mathematics has become a standard way of expressing the wave aberration function. Hamilton [4] was probably the first to apply the power series to aberration theory; but since his approach involves the characteristic function of Hamiltonian optics [5, 6], which we do not discuss, the power series expansion in this book is based on later work.

In the study of aberrations, the wave aberration function is not only a function of the coordinates (x'_S, y'_T) where the ray passes through the pupil sphere, but is also dependent upon where the object point is located in the object plane. If it is assumed that the aperture is circular and that the system is symmetrical about the axis, a convenient set of coordinates is $(r, \rho, \cos\varphi)$ in which r is the distance from the axis to the object point. By definition, of course, the object point is in the tangential plane, so a second coordinate is not required. Figure 4.3 shows the geometric relation between the rectangular and polar coordinates at the pupil sphere; the consequent expressions relating the independent variables are

$$x'_S = \rho \sin\varphi, \qquad y'_T = \rho \cos\varphi,$$

$$\rho = (x'^2_S + y'^2_T)^{1/2}, \qquad \varphi = \arctan(x'_S/y'_T). \tag{4-31}$$

In our discussions about expansion of the wave aberration function, the polar coordinates and the object plane coordinate are indicated:

$$W = W(r, \rho, \cos\varphi). \tag{4-32}$$

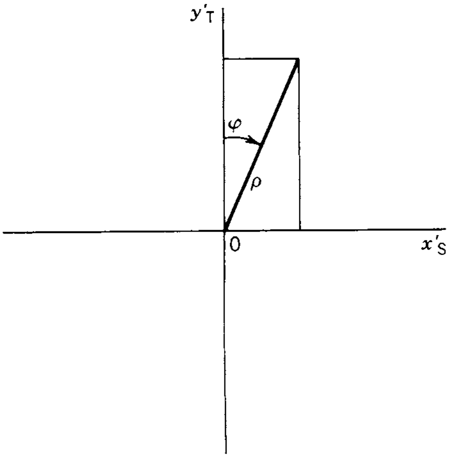


Figure 4.3. Polar coordinates (ρ, φ) on the pupil sphere.

In making up a power series involving r, ρ, and cos φ, all possible powers of each variable separately, products of two variables at the same or different powers, and products of three variables at all combinations of powers have to be considered. However, when the physical characteristics of the optical system to which they refer are taken into account, certain terms are seen to have zero coefficients under all conditions as pointed out in following paragraphs.

When the object point is close to the optic axis ($r = 0$), the r factor has to be omitted to allow the wave aberration function to take on values other than zero over the exit pupil. Furthermore, because of the symmetry that results from the object point on the axis, no variation of the function is possible with change of φ (or cos φ), so the cos φ factor must also be omitted when r is omitted. Finally, symmetry requires that the function be even; that is, $W(\rho) = W(-\rho)$. This means that when ρ is the only variable in a term, it must occur only in the even powers. In the previous chapter, it was shown that a term involving only ρ^2 (called h'^2 in Eq. (3-71)) represents a shift of the image plane, defocusing.

Some authors omit the ρ^2 term altogether in the series representing the wave aberration function because its coefficient would always be zero for a correctly focused system. However, during the design stage, the ρ^2 term is useful to account for a defect of focus. In fact, it is common practice under certain conditions to minimize the coefficient of this term, designated ${}_0C_{20}$ by a scheme discussed later, to find the best focal plane.

Another term, ${}_1C_{11}r\ \rho$ cos φ, which represents a lateral shift of the image point in the image plane, may also be included so that ${}_1C_{11}$ can be minimized to get best focus [11, 12].

As the reader may infer from the two examples already given, each coefficient in the power series is tagged with three subscripts to indicate the exponents on the three possible factors. If a, b, and c represent the three exponents, the correspondence is as follows: ${}_aC_{bc}r^a\,\rho^b\cos^c\varphi$.

Terms in the power series that omit all nonzero powers of ρ would represent only a constant phase increment over the wave front and, therefore, can be left out without any loss of generality.

After all possible terms in the power series are analyzed in terms of the optics system, it is found that the wave aberration function requires only powers of r^2, ρ^2, and the product $r\rho\cos\varphi$:

$$\begin{aligned} W(r,\rho,\cos\varphi) = {} & {}_0C_{00} + {}_1C_{11}r\rho\cos\varphi + {}_0C_{20}\,\rho^2 + {}_2C_{00}r^2 \\ & + {}_4C_{00}r^4 + {}_0C_{40}\,\rho^4 + {}_0C_{60}\rho^6 + \cdots \\ & + {}_1C_{31}r\rho^3\cos\varphi + \cdots \\ & + {}_2C_{22}r^2\,\rho^2\cos^2\varphi + \cdots \\ & + {}_2C_{20}r^2\rho^2 + \cdots \\ & + {}_3C_{11}r^3\rho\cos\varphi + \cdots \qquad (4\text{-}33) \end{aligned}$$

A question might be raised as to why a Z (coordinate in the direction of the optic axis) dependence is not explicitly shown in Eq. (4-33). Actually, the Z coordinate of $W(x'_S, y'_T)$ or $W(\rho, \varphi)$ is fixed by the other two coordinates inasmuch as the coordinate system for the wave aberration function is on the surface of the exit pupil sphere.

By definition of the wave aberration function, it coincides with the pupil sphere at the intersection with the optic axis, that is, where $\rho = 0$; so Eq. (4-33) at the axis would be written

$$\begin{aligned} W_0(r,\rho,\cos\varphi) &= {}_0C_{00} + {}_2C_{00}r^2 + {}_4C_{00}r^4 + \cdots \\ &= 0. \qquad (4\text{-}34) \end{aligned}$$

Since the series must be identically zero for all values of r, it follows that all the indicated coefficients in Eq. (4-34) must each be zero. If the terms indicated in Eq. (4-34) and the optional terms ${}_0C_{20}\rho^2$ and ${}_1C_{11}r\rho\cos\varphi$ (previously discussed in connection with focus) are dropped from the series given in Eq. (4-33), the general term for the power series expansion is

$$ {}_{l+m}C_{n+m,m}r^{l+m}\rho^{n+m}\cos^m\varphi, \qquad (4\text{-}35)$$

in which the terms having one of the three following combinations are omitted:

$$n = m = 0, \tag{4-36}$$

or

$$l = m = 0 \quad \text{and} \quad n = 2, \tag{4-37}$$

or

$$l = n = 0 \quad \text{and} \quad m = 1. \tag{4-38}$$

In general, l, m, and n are either positive integers or zero.

Since the variable r locates the point object in the object plane and the other two variables ρ and $\cos\varphi$, raised to their respective powers, determine the general shape of the wave aberration surface, we can think of $n + m$ and m, the powers of ρ and $\cos\varphi$ (which are the right-hand subscripts of the coefficient), as the indicators of the aberration type. The product of the coefficient and r raised to its power establishes the scale of the general shape. During the computation or the measurement of the optical transfer function, r is fixed and becomes part of the coefficient. The OTF thus determined is for the isoplanatism patch in the vicinity of the circle of r radius in the object plane and is valid only insofar as any variation of r has negligible effect on the OTF.

It is customary to group the terms of the wave aberration function series according to their *order*, which is defined as

$$\text{order} = (\text{sum of the powers of } r \text{ and } \rho) - 1. \tag{4-39}$$

According to this definition, the two optional focus terms, whose coefficients are respectively ${}_0C_{20}$ and ${}_1C_{11}$, would be first-order terms. The next five terms in the series are third-order terms and correspond to the Seidel aberrations:

Spherical aberration	${}_0C_{40}\rho^4$
Coma	${}_1C_{31}r\rho^3 \cos\varphi$
Astigmatism and Petzval curvature	${}_2C_{20}r^2\rho^2 + {}_2C_{22}r^2\rho^2 \cos^2\varphi$
Distortion	${}_3C_{11}r^3\rho \cos\varphi$

There are nine fifth-order terms, and they are usually grouped as follows:

Spherical aberration	${}_0C_{60}\,\rho^6$
Linear coma	${}_1C_{51}r\rho^5 \cos\varphi$
Elliptical coma	${}_3C_{31}r^3\rho^3 \cos\varphi + {}_3C_{33}r^3\rho^3 \cos^3\varphi$
Oblique spherical	${}_2C_{40}r^2\rho^4 + {}_2C_{42}r^2\rho^4 \cos^2\varphi$
Astigmatism and Petzval curvature	${}_4C_{20}r^4\rho^2 + {}_4C_{22}r^4\rho^2 \cos^2\varphi$
Distortion	${}_5C_{11}r^5\rho \cos\varphi$

There are 14 seventh-order terms, 20 ninth-order terms, and, in general, $2 + 3 + \cdots + \frac{1}{2}(N + 3)$ Nth order terms. Table 4.I gives the wave aberration power series terms through the ninth order. The terms are arranged according to their values of m and n as defined in Eq. (4-35).

As one notes the combinations of coefficient subscripts in the aberration groupings already given, it is apparent that other meaningful groupings are possible. For instance,

$$\text{spherical aberration} = {}_0C_{40}\rho^4 + {}_0C_{60}\rho^6 + {}_0C_{80}\rho^8 + \cdots$$

$$\begin{aligned}\text{distortion} &= {}_3C_{11}r^3\rho\cos\varphi + {}_5C_{11}r^5\rho\cos\varphi\\ &\quad + {}_7C_{11}r^7\rho\cos\varphi + \cdots\end{aligned}$$

Although the higher order aberrations must also be considered in highly corrected optical systems, we will confine our discussion in the following sections to just the third-order or Seidel aberrations. Welford [8, Chapters 6 and 7] can be pursued for a more extensive treatment. Also, in keeping with our historical and tutorial objectives, our discussions of aberrations and their correction will involve only simple lenses. The reader will understand that actual aberration correction procedures for multielement systems require complicated computer programming beyond the scope of this book.

As the surface shapes of the wave aberration function for the various aberrations are discussed, it is sometimes helpful in understanding the image defects to appreciate how the shape affects ray direction. As previous discussions of the function $W(r, \rho, \cos\varphi)$ have indicated, the surface shape is relative to the reference sphere at the exit pupil; so, if the function, not the wave front, were "flat," that is, having zero gradient over the whole pupil, the rays, which by definition are perpendicular to the wave front, would all converge radially to the center of the reference sphere and produce the ideal geometrical point image. (The present discussion ignores the spreading effects of diffraction.) As the flat disk is warped by the various aberrations, the slope or gradient of the surface is no longer everywhere zero, and most of the related rays no longer follow radial lines but are deflected from the ideal image point by the amount of the gradient.

SPHERICAL ABERRATION

As indicated earlier, the spherical aberration component of the wave aberration power series is itself an infinite power series in which each term is a positive even power of ρ with its coefficient. The Seidel spherical aberration term is ${}_0C_{40}\rho^4$, which has already been listed as one of the third-order terms. It is also

Table 4.I A Few Terms in the Wave Aberration Function Power Series

n	$m = 0$	$m = 1$	$m = 2$	$m = 3$	$m = 4$	$m = 5$
			Third-Order Aberrations			
0		${}_{3}C_{11}r^3\rho\cos\varphi$	${}_{2}C_{22}r^2\rho^2\cos^2\varphi$			
2	${}_{2}C_{20}r^2\rho^2$	${}_{1}C_{31}r\rho^3\cos\varphi$				
4	${}_{0}C_{40}\rho^4$					
			Fifth-Order Aberrations			
0		${}_{5}C_{11}r^5\rho\cos\varphi$	${}_{4}C_{22}r^4\rho^2\cos^2\varphi$	${}_{3}C_{33}r^3\rho^3\cos\varphi$		
2	${}_{4}C_{20}r^4\rho^2$	${}_{3}C_{31}r^3\rho^3\cos\varphi$	${}_{2}C_{42}r^2\rho^4\cos^2\varphi$			
4	${}_{2}C_{40}r^2\rho^4$	${}_{1}C_{51}r\rho^5\cos\varphi$				
6	${}_{0}C_{60}\rho^6$					
			Seventh-Order Aberrations			
0		${}_{7}C_{11}r^7\rho\cos\varphi$	${}_{6}C_{22}r^6\rho^2\cos^2\varphi$	${}_{5}C_{33}r^5\rho^3\cos^3\varphi$	${}_{4}C_{44}r^4\rho^4\cos^4\varphi$	
2	${}_{6}C_{20}r^6\rho^2$	${}_{5}C_{31}r^5\rho^3\cos\varphi$	${}_{4}C_{42}r^4\rho^4\cos^2\varphi$	${}_{3}C_{53}r^3\rho^5\cos^3\varphi$		
4	${}_{4}C_{40}r^4\rho^4$	${}_{3}C_{51}r^3\rho^5\cos\varphi$	${}_{2}C_{62}r^2\rho^6\cos^2\varphi$			
6	${}_{2}C_{60}r^2\rho^6$	${}_{1}C_{71}r\rho^7\cos\varphi$				
8	${}_{0}C_{80}\rho^8$					
			Ninth-Order Aberrations			
0		${}_{9}C_{11}r^9\rho\cos\varphi$	${}_{8}C_{22}r^8\rho^2\cos^2\varphi$	${}_{7}C_{33}r^7\rho^3\cos^3\varphi$	${}_{6}C_{44}r^6\rho^4\cos^4\varphi$	${}_{5}C_{55}r^5\rho^5\cos^5\varphi$
2	${}_{8}C_{20}r^8\rho^2$	${}_{7}C_{31}r^7\rho^3\cos\varphi$	${}_{6}C_{42}r^6\rho^4\cos^2\varphi$	${}_{5}C_{53}r^5\rho^5\cos^3\varphi$	${}_{4}C_{64}r^4\rho^6\cos^4\varphi$	
4	${}_{6}C_{40}r^6\rho^4$	${}_{5}C_{51}r^5\rho^5\cos\varphi$	${}_{4}C_{62}r^4\rho^6\cos^2\varphi$	${}_{3}C_{73}r^3\rho^7\cos^3\varphi$		
6	${}_{4}C_{60}r^4\rho^6$	${}_{3}C_{71}r^3\rho^7\cos\varphi$	${}_{2}C_{82}r^2\rho^8\cos^2\varphi$			
8	${}_{2}C_{80}r^2\rho^8$	${}_{1}C_{91}r\rho^9\cos\varphi$				
10	${}_{0}C_{100}\rho^{10}$					

referred to as *primary spherical aberration*. The succeeding terms in the series are respectively called *secondary*, *tertiary*, and so forth, spherical aberration.

Primary spherical aberration is plotted against ρ in Fig. 4.4 for an assumed lens. It will be noted that the practice is to use the ordinate rather than the abscissa for the independent variable in plots of this kind. An equal-increment contour plot of the primary spherical aberration for the same lens is shown in Fig. 4.5. The contours are in the unit-radius circle having the coordinates (ρ, φ) with the endwise view of the pupil ray as the center 0. A profile of the wave front and the pupil ray are shown in the tangential plane in Fig. 4.6. Figure 4.7 is a diagram of the primary spherical aberration in three dimensions. This shape, of course, is not quite the same as that of the wave front because it is relative to the reference sphere. Also, the magnitude of the aberration is exaggerated in this and similar diagrams to show its characteristics.

By showing the paths of equally spaced parallel rays in the tangential plane, Fig. 4.8 indicates the effects of spherical aberration for a spherical mirror. Although rays near the optic axis tend to converge to a common point, spherical aberration prevents a true focus. When spherical aberration is present in either a reflective or refractive system designed to produce an image on a plane perpendicular to the optic axis, a screen placed anywhere between the focus of the paraxial rays and the focus of the marginal rays will image an axial point source

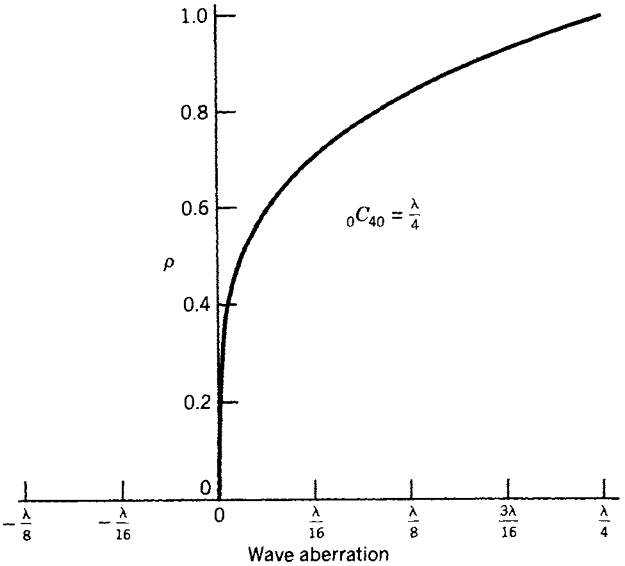


Figure 4.4. Wave aberration as a function of the distance from the pupil ray with one quarter wavelength maximum distortion at the edge of the pupil caused by primary spherical aberration.

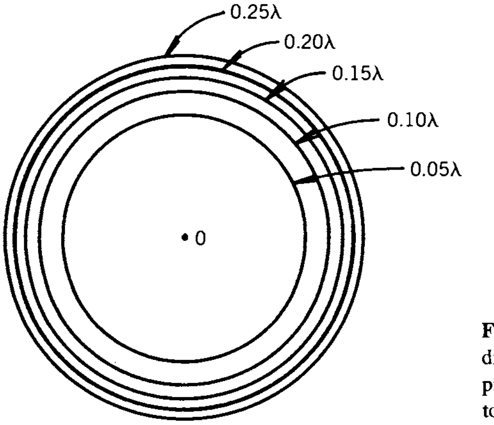


Figure 4.5. Contours of constant wave distortion produced by a lens having only primary spherical aberration with ${}_0C_{40}$ equal to λ/4.

as a circular patch of light. The minimum diameter patch that one finds by analytically or experimentally moving the image plane along the axis is called the *circle of least confusion*, but this location of the image plane is not necessarily the best for images of extended objects, which can be thought of as being made up of a large number of contiguous point sources. In the image, the patches

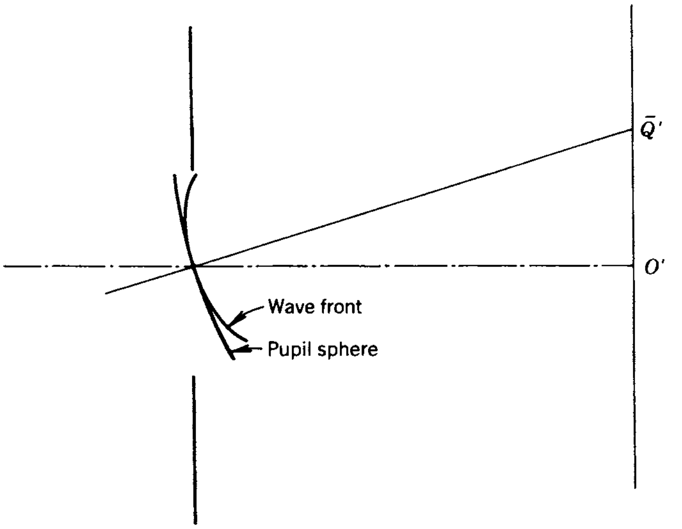


Figure 4.6. The wave-front profile in the tangential plane at the exit pupil for the spherical aberration represented in Figs. 4.4 and 4.5.

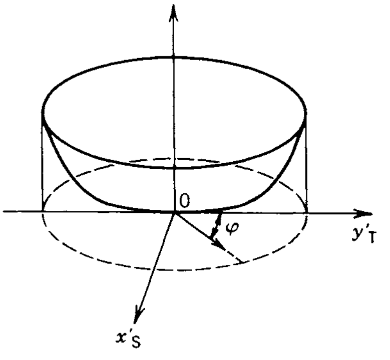


Figure 4.7. Three-dimensional representation of the wave distortion for primary spherical aberration. (Reproduced by permission from M. Born and E. Wolf, *Principles of Optics* (Pergamon Press, Oxford and New York), Third Revised Edition, 1965.)

from the point sources overlap considerably causing outlines and details in the object to be softened in the image; and details that would be smaller in the image than the size of a patch from a point source disappear altogether. The quality of an extended object image depends not only on the patch size from a point source but also on the distribution of light in the patch. The light patch may consist of a bright ring with a fainter center or of a small bright nucleus with a rather tenuous halo. Then, again, it might be a rather sharp central disk with a faint concentric fringe.

Figure 4.9 shows the modulation transfer function of a lens for which the coefficient ${}_0C_{40}$ is equal to one wavelength. The different curves correspond to different locations of the image plane between the marginal focus and the paraxial focus. The curve for the location midway between the foci, $B = -1$, is closest of those plotted to the broken line curve for no wave-front distortion; but the nature of the application would still have to be considered before a final decision could be made as to which image plane location to use. Further discussion of this kind of decision is left to later chapters.

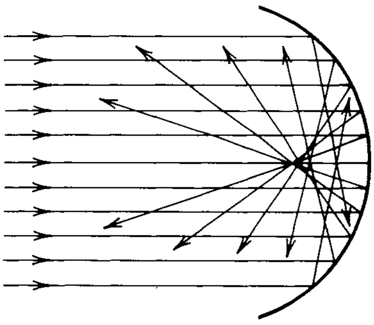


Figure 4.8. Spherical aberration of a spherical mirror.

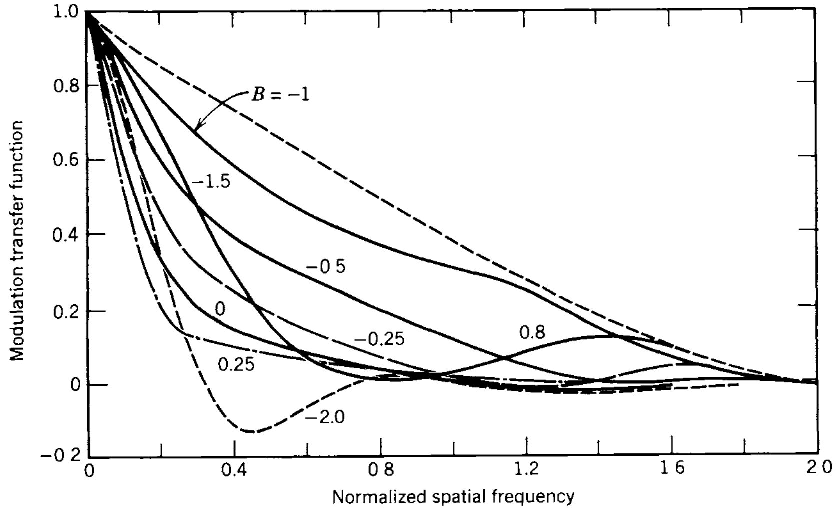


Figure 4.9. Modulation transfer function for a system with one wavelength of aberration, ${}_0C_{40} = \lambda$, caused by third-order spherical aberration. The curves are for different positions of the image plane between the marginal focus and the paraxial focus. The curve labeled $B = -1.0$ is for a setting midway between the two focal planes [15].

Figure 4.10 gives calculated modulation transfer function curves for different amounts of primary spherical aberration, that is, for different values of ${}_0C_{40}$ with the higher order coefficients fixed. For each value of ${}_0C_{40}$, the imaging plane was placed at the position of best low spatial frequency response.

A number of elementary techniques have been used for abating spherical aberration. Since this aberration is attributed to the spherical shape of the optical surfaces involved, it is not surprising that the introduction of one or more appropriately designed aspherical surfaces can bring the aberration within acceptable limits. For instance, the reflectors of astronomical telescopes are often made parabolic for this reason. The price paid is that the heavenly object must be kept very close to the optic axis or the imaging process becomes swamped by coma. When multiple lenses are involved in a system, a combination of positive and negative elements can sometimes end up with considerably less net spherical aberration than a simpler system for the same imaging application. Overcorrection in a negative lens is used to balance the undercorrection in a positive lens. When one has only the two surfaces of a single refractive element to work with, spherical aberration can still be minimized by selecting the best pair of curvatures to attain the desired focal length for the lens. With reference to Fig. 4.11, if a point source is at a great distance to the left and the glass in

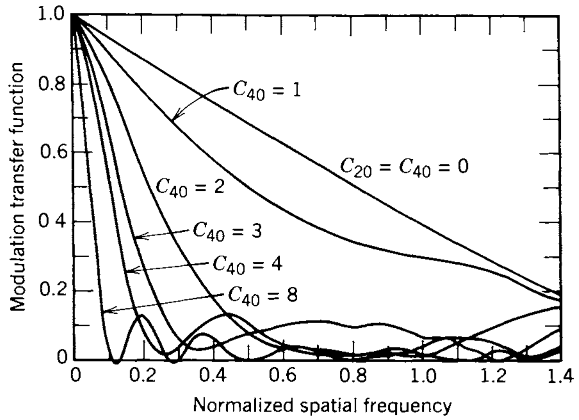


Figure 4.10. Modulation transfer function for an optical system having just primary spherical aberration with different values for ${}_0C_{40}$. For each value, the best image plane was found for low frequency response [16]. (Reproduced by permission of The General Electric Co., Ltd., of England.)

a spherical lens has a refractive index of 1.5, the shape with the least spherical aberration will be closest to the second lens in the bottom row (plano-convex with the bulge toward the object). Strictly speaking, however, it would be classified with the first lens (concavo-convex) with a ratio of radii of about 148 for a focal length of 3.8 in. and a thickness of 0.5 in. A typical focal length characteristic of a corrected lens is shown in Fig. 4.12 where the margin is slightly

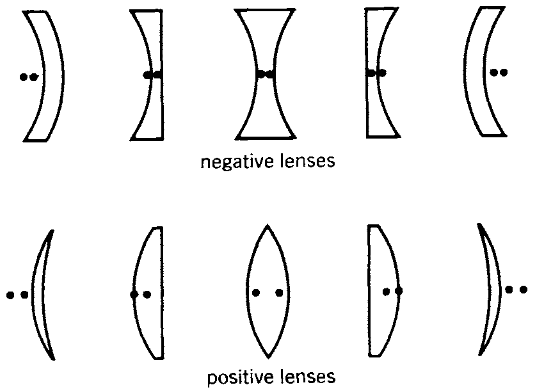


Figure 4.11. Lens shapes.

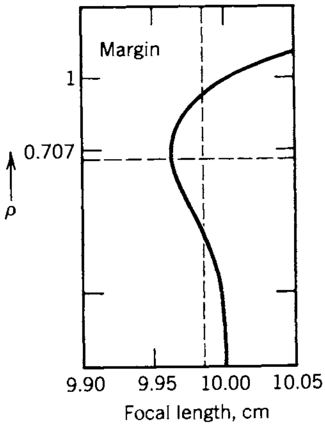


Figure 4.12. A plot of the aberration of a lens having spherical aberration with marginal overcorrection and zonal undercorrection. (From Jenkins and White, *Fundamentals of Optics*, 2d ed. McGraw-Hill Book Company, Inc., New York, 1950.)

overcorrected and the intermediate zone is undercorrected. The reverse is possible but unusual.

A well-corrected lens usually has its primary and higher order spherical aberrations balanced for a specific maximum entrance pupil radius. In cameras and other optical systems, a heavy penalty of missing detail is exacted when the system is forced to larger than design apertures. Figure 4.12 shows how rapidly the focal length deteriorates at the margin for one particular example.

COMA

The Seidel or third-order coma term is ${}_1C_{31}r\rho^3 \cos \varphi$. Because of the factors in this primary coma term, it is obvious that this aberration grows as the object is moved away from the optic axis and that the ray direction for a given object position is strongly dependent upon which zone of the exit pupil transmits the ray. As indicated in the previous section, coma is especially severe for reflecting paraboloids. In a laboratory test of the 200-in. Hale telescope at Mt. Palomar, coma became evident only 1 mm off axis, a field angle of only 13 seconds [14].

Figure 4.13 shows an equal-increment contour plot of primary coma where the maximum wave-front distortion is a quarter wavelength at the edge of the pupil in the tangential direction. As indicated, the aberration is negative in the lower half of the figure where φ ranges from $\pi/2$ to $3\pi/2$ radians and is positive in the upper half. In the sagittal plane (along the x_S'-axis where φ is equal to either $\pi/2$ or $3\pi/2$ radians), the wave front coincides with the pupil sphere.

However, a gradient in the tangential (y'_T) direction exists along this axis of coincidence; so the rays associated with this region of the wave front are not directed toward the Gaussian image point.

Figure 4.14 shows a profile of the coma-distorted wave front in the tangential plane. The radial distance between the pupil sphere and the wave front, the wave aberration function, is shown in three dimensions in Fig. 4.15.

The peculiar effects caused by coma on the image of a point source are shown in Fig. 4.16. The rays associated with each narrow zone of the wave front reach the image plane in a circle rather than at a point. As the zone radius ρ increases, the image circle becomes larger and is displaced more from the axis. In Fig. 4.16*a*, a narrow zone is shown in the wave front, and eight points in the zone are numbered. The rays associated with the zone produce the image shown in Fig. 4.16*b*. Pairs of rays from the eight zone points intersect at the correspondingly numbered points in the image, +1 and −1 rays at 1, +2 and −2 rays at 2, and so on. When the images through the various zones from a point source are superimposed, the image suggested by Fig. 4.16*c* results. Since the wave front is continuous rather than made up of discrete zones, the circles actually blend together forming a comet-shaped image of the point source. Because of

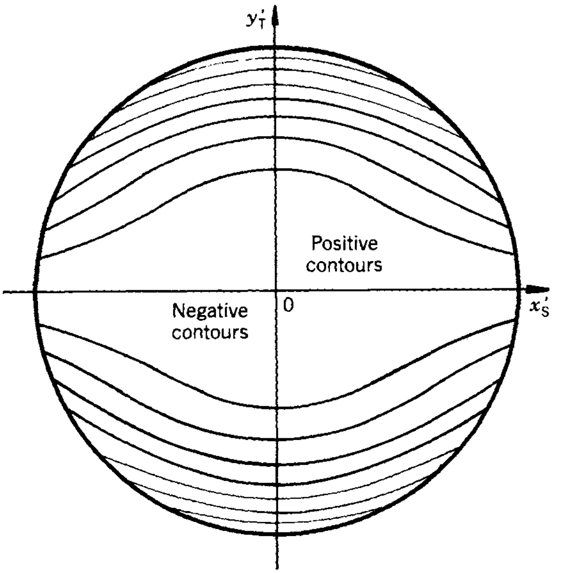


Figure 4.13. Contours of constant wave-front distortion for a lens having just primary coma with the coefficient ${}_1C_{31}$ equal to $\lambda/4$.

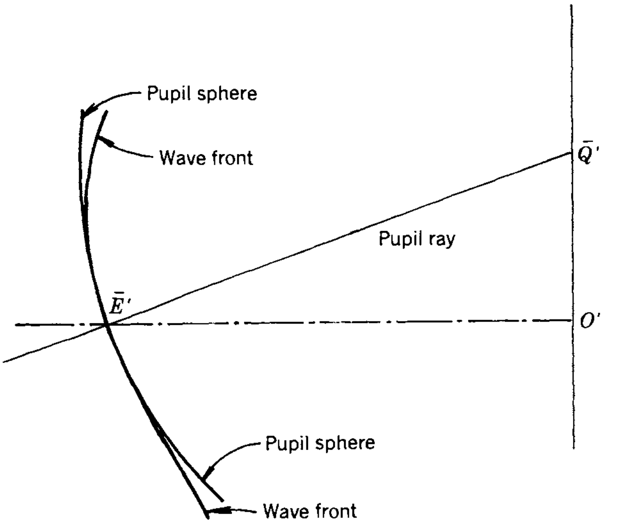


Figure 4.14. Profile of the wave front in the tangential plane showing the distortion produced by primary coma.

the r factor in the third-order coma term, the image grows as the object is moved away from the axis.

If crossed lines are used to locate an off-axis point in the object, the intersection can be located quite accurately in spite of considerable spherical aberration; but because of its asymmetrical nature, the blurring caused by coma can lead to large location errors.

Just as with spherical aberration, coma can be weakened in a simple lens by selecting the best pair of curvatures to provide a required focal length. Fortunately, the optimum combination for cropping coma is very close to that required for cropping spherical aberration. For relatively large apertures, higher

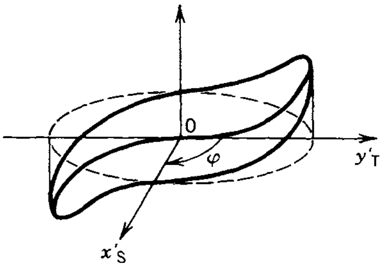


Figure 4.15. A three-dimensional representation of the aberration function produced by primary coma. (Reproduced by permission from M. Born and E. Wolf, *Principles of Optics* (Pergamon Press, Oxford and New York), Third Revised Edition, 1965.)

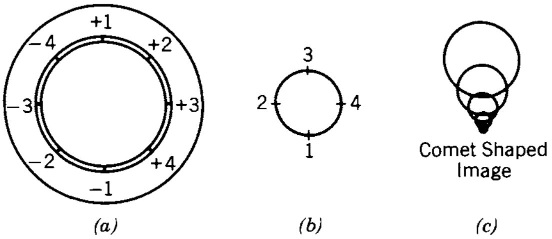


Figure 4.16. Formation of comet-shaped image of a point object. (From Jenkins and White, *Fundamentals of Optics*, 2d ed. McGraw-Hill Book Company, Inc., New York, 1950.)

order coma may become significant even though third-order coma has been softened. However, if a relation known as the *sine condition* can be realized for all zones, all coma must be absent [1, pp. 167–169; 8, p. 155].

Coma is minimized if the optical system can be made symmetrical about the aperture stop and the lateral magnification made unity. In the absence of spherical aberration, primary coma does not depend on the aperture position; but in

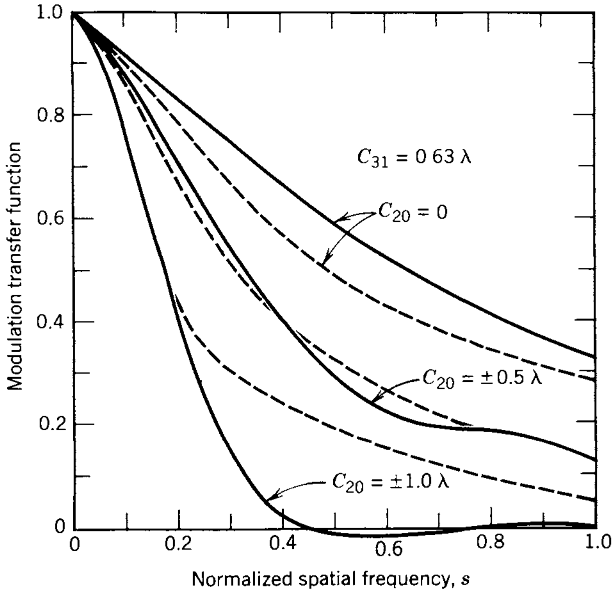


Figure 4.17. Modulation transfer function for a lens having just primary coma with ${}_1C_{31}$ equal to 0.63λ. The curves are for different amounts of defocusing. Solid lines are for $\psi = \pi/2$ and dashed lines for $\psi = 0$, where ψ is the angle of the line structure to the meridian plane as shown in Fig. 5.4 [17]. (Reproduced by permission of the General Electric Co., Ltd., of England.)

the presence of spherical aberration, a stop position can be found where primary coma vanishes, given complete freedom of movement. This procedure, of course, is of limited usefulness because of practical restrictions on stop position. In fact, when a high-quality optical system is designed, usually by computer methods, pat techniques to eliminate coma or any other single aberration are rarely successful; coexisting aberrations have to be treated jointly as suggested in later chapters.

Figures 4.17–4.19 show the calculated MTF for a lens having primary coma and differing amounts of defocusing. In the three figures, the coefficient ${}_1C_{31}$ has the values 0.63λ, 1.26λ and 1.89λ, respectively. Defocusing is measured from the paraxial image plane.

ASTIGMATISM

In the power series representing the wave aberration function, the term ${}_2C_{22}r^2\rho^2 \cos^2 \varphi$ is identified as the Seidel or third-order astigmatism term. The contour

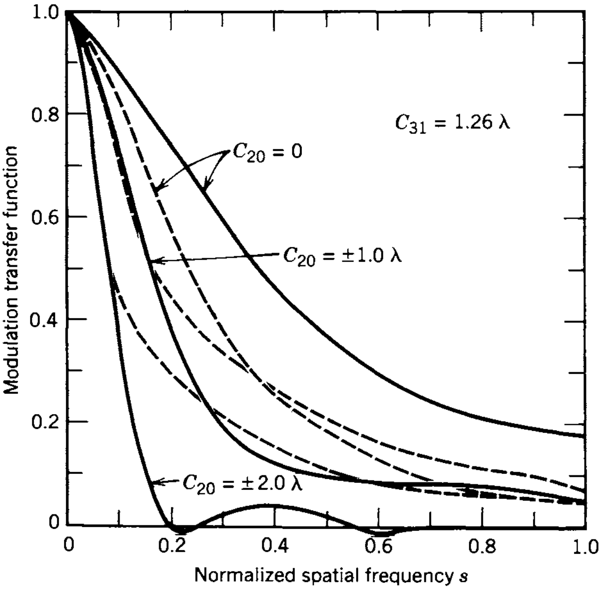


Figure 4.18. Modulation transfer function for a lens having just primary coma with ${}_1C_{31}$ equal to 1.26λ. The curves are for different amounts of defocusing. Solid lines are for $\psi = \pi/2$ and dashed lines for $\psi = 0$, where ψ is the angle of the line structure to the meridian plane as shown in Fig. 5.4 [17]. (Reproduced by permission of the General Electric Co., Ltd., of England.)

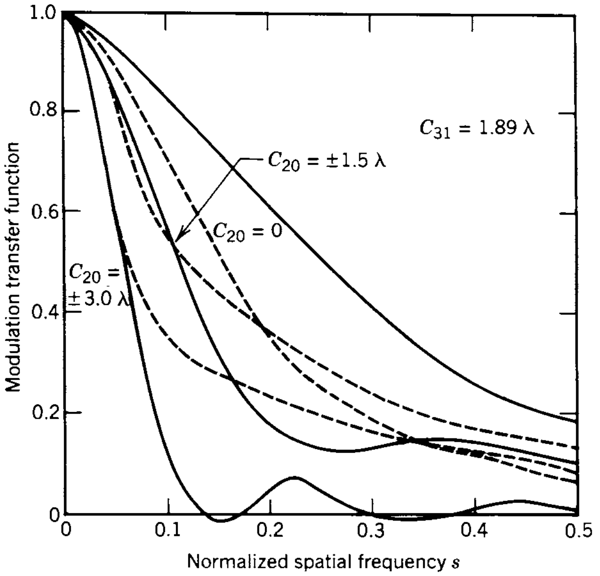


Figure 4.19. Modulation transfer function for a lens having just primary coma with ${}_1C_{31}$ equal to 1.89λ. The curves are for different amounts of defocusing. Solid lines are for $\psi = \pi/2$ and dashed lines for $\psi = 0$, where ψ is the angle of the line structure to the meridian plane as shown in Fig. 5.4 [17]. (Reproduced by permission of the General Electric Co., Ltd., of England.)

plot for primary astigmatism, Fig. 4.20, resembles the coma plot of Fig. 4.13 in some respects. The wave front and the reference sphere coincide along the x'_S-axis, and all the contour lines are perpendicular to the y'_T-axis. However, the contour lines for astigmatism are straight rather than curved, and the contours are positive relative to the x'_S-axis both above and below this axis. The profile of a wave front with astigmatism shown in the tangential plane (y'_T-axis of Fig. 4.20) is given in Fig. 4.21, and the wave aberration function with only primary astigmatism is shown in three dimensions in Fig. 4.22. A three-dimensional ray diagram is drawn in Fig. 4.23 to show image formation by a simple lens with primary astigmatism. The off-axis point source in the object plane is shown emitting four rays, two in the tangential plane and two in the sagittal plane, and the plane areas are indicated between the rays and the optic axis. When all rays from the point object are considered, they are found to focus in a line perpendicular to the tangential plane at the tangential focus and in a line perpendicular to the sagittal plane at the sagittal focus. In the region between the two foci, an image plane shows an elliptical spot, which becomes circular somewhere near the center of the region and collapses to a line as the image plane is moved to either focus.

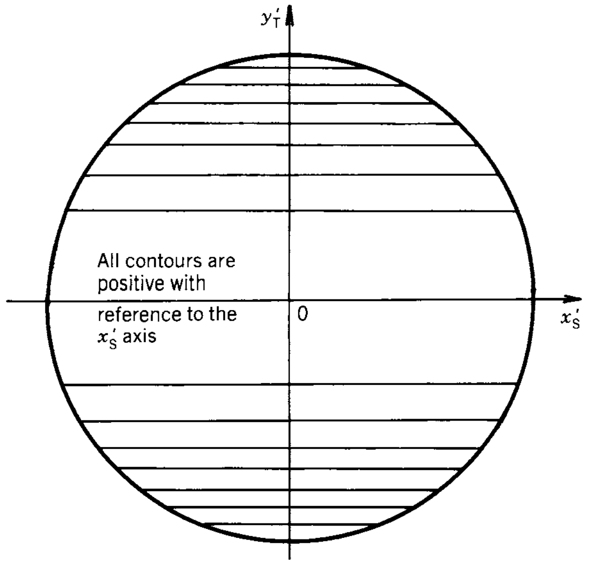


Figure 4.20. Contours of constant wave-front distortion for a lens having just primary astigmatism.

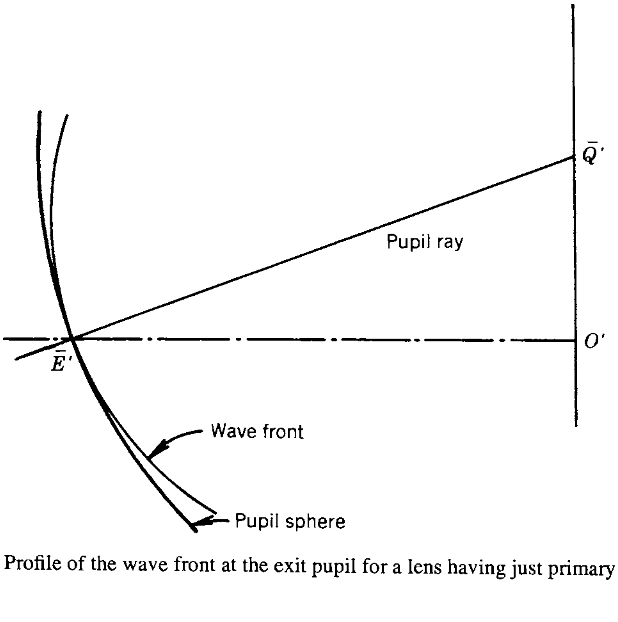


Figure 4.21. Profile of the wave front at the exit pupil for a lens having just primary astigmatism.

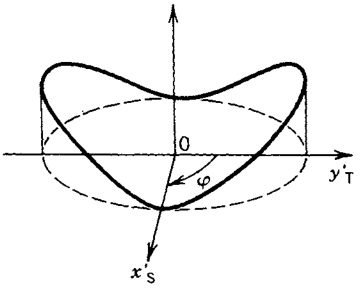


Figure 4.22. Three-dimensional representation of the aberration function for a lens having just primary astigmatism. (Reproduced by permission from M. Born and E. Wolf, *Principles of Optics* (Pergamon Press, Oxford and New York), Third Revised Edition, 1965).

Since the factor r^2 causes the astigmatic distortion of the wave front to increase as the square of the object point displacement from the axis, one would expect that aberration effects in the image would become more pronounced as the point source is moved away from the axis. Evidence of this is illustrated in Fig. 4.24 in which all points of the object plane are imaged on the right side of the lens. As in the imaging of a single point, two different sets of images result, now represented by two surfaces rather than two lines. With increasing r, the foci come closer to the lens and their separation increases. Both surfaces are paraboloidal, and they intersect the optic axis at the paraxial image point. Making the two surfaces coincide amounts to eliminating astigmatism.

As one might predict from our imaging discussion, astigmatism allows concentric circles on the object plane, centered on the optic axis, to be sharply imaged on the tangential image surface whereas radial lines are sharply imaged

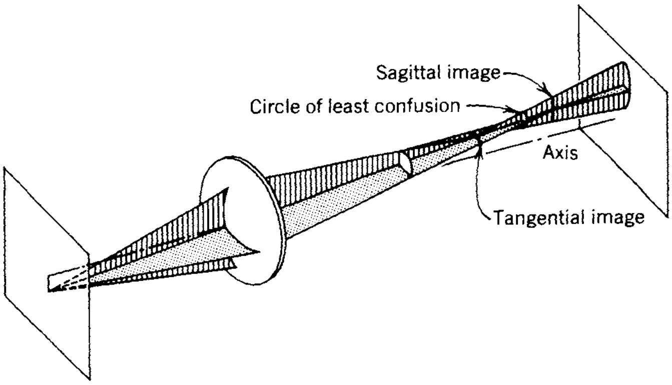


Figure 4.23. Formation of tangential and sagittal images in astigmatism.

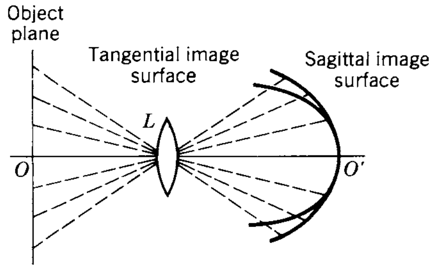


Figure 4.24. Astigmatic images of a plane object.

on the sagittal image surface. An elementary drawing of a wheel (Fig. 4.25) is a good object to illustrate these imaging characteristics. In the tangential image, Fig. 4.25*a*, both the rim and the hub are sharp, but the spokes get increasingly fuzzy as one goes outward from the hub. On the other hand, in the sagittal image, Fig. 4.25*b*, the spokes are sharp, but the rim is uniformly fuzzy. The hub circles, being close to the axis, remain sharp because of the low value of r^2. Astigmatism is a function of both the lens shape and the position of the aperture stop.

Figure 4.26 shows calculated MTF curves for two optical systems with different amounts of astigmatism specified by the constants p and q. These constants are defined in Chapter 9 where a detailed procedure for calculation is presented, and the numerical values for these MTF curves are given in Table 9.II. The OTF for one of these systems is plotted on the complex plane in Fig. 4.27. This method of presentation, called an *Argand diagram*, is discussed in connection with Fig. 5.11.

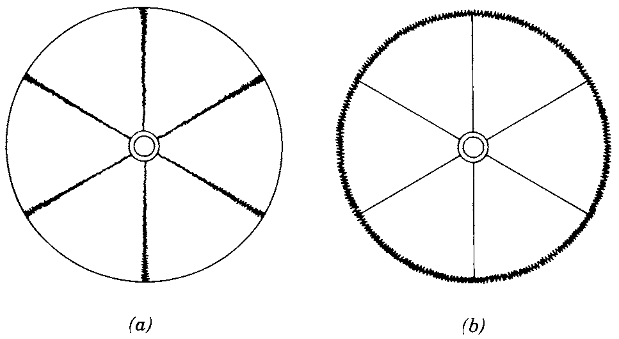


Figure 4.25. Astigmatic images of a spoked wheel: (*a*) tangential image and (*b*) sagittal image.

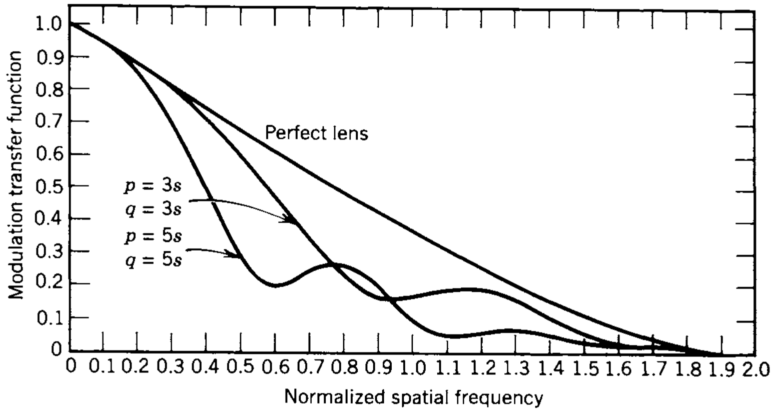


Figure 4.26. Calculated modulation transfer function for two assumed optical systems having different amounts of astigmatism compared with a "perfect" MTF. (See Table 9.II.)

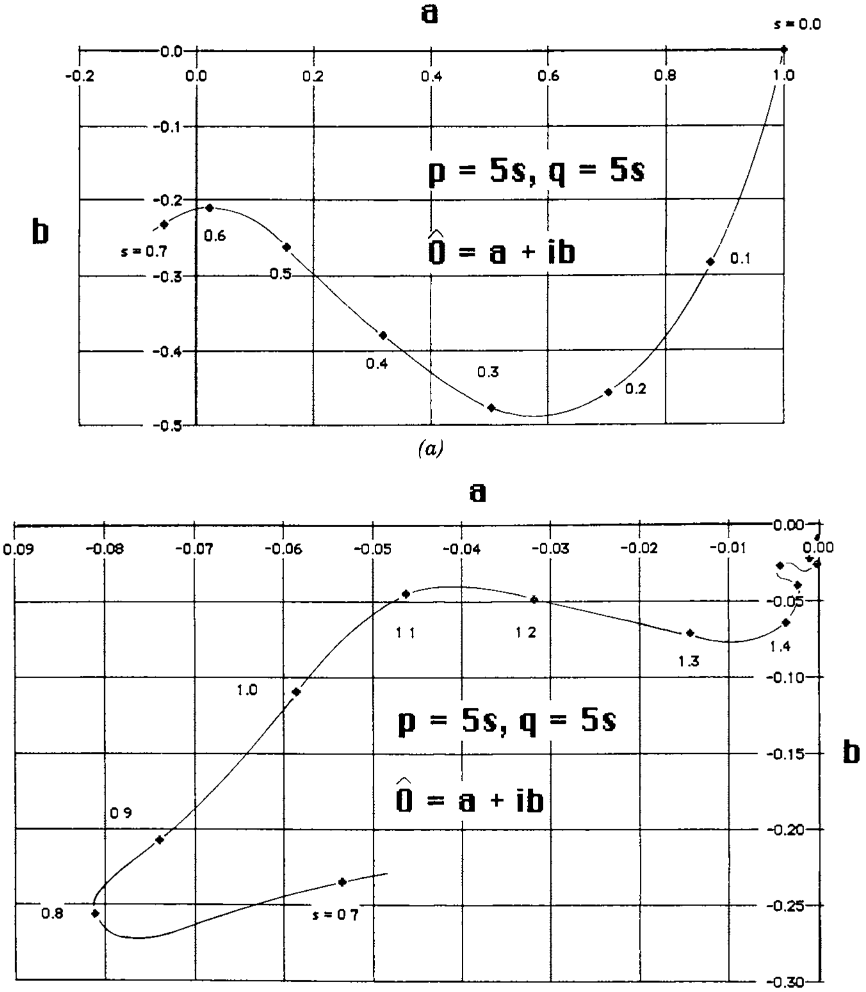
CURVATURE OF FIELD

Every optical system, without specific correction, has curvature of the image "plane." Field curvature, called Petzval curvature, is represented in the power series for the wave aberration function by the term ${}_2C_{20}r^2\rho^2$. In the absence of astigmatism, the sagittal and tangential surfaces coincide and lie on the Petzval surface. A three-dimensional plot of the Petzval curvature term is shown in Fig. 4.28.

Field curvature is especially objectionable in cameras, enlargers, and projectors because the film plane and the projection screen are typically flat. Correcting for curvature of field is referred to as "field flattening." Positive lenses introduce inward curvature of the Petzval surface (undercorrection), and negative lenses introduce outward curvature (overcorrection).

DISTORTION

If all other aberrations are eliminated and only the term ${}_3C_{11}r^3\rho \cos \varphi$ remains in the power series for the aberration function, the corresponding image is sharply defined; but as one examines the image positions of points farther and farther away from the axis, it is discovered that they are displaced more and more from their ideal positions because of the r^3 factor. This aberration is simply labeled *distortion*. Under its influence, images of straight lines that pass



(a)

(b)

Figure 4.27. The complex optical transfer function of one of the systems of Fig. 4.26 plotted on the complex plane (Argand diagram).

through the optic axis remain straight, but all other straight lines in the object produce curved images. The three-dimensional plot of the distortion term is shown in Fig. 4.29. Its effect on a uniform mesh diagram in the object plane can be seen in Fig. 4.30. Where distortion operates to move points more outward with increasing r, the pincushion effect of Fig. 4.30*a* results. On the other

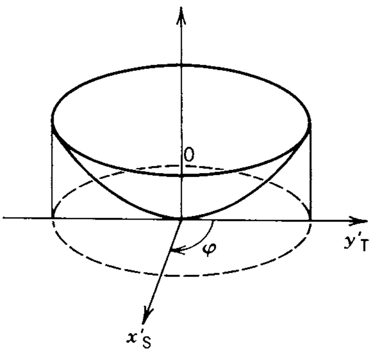


Figure 4.28. Three-dimensional representation of the aberration function for a lens having primary field curvature. (Reproduced by permission from M. Born and E. Wolf, *Principles of Optics* (Pergamon Press, Oxford and New York), Third Revised Edition, 1965.)

hand, if the aberration pulls the image points more and more inward from their proper location with increasing r, barrel distortion, as in Fig. 4.30*b*, results. Magnification is said to increase with r in pincushion distortion and to decrease with increasing r in barrel distortion.

An aperture stop located between a positive lens and the image increases pincushion distortion, and a stop on the side of the lens remote from the image increases barrel distortion. Distortion is minimized by making systems symmetrical about the aperture stop.

In summary, the first three of the primary or third-order aberrations—spherical aberration, coma, and astigmatism—cause lack of sharpness in the image. The last two—Petzval or field curvature and distortion—cause geometrical distortion of the image.

EXPANSION OF THE WAVE ABERRATION FUNCTION IN ZERNIKE POLYNOMIALS

Parallel to the expansion of the wave aberration function in a power series is the newer practice of expanding the function in a series of orthogonal polyno-

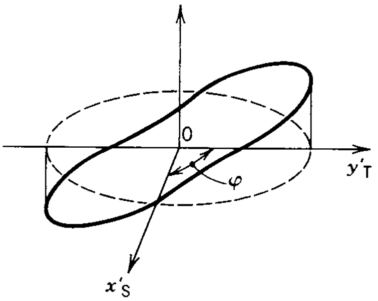


Figure 4.29. Three-dimensional representation of the aberration function for a lens having primary distortion. (Reproduced by permission from M. Born and E. Wolf, *Principles of Optics* (Pergamon Press, Oxford and New York), Third Revised Edition, 1965.)

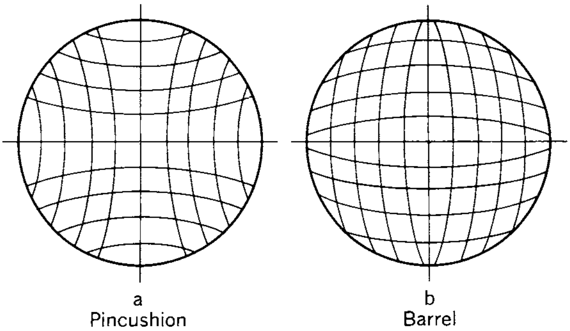


Figure 4.30. Distortion of a square mesh object pattern.

mials, particularly Zernike circle polynomials. Besides the introductory papers by Zernike and Nijboer [9, 24] and the references already mentioned by Kintner and Sillitto and by Hawkes, our necessarily brief treatment of Zernike polynomials can be extended by consulting Born and Wolf [1, pp. 464–468 and Appendix VII], the further discussions by Kintner [25, 26], and the article by Kim and Shannon [28].

Expansion in Zernike circle polynomials provides advantages in calculating diffraction integrals, arriving at the optical transfer function, and balancing aberrations. In the last application, effectiveness is greatest for very small aberrations where the objective is to maximize the Strehl ratio, which requires that the wave-front distortion be under one wavelength for the balancing technique to work.

Before defining the sets of functions known as the Zernike circle polynomials and Zernike radial polynomials, it is of interest to discuss the characteristics that would be desirable in functions used to build series equivalents of the wave aberration function.

The region of the wave aberration function, as we have defined it in previous discussions, is a circle of unit radius with rectangular coordinates (x_S', y_T') or polar coordinates $(\rho, \cos\varphi)$ in which $\cos\varphi$ is used instead of φ to take advantage of simplifications resulting from the symmetry about the tangential plane (y_T'-axis) (see Fig. 4.3). For the present discussion, the point object is considered fixed; so r, the coordinate of the point in the object plane, is not involved here as it was in the expansion in a power series.

Whatever set of functions is defined to serve as terms in a series expansion of the aberration function, it is apparent that the definition of the set must be broad enough to include all the kinds of terms necessary to represent any reasonably well-behaved function. Such a set is called *complete* in the jargon of mathematics.

Another property given the Zernike circle polynomials is *invariance in form*, which means that when the coordinate system is rotated, in general, as follows,

$$\begin{aligned} x' &= x \cos \phi + y \sin \phi, \\ y' &= -x \sin \phi + y \cos \phi, \end{aligned} \tag{4-40}$$

each polynomial $V(x, y)$ is transformed into a polynomial of the same form; that is, V satisfies the following relation under the rotation:

$$V(x, y) = G(\phi)\, V(x', y'), \tag{4-41}$$

where $G(\phi)$ is a continuous function with period 2π radians of the angle of rotation ϕ, and $G(0) = 1$.

A property that gives the Zernike circle polynomials advantages in mathematical manipulation is *orthogonality*. A system of functions $f_n(x)$, defined in the interval $a \leq x \leq b$ and continuous in it, is said to be pairwise orthogonal if

$$\int_a^b f_n(x) f_{n'}(x)\, dx = 0 \qquad \text{if } n \neq n'. \tag{4-42}$$

However, it is always assumed that

$$\int_a^b f_n^2(x)\, dx > 0. \tag{4-43}$$

The form of the general Zernike circle polynomial useful for the kind of optical system discussed in this chapter is $A_{mn} R_n^m(\rho) \cos m\varphi$, in which A_{mn} is the coefficient, $R_n^m(\rho)$ is an nth degree polynomial in ρ which contains no powers of ρ less than m, and the third factor is the cosine of the multiple angle $m\varphi$. As in the power series for the wave aberration function, ρ and φ are the polar coordinates for the exit pupil reference sphere. The positive integers m and n have values that are restricted as follows: $n > m$; $n - m$ is always even.

The Zernike circle polynomials exhibit their orthogonal property in the following equation:

$$\int_0^{2\pi} \int_0^1 [R_n^m(\rho) \cos m\varphi][R_{n'}^{m'}(\rho) \cos m'\varphi] \rho\, d\rho\, d\varphi$$

$$= \begin{cases} \dfrac{\pi}{2n + 2} & \text{for } n = n' \text{ and } m = m', \\ 0 & \text{for either or both } n \neq n' \text{ and } m \neq m'. \end{cases} \tag{4-44}$$

The factor $R_n^m(\rho)$ by itself is called a Zernike radial polynomial; the members of this set are also orthogonal functions:

$$\int_0^1 [R_n^m(\rho)][R_{n'}^{m'}(\rho)]\rho\, d\rho = \begin{cases} \dfrac{1}{2(n+1)} & \text{for } n = n', \\ 0 & \text{for } n \neq n'. \end{cases} \tag{4-45}$$

The formulas for the Zernike radial polynomials are

$$\begin{aligned} R_n^m(\rho) &= \frac{1}{[(n-m)/2]!\rho^m}\left\{\frac{d}{d(\rho^2)}\right\}^{(n-m)/2}\left\{(\rho^2)^{(n+m)/2}(\rho^2-1)^{(n-m)/2}\right\} \\ &= \sum_{s=0}^{(n-m)/2}(-1)^s\frac{(n-s)!}{s!\,[(n+m)/2-s]!\,[(n-m)/2-s]!}\rho^{(n-2s)}. \end{aligned} \tag{4-46}$$

These tedious formulas are often bypassed in practice by either using a *generating function* [1] or applying a *recurrence relation* that allows one to calculate a Zernike radial polynomial from two other Zernike radial polynomials, all of different degree n [26]. Table 4.II gives Zernike radial polynomials for m and n up through 8. An easily verified property is that $R_n^m(1) = 1$.

Once calculated, the Zernike polynomials are assembled in the series for $W(\rho, \varphi)$:

$$W(\rho, \varphi) = A_{00} + \frac{1}{\sqrt{2}}\sum_{n=2}^{\infty} A_{n0}R_n^0(\rho) + \sum_{n=1}^{\infty}\sum_{m=1}^{n} A_{nm}R_n^m(\rho)\cos m\varphi. \tag{4-47}$$

As indicated in earlier discussion, the coefficients A_{nm} are functions of r, the coordinate of the point object in the object plane. In the second term, $1/\sqrt{2}$ is factored out of A_{n0} to simplify derived formulas.

Although the third-order terms of the power series representation of the aberration function could be correlated quite closely with the historically important Seidel aberrations, the terms of the Zernike polynomial expansion do not match up well with the power series and Seidel classifications. One can appreciate why there are discrepancies by reviewing the relations between the multiple-angle cosine factors of the Zernike terms with the power cosine terms of the other series. For instance, from the following trigonometric identities, it is evident that a number of terms in one series would contribute to a single term

Table 4.II Zernike Radial Polynomials $R_n^m(\rho)$

m	n								
	0	1	2	3	4	5	6	7	8
0	1		$2\rho^2 - 1$		$6\rho^4 - 6\rho^2 + 1$		$20\rho^6 - 30\rho^4 + 12\rho^2 - 1$		$70\rho^8 - 140\rho^6 + 90\rho^4 - 20\rho + 1$
1		ρ		$3\rho^3 - 2\rho$		$10\rho^5 - 12\rho^3 + 3\rho$		$35\rho^7 - 60\rho^5 + 30\rho^3 - 4\rho$	
2			ρ^2		$4\rho^4 - 3\rho^2$		$15\rho^6 - 20\rho^4 + 6\rho^2$		$56\rho^8 - 105\rho^6 + 60\rho^4 - 10\rho^2$
3				ρ^3		$5\rho^5 - 4\rho^3$		$21\rho^7 - 30\rho^5 + 10\rho^3$	
4					ρ^4		$6\rho^6 - 5\rho^4$		$28\rho^8 - 42\rho^6 + 15\rho^4$
5						ρ^5		$7\rho^7 - 6\rho^5$	
6							ρ^6		$8\rho^8 - 7\rho^6$
7								ρ^7	
8									ρ^8

in the other:

$$\cos 2\varphi = 2\cos^2\varphi - 1,$$

$$\cos 3\varphi = 4\cos^3\varphi - 3\cos\varphi,$$

$$\cos 4\varphi = 8\cos^4\varphi - 8\cos^2\varphi + 1,$$

$$\cos 5\varphi = 16\cos^5\varphi - 20\cos^3\varphi + 5\cos\varphi, \ldots \qquad (4\text{-}48)$$

Although the new groupings of image errors in the Zernike polynomial expansion of the aberration function suggests that different classifications with new names might be desirable, the terminology listed earlier in this chapter for the third-order and the fifth-order terms of the power expansion still prevails in most discussions.

As indicated earlier, the Zernike polynomials have a built-in capacity for balancing aberrations to maximize the intensity at the Gaussian focus. For instance, $R_6^0(\rho) = 20\rho^6 - 30\rho^4 + 12\rho^2 - 1$ is recognized as a combination of third-order spherical aberration, fifth-order spherical aberration, and defocusing—according to the established nomenclature. A remarkable fact is that the coefficients of the three powers of ρ are in precisely the correct ratio to achieve maximum intensity. This useful characteristic is discussed in a later chapter.

As one considers the integers m and n as they occur in the Zernike circle polynomials, it is apparent that the main features of the aberration contour diagram such as Figs. 4.5, 4.13, and 4.20, symmetry, for instance, depend on m while details depend both on m and n. This suggests that in a new classification system, the value of m could determine the general type of aberration. On this basis, by analogy with the Seidel aberrations, the terms with $m = 0$ could be called spherical aberration; $m = 1$, coma; and $m = 2$, astigmatism. In such a system, curvature and distortion would appear as degenerate cases of spherical aberration and of coma. Terms with $m \geq 3$ have no parallel in the Seidel system. The value of n could translate into "primary," "secondary," and so on. For instance, $\rho^2 \cos 2\varphi$ would be primary astigmatism and $\rho^4 \cos 2\varphi$ would be secondary astigmatism.

REFERENCES

1. M. Born and E. Wolf, *Principles of Optics*, 3d ed. Pergamon, Oxford, 1965.
2. C. S. Williams and O. A. Becklund, *Optics: A Short Course for Engineers and Scientists*. Wiley-Interscience, New York, 1972, pp. 100–104.

3. W. T. Cathey, *Optical Information Processing and Holography.* Wiley, New York, 1974, pp. 7–13.
4. A. W. Conway and J. L. Synge (Eds.), *The Mathematical Papers of Sir W. R. Hamilton*, Vol. 1. Cambridge Univ. Press, London, 1931.
5. H. A. Buchdahl, *An Introduction to Hamiltonian Optics*. Cambridge Univ. Press, London 1970.
6. O. N. Stavroudis, *The Optics of Rays, Wavefronts, and Caustics*. Academic, New York, 1972, Chapter 12.
7. A number of significant papers by L. Seidel developing his aberration coefficients were published between the years 1852 to 1856 in the *Astronomische Nachrichten*; for example, see L. von Seidel, Sur Dioptrik. *Astron. Nachr.* **43,** 289, 304, and 321 (1856).
8. W. T. Welford, *Aberrations of the Symmetrical Optical System*. Academic, New York, 1974.
9. F. Zernike, Beugungstheorie des Schneidenver-Fahrens und Seiner Verbesserten Form, der Phasenkontrastmethode. *Physica* **1,** 689 (1934).
10. E. C. Kintner and R. M. Sillitto, A New "Analytic" Method for Computing the Optical Transfer Function. *Opt. Acta* **23,** 607 (1976).
11. H. H. Hopkins, The Use of Diffraction-Based Criteria of Image Quality in Automatic Optical Design. *Opt. Acta* **13,** 343 (1966).
12. R. S. Longhurst, *Geometrical and Physical Optics*. Wiley, New York, 1967.
13. H. H. Hopkins, *Wave Theory of Aberrations*. Oxford Univ. Press, Oxford, 1950.
14. I. S. Bowen, Optical Problems at Palomar Observatory. *J. Opt. Soc. Am.* **42,** 795 (1962).
15. G. Black and E. H. Linfoot, Spherical Aberration and the Information Content of Optical Images. *Proc. R. Soc. London Ser. A* **239,** 522 (1957).
16. A. M. Goodbody, The Influence of Spherical Aberration on the Response Function of an Optical System. *Proc. Phys. Soc. (London) Ser. B* **72,** 411 (1958).
17. A. M. Goodbody, The Influence of Coma on the Response Function of an Optical System. *Proc. Phys. Soc. (London) Ser. B* **75,** 667 (1960).
18. M. De, The Influence of Astigmatism on the Response Function of An Optical System. *Proc. R. Soc. London Ser. A* **233,** 91 (1955).
19. R. Barakat, Computation of the Transfer Function of an Optical System from the Design Data for Rotationally Symmetric Aberrations, I. Theory. *J. Opt. Soc. Am.* **52,** 985 (1962).
20. B. Tatian, Aberration Balancing in Rotationally Symmetric Lenses. *J. Opt. Soc. Am.* **64,** 1083 (1974).
21. Reference 5, pp. 265–268.
22. M. De, L. N. Hazra, and P. K. Purkait, Walsh Functions in Lens Optimizations, I. FEE-Based Criterion. *Opt. Acta* **25,** 573 (1978).
23. P. W. Hawkes, The Diffraction Theory of the Aberrations of Stigmatic Ortho-

morphic Optical or Electron Optical Systems Containing Toric Lenses or Quadrupoles. *Opt. Acta* **12,** 237 (1965).

24. B. R. A. Nijboer, Thesis, University of Groningen, 1942.
25. E. C. Kintner, A Recurrence Relation for Calculating the Zernike Polynomials. *Opt. Acta* **23,** 499 (1976).
26. E. C. Kintner, On the Mathematical Properties of the Zernike Polynomials. *Opt. Acta* **23,** 679 (1976).
27. Richard Barakat, Optimum Balanced Wave-Front Aberrations for Radially Symmetric Amplitude Distributions: Generalizations of Zernike Polynomials. *J. Opt. Soc. Am.* **70,** 739 (1980).
28. C. J. Kim and R. R. Shannon, Catalog of Zernike Polynominals. In *Applied Optics and Optical Engineering*, Vol. 10, R. R. Shannon and J. C. Wyant (Eds.). Academic, San Diego, 1987.

5

Mathematical Theory of OTF

INTRODUCTION

As we have already indicated in discussing geometrical optics in the previous two chapters, rays from a point object in a perfect optical system form a homocentric pencil of rays that converge to a point image; and the wave fronts (which, by definition, are perpendicular to the rays) are spherical. However, if in this approach aberrations keep the rays from passing through a common point, the wave fronts cannot be spherical.

In geometrical optics, analysis of aberrations becomes a study of differences between particular nonhomocentric pencils of rays and the ideal homocentric pencil. In the OTF approach, on the other hand, aberrations are identified with how the wave front departs from a spherical shape. The link between these departures and the image is the diffraction integral, which relates the complex amplitude of light on the image plane to the discrepancies between the actual wave front and a reference sphere. A significant difference between the geometrical and OTF approaches is that the ''perfect'' optical system for the latter does not form a point image of a point object but rather a diffraction pattern (sometimes referred to as an *Airy pattern*) consisting of a disk surrounded by a series of rings. The pattern results from limiting the solid angle of an actual wave front to something considerably less than a complete sphere. Addition of aberration effects further complicates the image pattern.

If we broaden our consideration of a point object to an extended object, which may be thought of as an assembly of point objects, it soon becomes apparent that the job of calculating the corresponding extended image by superposing aberrated diffraction patterns is extremely difficult except for geometrically simple objects. For example, consider Fig. 2.16 and try to visualize the task of superposing just a few of these, each produced by one of a cluster of point sources in the object. Specifying an aberration tolerance in terms of image quality by this approach is virtually impossible. Therefore, as a practical matter, it becomes desirable to calculate the OTF from the wave-front aberration because the OTF can be related in meaningful ways to image quality.

DEFINITIONS, NOMENCLATURE, AND CONVENTIONS

The foundation for the mathematical theory of OTF has been laid in the earlier chapters and in the appendix of this book, and now we distill some of the ideas from these pages in preparation for further mathematical development.

As suggested in the introduction of this chapter, a key relation for the study of the OTF is the diffraction integral

$$\hat{U}_0(u'_S, v'_T) = \hat{C}\hat{f}_p \iint_{\mathcal{A}} \hat{G}(x'_S, y'_T) \exp[i2\pi(x'_S u'_S + y'_T v'_T)]\, dx'_S\, dy'_T, \tag{5-1}$$

which is a slightly abbreviated form of Eq. (4-22). The expression is for $\hat{U}_0$, the complex amplitude at a point on the image plane at the reduced coordinates (u'_S, v'_T). Instead of the reduced coordinates, the real-space coordinates (ξ_0, η_0) are sometimes indicated, which have their origin on the image plane at the ideal image point; the ideal image point has, in turn, the real-space coordinates $(0, \eta')$ with the optic axis as the origin. (Alternative symbols for ξ_0 and η_0 are $\Delta\xi'$ and $\Delta\eta'$, respectively.) The first term, $\hat{C}$, in the expression for $\hat{U}_0$ is a complex constant. The second term, $\hat{f}_p$, is a function of the coordinates of the point on the image plane for which $\hat{U}_0$ is expressed. Again, these coordinates may be indicated in any of the several ways already discussed in connection with $\hat{U}_0$. The double integral is taken over the surface $\mathcal{A}$ of the wave front as it emerges from the exit pupil and is written in terms of the coordinates (x'_S, y'_T) on the reference sphere at the exit pupil. Because the wave front is abruptly cut off by the aperture, the pupil function $\hat{G}(x'_S, y'_T)$ in the integrand has to be written in two ways:

$$\hat{G}(x'_S, y'_T) = G(x'_S, y'_T) \exp[-ikn'W(x'_S, y'_T)] \quad \text{when } (x'^2_S + y'^2_T) \le 1, \tag{5-2}$$

and

$$\hat{G}(x'_S, y'_T) = 0 \qquad \text{when } (x'_S + y'_T) > 1, \tag{5-3}$$

which were originally written as Eqs. (4-29) and (4-30). Because the pupil function has a zero value everywhere beyond the unit circle (Eq. (5-3)), the double integral in Eq. (5-1) can be shown with the limits $-\infty$ and $+\infty$ instead of over the surface $\mathcal{A}$.

The exponential factor of the pupil function has already been discussed at great length as the wave aberration function, which was shown expanded in the previous chapter both as a power series and in Zernike polynomials. Because of the symmetry of the assumed optical system, the indicated rectangular coordinates (x'_S, y'_T) give way to derived polar coordinates $(\rho, \cos\varphi)$ or $(\rho, \cos m\varphi)$ in the expansions.

In spite of the approximations involved in the derivation of Eq. (5-1), the diffraction integral, this formula is found highly accurate for practical applications in optical system study.

The normalized coordinates (x'_S, y'_T) for the exit pupil were first defined in Chapter 3 by Eq. (3-27). In all subsequent discussions, the subscripts S and T (signifying *sagittal* and *tangential*, respectively) have been retained to distinguish the normalized coordinates from the real-space paraxial coordinates (x', y') introduced earlier in Chapter 3. For consistency, the same subscripts have been applied to the reduced image plane coordinates (u'_S, v'_T) defined in Eq. (3-40). The primes, indicating image space, have been retained to distinguish the coordinates from those (unprimed) for object space. Now we propose to simplify expressions involving these symbols by dropping both the subscripts and the primes, and we offer the following justification: The mathematics in the remainder of this book will not involve paraxial assumptions; so dropping the subscripts should cause no confusion. Unless we indicate otherwise, our discussions will assume the necessary conditions, including isoplanatism (see Eq. (3-29)), so that

$$x = x', \qquad y = y', \qquad u = u', \quad \text{and} \quad v = v'. \tag{5-4}$$

Thus the unprimed symbols, in most instances, can represent image space coordinates as well as object space coordinates.

As indicated in the development of the normalized and reduced coordinate systems, the x-axes in both object and image spaces are parallel to the sagittal plane and are perpendicular to the optic (z) axis; the y-axes lie in the tangential plane and are also perpendicular to the optic axis. The u-axes in both object and image spaces are parallel to the x-axes, and the v-axes are parallel to the y-axes. Whenever a single off-axis object point is being considered in a rotationally symmetrical (about the optic axis) optical system, no loss of generality is suffered by placing the object point $\overline{Q}$ in the tangential plane as in Fig. 5.1.

Comparison of the diffraction integral, Eq. (5-1), and the Fourier integral, Eq. (B-16) of Appendix B, show them to have identical mathematical forms. This likeness accounts for many of the useful relations in OTF theory. Discussions of Fourier transforms, particularly in Appendix B, have shown that the indicated transformation of function, $\hat{G}(x'_S, y'_T)$ to $\hat{U}_0(u'_S, v'_T)$ in Eq. (5-1), can be reversed by an expression that differs in form only in the sign of the exponent. (Compare Eqs. (B-15) and (B-16).) Since these reversible transformations occur frequently in OTF theory, Eq. (5-1) and its reverse are often shortened to

$$\hat{U}_0(u'_S, v'_T) \leftrightarrow \hat{G}(x'_S, y'_T), \tag{5-5}$$

or, with the omissions of subscripts and primes,

$$\hat{U}_0(u, v) \leftrightarrow \hat{G}(x, y). \tag{5-6}$$

The quantity $\hat{U}_0(u, v)$ has already been identified in Eq. (5-1) as the complex amplitude of the electromagnetic disturbance at the general point (u, v) in the image plane near $\bar{Q}'$ (Fig. 5.1) and with the origin of the coordinates (u, v) at $\bar{Q}'$. The function representing the complex amplitude, $\hat{U}_0(u, v)$, is called the *amplitude point spread function* whose square is proportional to the flow rate of energy at (u, v), that is, flux density (in W/cm^2, for example). The mathematical expression for this relation is

$$B_0(u, v) = [\hat{U}_0(u, v)][\hat{U}_0^*(u, v)] = \left|[\hat{U}_0(u, v)]\right|^2. \tag{5-7}$$

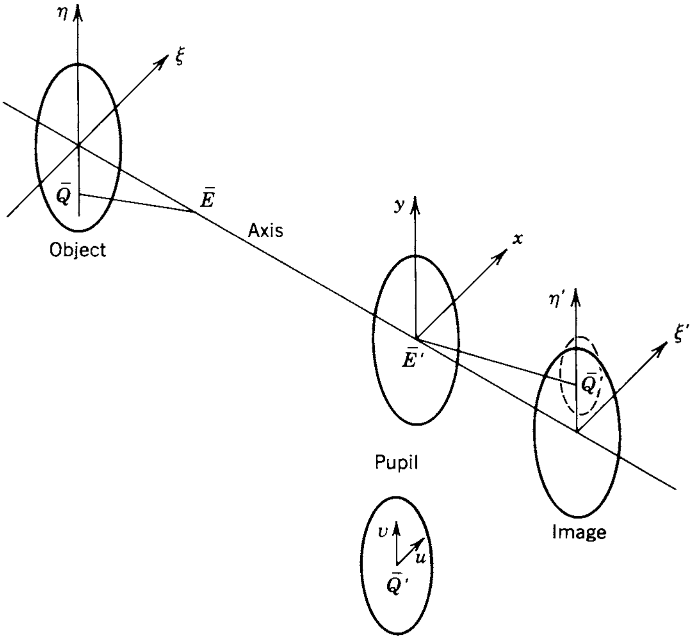


Figure 5.1. The coordinate systems.

The product of $\hat{U}_0$ times its conjugate is by definition the square of U_0 and, by virtue of this mathematical operation, is always a real quantity. The function $B_0(u, v)$ is called the *flux-density point spread function* or PSF. It describes the distribution of flux density in the image plane produced by a point source in the object plane. The integral of the point spread function over the image plane is often arbitrarily set equal to unity, which is equivalent to assuming that the total light flux passing through the system from the point object is a unit quantity.

Because the flux-density point spread function is a two-dimensional distribution of a quantity over the image plane, a Fourier transform as in Eq. (B-15) can be applied to $B_0(u, v)$ to find the equivalent two-dimensional spatial frequency function at the image plane:

$$\hat{b}_0(s_S, s_T) = \iint_{-\infty}^{+\infty} B_0(u, v) \exp[-i2\pi(s_S u + s_T v)]\, du\, dv. \tag{5-8}$$

This, when normalized with respect to its value at $s_S = s_T = 0$, is called the *optical transfer function*. Because the coordinates (u, v), in more explicit notation, were defined as normalized to unity wavelength in Eq. (3-40), the frequencies (s_S, s_T) are normalized to unit frequency and correspond to the frequencies defined in Eq. (3-47).

Definite integrals with positive and negative infinity as limits require special functions in the integrand to be practical, that is, the value of the function must become negligible at reasonably small values of the independent variables. We have already discussed the diffraction integral and the nature of the pupil function that permits the infinity limits. In integrals like that in Eq. (5-8), where integration is over the image plane and the integrand function is usually a flux density, practically all of the optical power is confined to a relatively small area whose boundaries are apparent from the parameters of the optical system.

By normalizing the constants preceding the integral in Eq. (5-1) and reviewing the discussion leading to Eq. (5-6), we can write

$$\hat{G}(x, y) = \iint_{-\infty}^{+\infty} \hat{U}_0(u, v) \exp[-i2\pi(ux + vy)]\, du\, dv, \tag{5-9}$$

which indicates that the pupil function $\hat{G}(x, y)$ can be derived from the complex amplitude function $\hat{U}_0(u, v)$ on the image plane.

From the relations expressed in Eqs. (5-8) and (5-9), the spatial frequency function $\hat{b}_0(s_S, s_T)$ at the image plane can be written in terms of the pupil function $\hat{G}(x, y)$. In brief notation, Eq. (5-8) can be expressed as

$$\hat{b}_0(s_S, s_T) \leftrightarrow B_0(u, v). \tag{5-10}$$

Just as $\hat{U}_0$ is related to $\hat{G}$ in Eq. (5-6), we can assume that the second factor $\hat{U}_0^*$ of the middle expression of Eq. (5-7) has a similar relation to some function $\hat{G}_1$:

$$\hat{U}_0^*(u, v) \leftrightarrow \hat{G}_1(x, y). \tag{5-11}$$

By applying the convolution theorem (discussed in connection with Eq. (B-31) and the following equations in Appendix B) to Eqs. (5-6), (5-7), (5-10), and (5-11), one can write

$$\hat{b}_0(s_S, s_T) = \iint\limits_{-\infty}^{+\infty} [\hat{G}(x, y)][\hat{G}_1(s_S - x, s_T - y)]\, dx\, dy, \tag{5-12}$$

which, in words, states that $\hat{b}_0$ is the convolution of $\hat{G}$ and $\hat{G}_1$. From the theory of Fourier integrals, it is known that the transform of the complex conjugate of a function is the reversed complex conjugate of the transform of the function [1, p. 16]; that is, if

$$\hat{f}(x) \leftrightarrow \hat{F}(\omega), \tag{5-13}$$

then

$$\hat{f}^*(x) \leftrightarrow \hat{F}^*(-\omega). \tag{5-14}$$

This indicates that Eq. (5-12) can be written

$$\hat{b}_0(s_S, s_T) = \iint\limits_{-\infty}^{+\infty} [\hat{G}(x, y)][\hat{G}^*(x - s_S, y - s_T)]\, dx\, dy. \tag{5-15}$$

This equation is recognized, by comparison with Eq. (B-84), as the autocorrelation function of $\hat{G}(x, y)$. When a constant is added to or subtracted from each independent variable as in the second factor in the integrand of Eq. (5-15), the effect is to shift the function on the coordinate axis. (See the discussion under "Significance of the Convolution Integral" in Appendix B.) Any limits that apply to the function are correspondingly shifted. Because a pupil function has

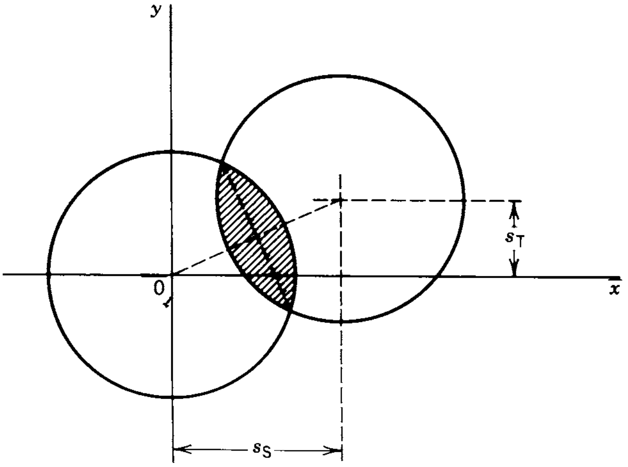


Figure 5.2. Common, or overlapping, area of two sheared circles as in the autocorrelation integral.

nonzero values only within a circle of unit radius as stated in Eqs. (5-2) and (5-3), the nonzero areas for the integrand factors in Eq. (5-15) appear graphically as in Fig. 5.2. Since both factors have to be nonzero simultaneously to give the integrand a nonzero value, the integration is taken over only the overlapping region (shaded in the figure) of the two circles. Two functions offset as indicated in Fig. 5.2 are said to be "sheared." From the diagram, it is apparent that the distance between the circle centers is $(s_S^2 + s_T^2)^{1/2}$. A set of examples based on this figure is worked out in a later section of this chapter.

Equation (5-1) and the subsequent discussion of various optical mathematical concepts have been based on a point source in the object plane without any attempt to write a mathematical expression for the point source. Such an expression can be set up with the application of the Dirac delta function as discussed in Appendix B, where it is shown that this function is zero everywhere except at a point and that its integral over the area in the vicinity of the point is unity. Thus, the flux density in the *object* plane can be represented by $B_1(u, v)$ as

$$B_1(u, v) = C_1\delta(u, v) = C_1[\delta(u)][\delta(v)], \tag{5-16}$$

where C_1 is a constant.

Also, in Appendix B, it is shown that the Fourier transform of the delta function is unity, a real constant. So, the spatial frequency spectrum produced by a point source of light is continuous with constant power level; but this spectrum represents the "input" to the optical system in terms of frequency.

Application of the Fourier transform to the point source may be expressed as

$$B_1(u, v) \leftrightarrow C_1. \qquad (5\text{-}17)$$

When we examine the resulting image plane distribution of spatial frequencies (as we do in later sections of this chapter) by properly applying the Fourier transform to the flux-density point spread function, we find that the amplitude in the image generally declines with increasing frequency (although in some systems, the level might increase with increasing frequency over short intervals), and a cutoff frequency is finally reached beyond which the power remains zero at all frequencies.

Borrowing from the language of communication engineering, we can say that image forming optical systems invariably behave like low-pass filters. The particular characteristic of any spread function $B_0(u, v)$ is a manifestation of the imaging properties of the optical system including the aberrations that it produces. The amplitude and phase over a wave front at the exit pupil determine the performance of the optical system both in terms of the impulse (point source) response, which is the point spread function, and in terms of the Fourier transform of the spread function, which is the unnormalized optical transfer function. The spectrum of the distribution received at the image $b_i(s_S, s_T)$ can be expressed in terms of the input spectrum C_1 and the optical transfer function, before normalization at $(0, 0)$, $\hat{b}_0(s_S, s_T)$, as

$$\hat{b}_i(s_S, s_T) = [\hat{b}_0(s_S, s_T)]\, C_1. \qquad (5\text{-}18)$$

The transforms of the three terms in Eq. (5-18) are

$$\hat{b}_i(s_S, s_T) \leftrightarrow B_i(u, v), \qquad (5\text{-}19)$$

$$\hat{b}_0(s_S, s_T) \leftrightarrow B_0(u, v), \qquad (5\text{-}20)$$

$$C_1 \leftrightarrow C_1 \delta(u, v). \qquad (5\text{-}21)$$

The relation in Eq. (5-21) is discussed in connection with Eqs. (B-25) and (B-26) in Appendix B. By applying the convolution theorem, as was done to write Eq. (5-12),

$$B_i(u', v') = \iint_{-\infty}^{+\infty} B_0(u, v)\,[\delta(u' - u, v' - v)]\, du\, dv. \qquad (5\text{-}22)$$

In words, Eq. (5-22) says that the distribution of flux density in the image plane is equal to the convolution of the point spread function (a property of the optical system) and the distribution of flux in the object plane (a point source), which distribution happens here to be described by the delta function. Since a general distribution in the object plane can be thought of as the superposition of many points, one might expect that Eq. (5-22) could be generalized to

$$B_i(u', v') = \iint_{-\infty}^{+\infty} B_0(u, v)\, B_1(u' - u, v' - v)\, du\, dv, \qquad (5\text{-}23)$$

where $B_1(u, v)$ is a function describing a general distribution of flux density in the object plane. It may be useful to visualize the distribution as a juxtaposition of an infinity of point sources, each incoherent within itself and each incoherent in relation to every other point source. Each point source has the appropriate intensity and enough "spread" to render the object-plane flux-density distribution continuous. This powerful optical relation, Eq. (5-23), does indeed hold provided certain restrictions are observed, which are discussed in the next section.

LINEARITY AND ISOPLANATISM

The mathematical statements of the preceding section are based on a number of tacit assumptions that should be reviewed whenever OTF theory is applied. The bases for these assumptions are discussed in Chapter 2 in connection with isoplanatism, linearity, and coherence.

The light beam passing through the optical system is assumed in our developments to be a constant single-frequency wave train, that is, a perfectly coherent beam. Actually, the usual situation is that a spectrum of frequencies, variable both in amplitude and spectral content, compose the beam. This is true even when a nominally single-frequency, incoherent, finite, or damped wave train is involved.

Fortunately for the application of OTF theory, most optical systems are linear in the sense that response is proportional to an appropriately selected input. This property allows one to apply each component frequency or beam individually and then to combine the results to obtain the complete response function for the system. Some written discussions of optical systems responses to coherent and incoherent light give the impression that a system reacts differently to the two kinds of light. However, an optical system, in general, is strictly passive. The fundamental response of the system to any given frequency component of incoherent light is identical to the response to the single frequency in coherent light, provided the two frequencies are the same. The differences exist because the two kinds of light must be superposed differently.

Before we can follow specific rules for superposition, we must be assured that we are operating within a single isoplanatism patch as discussed in Chapter 2. For instance, if two identical point sources, one at (u_1, v_1) and the other at $(u_1 + \Delta u, v_1 + \Delta v)$, produce displaced patterns on the image plane that are precisely identical, the two points are said to be in the same isoplanatism patch; and the same point spread function can be convolved with the delta function representing each point to produce the mathematical expression for the corresponding pattern as in Eq. (5-22). According to the nomenclature already adopted, the two intensity distributions in the image are

$$B_1(u_1, v_1) \quad \text{and} \quad B_1(u_1 + \Delta u, v_1 + \Delta v). \tag{5-24}$$

Similarly, by convolving the amplitude point spread function with the delta function, the corresponding pair of identical (except for location) amplitude distribution functions result:

$$\hat{U}_1(u_1, v_1) \quad \text{and} \quad \hat{U}_1(u_1 + \Delta u, v_1 + \Delta v). \tag{5-25}$$

From the described characteristics of an isoplanatism patch, it can be said to be a region in the object plane over which one wave-front aberration function holds for all points.

Having satisfied the isoplanatism condition, we face the question of how to combine the patterns produced through the optical system by the two point sources. Obviously, superposition is achieved by adding some property of the two patterns, point by point, on the image plane—But should we add intensities (B_1) or amplitudes $(\hat{U}_1)$? The answer depends on whether the two sources are incoherent or coherent. If incoherent, intensities (real) must be superposed; if coherent, the known phase relations must be taken into account and the amplitudes (complex) superposed. Because of the fixed phase relations characteristic of coherent light, the superposed patterns will show diffraction effects not evident in the superposed incoherent patterns; however, if we could observe the superposed incoherent patterns instant by instant, we would see a dynamic sequence of diffraction effects similar to the stationary effects noted for coherent light. It is the short-term dynamic averaging that permits the direct superposition of intensities:

$$B_i(u_1, v_1) = B_1(u_1, v_1) + B_1(u_1 + \Delta u, v_1 + \Delta v). \tag{5-26}$$

However, because of the "stop-action" nature of the coherent light situation, the more tedious superposition of amplitudes, with appropriate attention to phase, has to be performed:

$$\hat{U}_{\mathrm{i}}(u_1, v_1) = \hat{U}_1(u_1, v_1) + \hat{U}_1(u_1 + \Delta u, v_1 + \Delta v). \qquad (5\text{-}27)$$

Then by squaring the total amplitude point by point, as in Eq. (5-7), the $B_{\mathrm{i}}(u_1, v_1)$ for coherent light can be determined. Theoretically, the incoherent superposition problem could also start with Eq. (5-27) if the instantaneous phase relations were known, but this approach has no practical application.

For convenience in the superposition discussion, the displaced identical patterns previously postulated for the isoplanatism discussion were used for illustration. The reader should realize that the conclusions reached about superposition apply as well to unlike patterns.

Unfortunately, no clear-cut method of image plane superposition exists for partially coherent light. Solutions would require some kind of combination of results reached by first assuming coherent and then incoherent light. Image patterns for coherent and incoherent light in otherwise identical optical systems have no direct relation to each other. Kintner and Sillitto [2] suggest a performance indicator, unrelated to OTF principles, that is based on the oscillatory ringing observed in images formed with partially coherent light. In another approach [4–12 of Chapter 2], a calculated mutual coherent function, based on pairs of points in the object, does have a linear relation in the image, even for partially coherent light, and can be used as a performance indicator. However, because it has no readily apparent characteristic discernible in the image and its value carries no intuitive significance of image quality, the mutual coherence function does not seem as popular as performance tests that concentrate on a particular characteristic (like ringing) of the imaging process. Because the Kintner and Sillitto treatment and other performance evaluations for partially coherent light are not related to OTF theory, further discussion of them is limited to remarks accompanying Eqs. (10-4) and (10-5) [1 of Chapter 10].

If no reference is made to coherence, the optical analyst must make an assumption concerning this characteristic when conclusions are dependent on which kind of light is involved as in the formation of images. Also, the optical experimenter must be alert to the nature of test discrepancies that result from partial coherence in a system assumed to be operating with incoherent light.

IMAGE OF A GENERAL DISTRIBUTION

The superposition of the images of two object points, discussed in the previous section, is the first step toward developing the expressions that describe the formation of an image from a general extended object. If such an object is a self-luminous area of incoherent light on the object plane, it can be described by the intensity (flux-density) distribution $B_1(u, v)$. It is assumed to lie within a single isoplanatism patch.

Radiant flux from each element of area (point) on the object produces its particular intensity distribution on the image plane; that is, any object distribution $B_1(u, v)$ may be regarded as a spatial distribution of an infinity of delta functions, each being multiplied by $B_1(u, v)$ at its location. The convolution integral is applied to express the interaction of two functions like the intensity distribution in the object $B_1(u, v)$ and the point spread function $B_0(u, v)$ of the optical system to produce the intensity distribution $B_i(u', v')$ in the image. The principles of this application are discussed in Appendix B under "Significance of the Convolution Integral." The mathematical expression in the present instance has already been anticipated in Eq. (5-3). As indicated in the identity of Eqs. (B-47) and (B-48) in Appendix B, Eq. (5-3), can also be written

$$B_i(u', v') = \iint_{-\infty}^{+\infty} B_1(u, v)\, B_0(u' - u, v' - v)\, du\, dv. \tag{5-28}$$

As in other integrals showing $-\infty$ and $+\infty$ as the limits, the practical limits here between which the integration has to be performed define a comfortably small span. The outer elements on the object plane that need be considered are only those that make an appreciable contribution to the light that passes through the optical system and reach the area of interest on the image plane. When the isoplanatism patch is defined, its limits are sometimes substituted for the infinity signs.

By applying the convolution theorem as was demonstrated earlier in this chapter, Eq. (5-28) is found equivalent to

$$\hat{b}_i(s_S, s_T) = [\hat{b}_1(s_S, s_T)]\,[\hat{b}_0(s_S, s_T)], \tag{5-29}$$

where $\hat{b}_1(s_S, s_T)$ represents the spatial frequency distribution of intensity in the object, $\hat{b}_0(s_S, s_T)$ is the frequency response of the optical system, and $\hat{b}_i(s_S, s_T)$ represents the resulting frequency distribution of intensity in the image. Typically Eq. (5-29) describes the low-pass filter operation of the optical system on the general spectrum in the object to produce the modified spectrum in the image. Inasmuch as the spectra are described by complex numbers, both amplitude and phase are involved in each value. The optical system frequency response function $\hat{b}_0(s_S, s_T)$, when normalized to unity at zero spatial frequency, is defined as the OTF.

The pupil function $\hat{G}(x, y)$ (discussed in connection with Eqs. (5-2) and (5-3)) is sometimes also called a frequency response function; but here the reference applies to the amplitude and phase of the electromagnetic disturbance (neither observable by the eye, detectable by any sensor, nor recordable by photographic film) whereas $\hat{b}_0(s_S, s_T)$ refers to spatial frequency properties of flux density distributions. Unless indicated otherwise, *frequency response function* in this text has reference to a *spatial frequency response function of flux density* as in Eq. (5-29).

ONE-DIMENSIONAL ANALYSIS

Two-dimensional distributions such as $\hat{b}_i(s_S, s_T)$ and $\hat{b}_1(s_S, s_T)$ of Eq. (5-29) are sometimes hard to handle in analysis and computation, and they are especially troublesome conceptually as guides in measurements and experiments. Often, without any loss in generality, a one-dimensional approach can be substituted for the two-dimensional mathematics to simplify and clarify analysis and visualization of the optical processes.

In Fig. 5.3 a sinusoidal distribution of flux density in the object plane is represented schematically by the dashed lines, which are loci of maximum flux density. This distribution is designated $B_1(u, v)$ with the spatial frequencies s_S along the u-axis and s_T along the v-axis. When the lines of constant flux density are parallel to neither axis, as shown in the figure, the axes can be rotated about the origin through an angle ψ to make the lines parallel, in this instance, to the

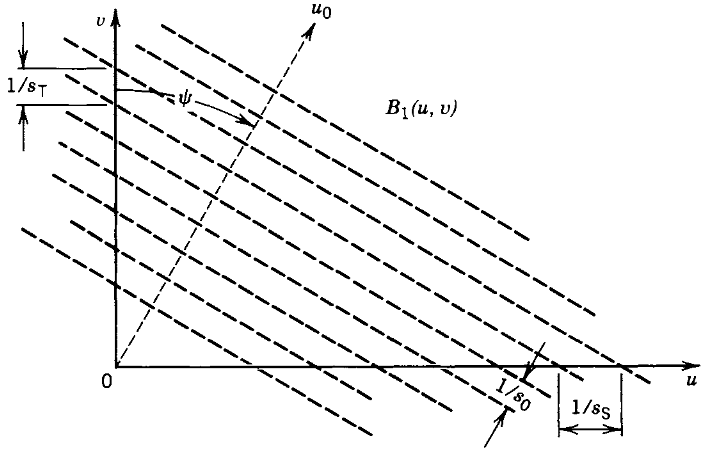


Figure 5.3. Schematic of a one-dimensional sinusoidal distribution.

u-axis. From the geometry of the figure,

$$\psi = \arctan(s_S/s_T). \tag{5-30}$$

The spatial period along the u_0-axis, which is perpendicular to the dashed lines, is $1/s_0$. Further relations in Fig. 5.3 are

$$\sin\psi = (1/s_0)/(1/s_S) = s_S/s_0, \tag{5-31}$$

$$\cos\psi = s_T/s_0, \tag{5-32}$$

$$\sin^2\psi + \cos^2\psi = (s_S/s_0)^2 + (s_T/s_0)^2 = 1, \tag{5-33}$$

$$s_0^2 = s_S^2 + s_T^2. \tag{5-34}$$

Since in most analyses only positive frequencies are considered, the angle ψ is usually confined to the first quadrant.

For illustration in Fig. 5.3 we have chosen a single-frequency pattern for which we can write the Fourier transform relation and the coordinate shifts as

$$B_1(u, v) \leftrightarrow \hat{b}_1(s_S, s_T), \tag{5-35}$$

$$B_1(u, v) = B_1(u_0, \psi), \tag{5-36}$$

$$\hat{b}_1(s_S, s_T) = \hat{b}_1(s_0, \psi). \tag{5-37}$$

Although all the functions shown in the three relations above have two independent variables, $B_1(u_0, \psi)$ and $\hat{b}_1(s_0, \psi)$ have the advantage of easily visualized functions of the single variable u_0 or s_0 once the axes are shifted through the angle ψ.

This single-frequency example can be expanded to the general situation, with no revision in the three relations, were $B_1(u, v)$ is any distribution of flux density. Then $\hat{b}_1(s_0, \psi)$ gives the amplitude and phase of a single spatial frequency wave in the azimuth direction ψ for each pair of s_S and s_T. The total $B_1(u, v)$ is the superposition of all these single-frequency waves with appropriate attention to azimuth and phase. Each component wave of frequency s_0 in the superposition has its azimuth direction ψ and is a space function of the form

$$B_1(u_0) = \alpha + \beta\cos(2\pi s_0 u_0 + \gamma), \tag{5-38}$$

where β is the amplitude of the wave, α (which is equal to or greater than β) is the constant level necessary to keep the wave from going negative (because negative flux density has no physical meaning), and γ is the relative phase angle of the component wave. Actually, the inclusion of a constant such as α to prevent negative values of flux density need only be considered *after* all alternating waves are algebraically added; it is their *sum* that must not be allowed to swing to negative values. In the Fourier analysis, the total of all α terms occurs as a single term.

A particular consequence of the one-dimensional treatment is that slits and gratings—a sinusoidal distribution is one kind of grating—oriented successively in different azimuths may be used for testing the imaging properties of optical systems. Experimentally, slits rather than point sources are desirable to get a high signal-to-noise ratio at the detector. In practical setups, a slit can transmit up to 10^4 times as much radiant power as a pinhole.

Another benefit from the one-dimensional treatment is that a single independent variable suffices for a plot of the computed frequency response function $\hat{b}_0(s_S, s_T)$ in Eq. (5-29).

When the sagittal magnification m_S differs from the tangential magnification m_T, because of distortion by the optical system, the expressions for one-dimensional treatment and the observations in related testing procedures are modified. A direct result of the difference in magnifications is that the spatial periods in the image plane, $1/s_S'$ and $1/s_T'$, bear different ratios to the corresponding periods, $1/s_S$ and $1/s_T$, in the object plane:

$$1/s_S' = m_S/s_S, \tag{5-39}$$

$$1/s_T' = m_T/s_T. \tag{5-40}$$

The expression for the image azimuth angle ψ', corresponding to the object azimuth angle ψ in Eq. (5-30), is

$$\begin{aligned}\psi' &= \arctan(s_S'/s_T') \\ &= \arctan[(m_T/m_S)(s_S/s_T)],\end{aligned} \tag{5-41}$$

so

$$\tan\psi' = (m_T/m_S)\tan\psi. \tag{5-42}$$

In a test setup where a grating is in the object plane, the grating azimuth angle will generally differ from the image azimuth angle when distortion is present. As Eq. (5-42) indicates, exceptions occur at angles of 0 and 90°.

When the image frequency is to be calculated and either or both the magnification constants differ from unity, Eq. (5-34) has to be modified as follows:

$$(s_0')^2 = (s_S/m_S)^2 + (s_T/m_T)^2. \tag{5-43}$$

OPTICAL TRANSFER FUNCTION

The optical transfer function (OTF) is one of the concepts presented in Chapter 2. The illustrating expression, Eq. (2-19), gives the OTF as a function of a single real-space spatial frequency ω. The magnitude (modulus) T of the OTF is shown in the discussion preceding the equation to be a ratio of contrasts C'/C in the image and object.

Now, with the benefit of considerable concept and mathematical development beyond Chapter 2, the definition of the OTF can be brought into a more conventional form based on the function $\hat{b}_0(s_S, s_T)$ in Eqs. (5-10) and (5-15). An expression for $\hat{b}_0(s_S, s_T)$ that closely parallels the expression for the OTF in Chapter 2 is

$$\hat{b}_0(s_S, s_T) = b_0(s_S, s_T) \exp[i\phi(s_S, s_T)]. \tag{5-44}$$

The principal difference between Eqs. (2-19) and (5-44) is that two reduced spatial frequencies (defined in Eq. (3-47)) have been substituted in Eq. (5-44) for the single real-space frequency. A further modification to $\hat{b}_0(s_S, s_T)$ to bring it into a common OTF (symbol: $\hat{O}$) form is to normalize it with reference to its value at zero frequency:

$$\begin{aligned} \hat{O}(s_S, s_T) &= \hat{b}_0(s_S, s_T)/\hat{b}_0(0, 0) \\ &= [b_0(s_S, s_T)/b_0(0, 0)] \exp[i\phi(s_S, s_T)] \\ &= T(s_S, s_T) \exp[i\phi(s_S, s_T)], \end{aligned} \tag{5-45}$$

where $T(s_S, s_T)$ is the modulation transfer function,

$$\text{MTF} = T(s_S, s_T), \tag{5-46}$$

and $\phi(s_S, s_T)$ is the phase transfer function,

$$\text{PTF} = \phi(s_S, s_T). \tag{5-47}$$

By the method demonstrated in the previous section on one-dimensional analysis, these functions can be indicated as functions of a single frequency s, whose subscript is dropped to simplify notation:

$$\hat{O}(s) = T(s) \exp[i\phi(s)], \tag{5-48}$$

$$\text{MTF} = T(s) = b_0(s)/b_0(0), \tag{5-49}$$

$$\text{PTF} = \phi(s). \tag{5-50}$$

Because the zero-frequency values have been taken as references in the normalizing procedure, it follows that the normalized value of the MTF approaches unity and the normalized value of the PTF approaches zero as the frequencies approach zero.

With reference to Fig. 5.1, which shows the off-axis image point $\overline{Q}'$ as the origin of the (u, v) coordinates, a number of assumptions pertinent to the present discussion can be reviewed. An OTF for an optical system holds for only one location of $\overline{Q}'$, which would seem to require two coordinate values in the image plane; however, since circular symmetry of the system is assumed, only a radial value is needed. Furthermore, from the definition of the tangential plane, whose intersection with the image plane is the vertical coordinate axis, the radial value is identical with the η coordinate because $\overline{Q}'$ is always, by definition, in the tangential plane.

The OTF, which can be thought of as an operator on each spatial frequency in the object plane to produce a corresponding result in the image plane, is dependent on the orientation of the object plane frequency pattern. The orientations favored in most optical test procedures are shown in Fig. 5.4. In Fig. 5.4*a* the grating is said to be in the *radial direction* ($\psi = 0$) with lines of constant phase parallel with the sagittal plane (ξ-axis); in Fig. 5.4*b* the grating is in the *tangential direction* ($\psi = 90° = \pi/2$ radians) with lines of constant phase parallel with the tangential plane (η-axis) [14]. Each diagram has an inset of defining the general angle ψ on the (u, v) coordinate axes. Except where a standard specifies the nomenclature of particular grating orientations as in the statements above, the practice in this book is to describe spatial frequency orientation either by giving the angle ψ of the line perpendicular to the lines of constant flux density in the spatial frequency pattern or by indicating by subscript whether the lines of constant flux density are perpendicular to an axis in the tangential plane (T) or in the sagittal plane (S). Specifying *radial direction* according to the standard is consistent with our practice because the grating lines are indeed perpendicular to a radial line from the optic axis. However, its counterpart, *tangential direction*, refers to perpendicularity to a line tangent to a circle of symmetry (*not* to a line in the tangent plane). Hence,

$$\omega \text{ in ``radial direction''} \equiv \omega_{\mathrm{T}}, \tag{5-51}$$

$$\omega \text{ in ``tangential direction''} \equiv \omega_{\mathrm{S}}. \tag{5-52}$$

Although, the OTF, MTF, and PTF are usually indicated as functions of the single frequency variable s as in Eqs. (5-48)–(5-50), the above discussions about radial and orientation dependence imply that an explicit indication of dependence would require

$$\mathrm{OTF} = \hat{O}(s, r, \psi) = T(s, r, \psi)\exp[i\phi(s, r, \psi)], \tag{5-53}$$

with similar independent variable expressions for the MTF and PTF.

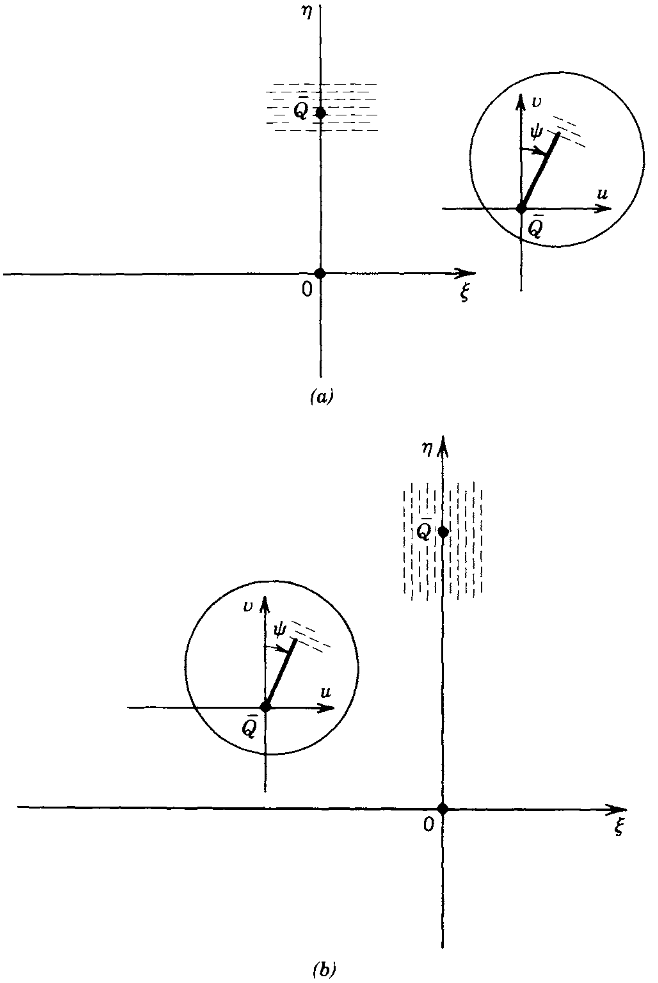


Figure 5.4. Spatial frequency lines oriented (*a*) in radial direction and (*b*) in tangential direction. Insets show orientation in a general intermediate direction ψ. Insets also show the (u, v) coordinates, which have their origin at $\bar{Q}$.

Another independent variable expression that requires comment is in the definition of *relative modulation*, $M(s, \psi)$:

$$M(s, \psi) = T(s, \psi)/T(s, \psi; 0), \tag{5-54}$$

where $T(s, \psi; 0)$ denotes the value of the modulation transfer function at frequency s and orientation ψ if the system under consideration is limited only by diffraction. The relative modulation function is often used in specifications of optical systems and is discussed in a subsequent section in connection with the variance of the aberration difference function.

THE PERFECT OTF

As indicated in the previous section, our definition of the OTF is a normalized $\hat{b}_0(s_S, s_T)$, which can be calculated for a particular optical system by integration in either Eq. (5-10) or Eq. (5-15). Numerical methods of integration are usually applied. Some of these methods are discussed in Chapter 10. However, a *general* expression involving only algebraic and transcendental functions without calculus operations for $\hat{b}_0(s_S, s_T)$ cannot be derived from either integral expression. To arrive at a manageable OTF expression, certain reasonable limiting assumptions can be applied without eliminating all generality in the results.

In the earlier chapters of this book and indeed in most general optical treatments elsewhere, circular (rotational) symmetry about the optical axis is usually assumed. This means that optical elements (lenses, stops, mirrors, etc.) in a system are centered on the axis, and the circular contours and edges of the elements are in planes perpendicular to the axis. A property of such a system is that the image of a fixed object is unaffected by rotation of the system about its axis. For a general off-axis point object, a *tangential* plane is defined in Chapter 3 to include the point object and the optical axis. A result is that the pupil ray from the point object to the ideal image point on the image plane remains in the tangential plane throughout the system. Where it is convenient to set up a coordinate system in the object plane, the image plane, or an intermediate surface, the intersection of the tangential plane and each surface is conventionally one of the coordinate axes. Because of the way the tangential plane is defined, this axis is an axis of symmetry in each surface coordinate system.

Besides assuming circular symmetry and setting up a convenient coordinate system, we have to make some simplifying assumptions concerning $\hat{G}(x, \mathrm{y})$ in Eq. (5-15) to reach a useful "general" expression for the OTF.

In Chapter 4, the significance of $\hat{G}(x, y)$ is discussed under "The Pupil Function." Near the beginning of the same chapter, the idea of eliminating wave-front distorting aberrations but accepting the image faults due to diffraction in a corrected system is discussed in explaining the term, "diffraction-limited system." The OTF for such a near-perfect system is the one we here call the *perfect OTF*. By reviewing the pupil function, we see that aberrations are eliminated by making the aberration function $W(x, y)$ zero and the modulus $G(x, y)$ a real constant. These conditions can be most closely approached in an actual system by working with an on-axis point object and paraxial rays (small aperture). Then Eqs. (4-29) and (4-30) reduce to

$$\hat{G}(x, y) = G, \text{ a constant,} \quad \text{when} \quad (x^2 + y^2) \leq 1, \tag{5-55}$$

$$\hat{G}(x, y) = 0 \quad \text{when} \quad (x^2 + y^2) > 1. \tag{5-56}$$

When these values are substituted for $\hat{G}(x, y)$ and $\hat{G}^*(x - s_S, y - s_T)$ in Eq. (5-15) and the graphical representation of the integration is considered in Fig. 5.2, it is apparent that the value of the integral is G^2 times the shaded common area of the two unit-radius circles:

$$\hat{b}_0(s_S, s_T) = G^2[2A(s_S, s_T)], \tag{5-57}$$

where $A(s_S, s_T)$ is the area of the shaded segment of each circle, two of which make up the total shaded area. The geometric formula for a unit-radius circle segment whose chord is $d/2$ from the center is

$$A(d) = \arccos(d/2) - (d/2)\left[1 - (d/2)^2\right]^{1/2}. \tag{5-58}$$

From the earlier discussion of Fig. 5.2,

$$d = (s_S^2 + s_T^2)^{1/2}. \tag{5-59}$$

For $s_S = s_T = 0$,

$$d = 0, \tag{5-60}$$

$$A(0) = \arccos 0 = \pi/2, \tag{5-61}$$

$$\hat{b}_0(0, 0) = \pi G^2. \tag{5-62}$$

The perfect OTF (diffraction limited) by definition is

$$\hat{O}(s_S, s_T) = \hat{b}_0(s_S, s_T)/\hat{b}_0(0, 0). \tag{5-63}$$

By substituting expressions derived for numerator and denominator, we obtain

$$\begin{aligned}\hat{O}(s_S, s_T) &= G^2[2A(d)]/(\pi G^2)\\ &= (2/\pi)\left\{\arccos(d/2) - (d/2)\left[1 - (d/2)^2\right]^{1/2}\right\}.\end{aligned} \tag{5-64}$$

(Note that the angle must be expressed in radians.) By reference to the earlier discussion in this chapter on one-dimensional analysis where it is shown that, with appropriate rotation of axes, a single frequency s (subscript dropped) can be substituted for s_S and s_T,

$$s^2 = s_S^2 + s_T^2. \tag{5-65}$$

It is thus apparent that in a single-frequency analysis, from Eq. (5-59),

$$s = d, \tag{5-66}$$

and the perfect OTF is

$$\hat{O}(s) = (2/\pi)\left\{\arccos(s/2) - (s/2)\left[1 - (s/2)^2\right]^{1/2}\right\}. \tag{5-67}$$

Values for this function are given in Table 5.I, and are plotted in Fig. 5.5. Since the radii of the sheared circles are each unity, it is apparent that the range over which the circles overlap and, as a consequence, $\hat{O}(s)$ has a nonzero value is

$$0 \le s \le 2. \tag{5-68}$$

This range is unaffected by reverting back to the more general situation where $\hat{G}(x, y)$ is neither real nor a constant; the range calculation depends only upon the rotational symmetry assumption and the normalized nature of the reduced spatial frequency s (see the discussion leading up to Eq. (3-47)). Therefore, any quantity directly derived from the nonzero frequency range (such as the cutoff

Table 5.I The Perfect MTF as a Function of Normalized Spatial Frequency

s	MTF	s	MTF
0.0	1.00000	1.1	0.33683
0.1	0.93636	1.2	0.28476
0.2	0.87289	1.3	0.23507
0.3	0.80973	1.4	0.18812
0.4	0.74706	1.5	0.14429
0.5	0.68504	1.6	0.10409
0.6	0.62384	1.7	0.06815
0.7	0.56364	1.8	0.03739
0.8	0.50463	1.9	0.01332
0.9	0.44701	2.0	0.00000
1.0	0.39100		

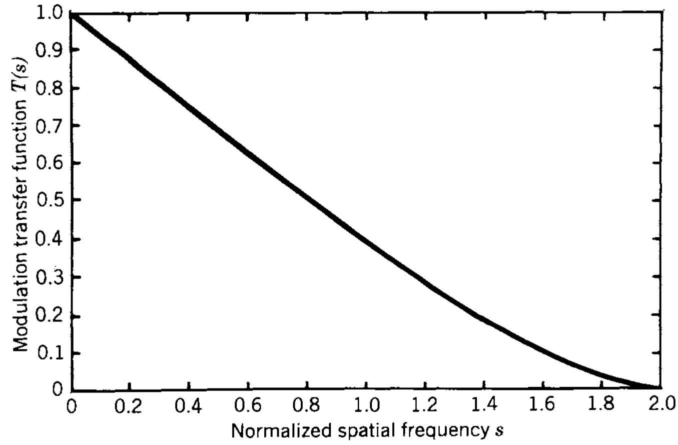


Figure 5.5. A plot of the diffraction-limited, or perfect, MTF.

frequency) has a significance for a more general OTF than the value of the perfect OTF given by Eq. (5-67).

The OTF plot of Fig. 5.5 shows that every optical system within our assumptions functions as a low-pass filter with a cutoff frequency, $s_c = 2$. This can be converted to a real-space cutoff frequency by solving for ω in Eq. (3-52) or (3-53). Near the optic axis, ρ_S and ρ_T are close to unity; so,

$$\omega_c = (2n \sin \alpha)/\lambda, \tag{5-69}$$

and by the definition in Eq. (3-15),

$$\omega_c = 2 \text{ N.A.}/\lambda, \tag{5-70}$$

where n is the refractive index of image space, α (often called *angular aperture*) is half the cone angle subtended by the exit pupil at the image point, λ is the free space wavelength of light, and N.A. is the numerical aperture.

When anamorphotic stretching of the diffraction pattern (see the anamorphotic stretching discussion in Chapter 4) for off-axis object points has to be considered, that is, where ρ_S and ρ_T are not equal, the cutoff frequency will depend on the orientation of the spatial frequency pattern, and

$$\omega_{Sc}/\omega_{Tc} = \rho_T/\rho_S. \tag{5-71}$$

In most optical systems, $\omega_{Tc} > \omega_{Sc}$; so in a test with an optical grating just under the nominal cutoff frequency, the best response usually results when the grating is oriented with its lines of uniform phase perpendicular to a radius line (''radial direction'').

All conceivable (but not necessarily realizable) OTFs can be sorted into three classes according to the imperfections in the optical systems they represent:

1. Ideal (no diffraction, no aberrations).
2. ''Perfect'' (diffraction present, no aberrations).
3. Usual (diffraction and aberrations both present).

The second classification is the one we have dwelt on in this section.

An ideal image-forming optical system (first classification) would be one that functioned so perfectly that the image would be identical to the object even to maintaining the contrast in the finest structure of the object. In terms of the OTF as represented by a plot like Fig. 5.5, the ideal OTF would be graphed as a horizontal line at the ordinate value of unity for all values of the frequency s. Like the ''perfect'' OTF, the ideal OTF would be real for all spatial frequencies; that is, there would be no phase shift and the PTF would be zero for all s. Obviously, since the contrast would be undiminished however great the spatial frequency (however fine the structure), there would be no cutoff frequency. Such an OTF is unattainable and is useful only to let us appreciate by comparison what shortcomings we must accept in a physically realizable system.

In discussing the ''perfect'' OTF, quotation marks are used wherever it seems necessary to emphasize that the characteristic so designated is actually far short of perfection but is rather ''an OTF as perfect as the imperfection caused by the wave nature of light will allow.'' It is the limit (not quite attainable) to which highly corrected systems aspire. The mathematical expression for the perfect OTF based on Eq. (5-15) has shown $\hat{b}_0(s_S, s_T)$ to be a real function. Further, it is an even function symmetric about the origin of the (s_S, s_T) coordinate system (Fig. 5.6). However, values of the function outside the first quadrant (where s_S and s_T are both positive) are of questionable significance unless the meaning of a negative spatial frequency is defined.

All realizable image-forming optical systems have OTFs in our third classification where the parent function $\hat{b}_0(s_S, s_T)$ of Eq. (5-15) is complex, which means not only that the amplitudes of the frequency components are diminished in different amounts but that the phases of the components are shifted according to some PTF. The significance of variable MTFs and PTFs can be discussed by referring to Fig. 5.7, which is a plot of three sine waves of the same frequency. Each wave can be represented by an expression in the form of Eq. (5-38). Curve 1 in the figure represents a frequency component in the object plane. If an ideal OTF applied, curve 1 would also represent the same frequency component in the image plane. However, if the perfect OTF (second classification) applied, curve 2, which is in phase with curve 1 but is of reduced am-

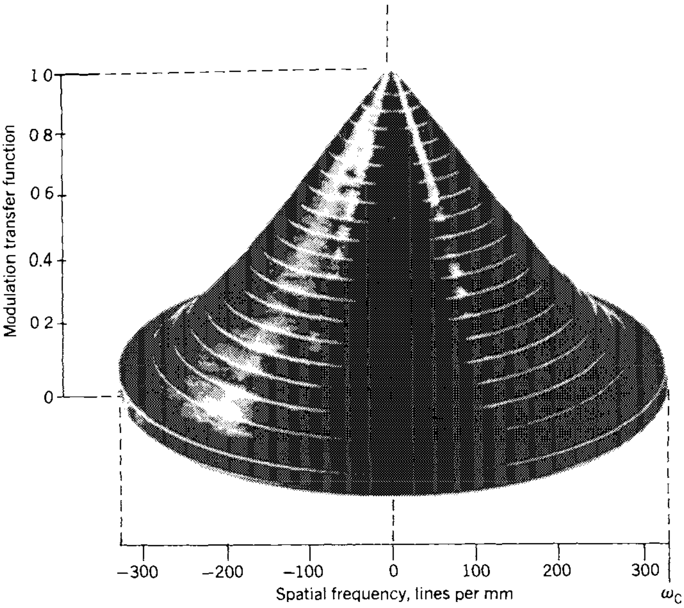


Figure 5.6. A representation of the two-dimensional, perfect OTF [6].

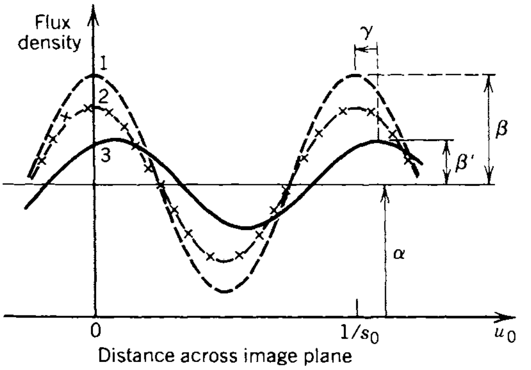


Figure 5.7. Plot of intensity distribution in the image of a sinusoidal test object showing (1) an ideal image, (2) a "perfect" image having reduced amplitude, and (3) an aberrated image having a further reduced amplitude and a phase shift. (From H. H. Hopkins, The Application of Frequency Response Techniques in Optics. In *Proc. of the Conference on Optical Instruments and Techniques, London, 1961*, K. J. Habell (Ed.). Wiley, New York, 1963, p. 487.)

plitude, would represent the corresponding frequency in the image plane. Finally, in the usual system, curve 3, which lags curve 1 by γ and is of reduced amplitude, would be the type of frequency component one would expect to find in the image plane. Of course, in specific instances the amount of reduction in amplitude and both the amount and sign of the phase shift γ (negative in the figure) could be expected to differ from our illustration.

PERFECT OTF FROM SPREAD FUNCTION

In the previous section, the perfect OTF has been formulated as the normalized autocorrelation function of the idealized form of the pupil function $\hat{G}(x, y)$. The same OTF can be derived from the flux-density point spread function, which has already been described as the image of an axial point object formed by a diffraction limited system (no aberrations).

The flux-density point spread function is found by first applying the diffraction integral, Eq. (5-1), to a constant real $\hat{G}(x, y)$ and then squaring the resulting expression for $\hat{U}_0(u, v)$ according to Eq. (5-7) to get the desired $B_0(u, v)$. When symmetry about the axis is taken into account, the one-dimensional expression already discussed as Eq. (2-24) can be substituted for $B_0(u, v)$, with some revision of symbols:

$$\begin{aligned} B_0(2\pi u_0) &= \mathcal{W}_{c0}\left[2J_1(2\pi u_0)\right]^2/(2\pi u_0)^2 \\ &= (\mathcal{W}_{c0}/4\pi^3)\left[2\pi J_1(2\pi u_0)\right]^2/u_0^2 \\ &= (\mathcal{W}_{c0}/4\pi^3)f(u_0). \end{aligned} \tag{5-72}$$

The product $2\pi u_0$ is chosen as the independent variable to accommodate later mathematical manipulation. The grouping of terms to define $f(u_0)$ is done also with a future substitution in mind. $\mathcal{W}_{c0}$ is the flux density at the on-axis image point. The symbol J_1 indicates a Bessel function of the first kind and order one with the argument $2\pi u_0$. The variable u_0 is defined as

$$u_0 = rn' \sin \alpha'/\lambda, \tag{5-73}$$

where $r = (\xi^2 + \eta^2)^{1/2}$ is the real-space radial distance in the image plane from the on-axis image point, n' is the index of refraction in image space, α' is the angle made with the axis by the edge ray to the on-axis image point, and λ is the wavelength of light in vacuum. A review of the definitions of reduced and canonical coordinates in Chapter 3 affirms that Eq. (5-73) is consistent with

$$u_0 = (u^2 + v^2)^{1/2},$$
$$u = u_0 \sin \psi, \qquad v = u_0 \cos \psi. \tag{5-74}$$

The basis for these relations is shown geometrically in Fig. 5.3.

In the introductory statements to Eq. (5-5), it is pointed out that the diffraction integral, Eq. (5-1), is identical in mathematical form to the Fourier integral, Eq. (B-16). Thus, in general, we may relate the complex amplitude in the image plane to the complex amplitude (modulus) of the wave front at the exit pupil by the Fourier transform and its inverse:

$$\hat{U}_0(u, v) \leftrightarrow \hat{G}(x, y), \tag{5-75}$$

which has also been expressed as Eq. (5-6). For the perfect OTF, as in the previous section, $\hat{G}(x, y)$ is assumed a real constant, and the object point is on the optic axis. In the discussion of the Fourier transform in Appendix B, it is shown that if a function is both real and even, its transform must also be real and even; so $\hat{U}_0(u, v)$ is real and even. That $\hat{U}_0$ is even could also have been concluded from the symmetry of the assumed optical system about the optic axis.

In Eq. (5-75), (u, v) are space coordinates in the image plane; so we would expect the independent variables (x, y) of the transform to be spatial frequencies in accordance with the usual significance of Fourier transformation. However, (x, y) are, in fact, space coordinates on the reference sphere at the exit pupil. For comparison we combine Eqs. (5-7) and (5-10) and write

$$B_0(u, v) = \left[\hat{U}_0(u, v)\right]\left[\hat{U}_0^*(u, v)\right] \leftrightarrow \hat{b}_0(s_S, s_T). \tag{5-76}$$

It is apparent that the transform of $\hat{U}_0$ is a function in the space variables (x, y) and the transform of $\hat{U}_0^2$, as indicated in Eq. (5-76), is a function in the spatial frequency variables (s_S, s_T). This suggests an equivalence between the spatial frequencies in the image plane to the space coordinates on the pupil sphere. Making this assumption, we write Eq. (5-8) as

$$\hat{b}_0(x, y) = \iint_{-\infty}^{+\infty} B_0(u, v) \exp\left[-i2\pi(xu + yv)\right] du\, dv. \tag{5-77}$$

On the left side is the function that, when normalized, is the OTF; the right side involves $B_0(u, v)$, which is the spread function. Our present purpose is to apply the perfect OTF assumptions to Eq. (5-77) to get the expression for the perfect OTF already reached in the previous section by another method.

As suggested by the form of Eq. (5-72) for B_0, a transformation of coordinates, rectangular to polar, is necessary in Eq. (5-77). Transformation expressions for the image plane are already given as Eq. (5-74). The corresponding expressions for the exit pupil are

$$\begin{gathered} \rho = (x^2 + y^2)^{1/2}, \\ x = \rho \sin \varphi, \qquad y = \rho \cos \varphi. \end{gathered} \tag{5-78}$$

Application of differential formulas to the expressions for u and v in Eq. (5-74) leads to

$$du\, dv = u_0\, du_0\, d\varphi. \tag{5-79}$$

From the various transformation expressions, Eq. (5-77) can be written as

$$\begin{aligned} \hat{b}_0(\rho) &= \int_{-\infty}^{+\infty} \int_0^{2\pi} B_0(u_0) \exp[-i2\pi\rho u_0(\sin \varphi \sin \psi + \cos \varphi \cos \psi)] \\ &\quad \cdot u_0\, d\psi\, du_0 \\ &= \int_{-\infty}^{+\infty} \int_0^{2\pi} u_0 B_0(u_0) \exp[-i2\pi\rho u_0 \cos(\psi - \varphi)]\, d\psi\, du_0. \end{aligned} \tag{5-80}$$

Breaking down $B_0(u_0)$ as indicated in Eq. (5-72) yields

$$\hat{b}_0(\rho) = (\mathcal{W}_{c0}/4\pi^3) \int_{-\infty}^{+\infty} \int_0^{2\pi} u_0 f(u_0) \exp[-i2\pi\rho u_0 \cos(\psi - \varphi)]\, d\psi\, du_0. \tag{5-81}$$

A well-known formula for Bessel functions [5] can be applied to perform the integration on ψ:

$$J_0(b) = (1/2\pi) \int_0^{2\pi} \exp[-ib \cos(\psi - \varphi)]\, d\psi. \tag{5-82}$$

This reduces Eq. (5-81) to

$$\hat{b}_0(\rho) = (\mathcal{W}_{c0}/2\pi^2) \int_{-\infty}^{+\infty} u_0 f(u_0)\, J_0(2\pi\rho u_0)\, du_0. \tag{5-83}$$

This can be integrated by applying a standard Hankel transform formula [5, p. 145]:

$$\bar{f}(\omega) = \int_0^\infty rf(r)\, J_0(\omega r)\, dr, \tag{5-84}$$

where

$$f(r) = (1/2\pi)\left[J_1(ar)\right]^2/r^2, \tag{5-85}$$

and

$$\bar{f}(\omega) = 2\arccos(\omega/2a) - (\omega/a)\left[1 - (\omega/2a)^2\right]^{1/2} \quad \text{when } 0 < \omega \leq 2a, \tag{5-86}$$

and

$$\bar{f}(\omega) = 0 \qquad \text{when } \omega > 2a. \tag{5-87}$$

When the Hankel transform formula is written in terms of the nomenclature of Eqs. (5-83) and (5-72),

$$\begin{aligned}\bar{f}(\rho) &= \int_{-\infty}^{+\infty} u_0 f(u_0)\, J_0(2\pi\rho u_0)\, du_0 \\ &= 2\arccos(\rho/2) - \rho\left[1 - (\rho/2)^2\right]^{1/2} \\ &= 2\left\{\arccos(\rho/2) - (\rho/2)\left[1 - (\rho/2)^2\right]^{1/2}\right\}.\end{aligned} \tag{5-88}$$

Then, from Eq. (5-83),

$$\begin{aligned}\hat{b}_0(\rho) &= (\mathcal{W}_{c0}/2\pi^2)\bar{f}(\rho) \\ &= (\mathcal{W}_{c0}/2\pi)(2/\pi)\left\{\arccos(\rho/2) - (\rho/2)\left[1 - (\rho/2)^2\right]^{1/2}\right\}.\end{aligned} \tag{5-89}$$

Although ρ was originally defined as a radial coordinate at the exit pupil, Eq. (5-78), it is also equivalent to spatial frequency, as pointed out in the discussion preceding Eq. (5-77); so Eq. (5-89) may be written

$$\hat{b}_0(s) = (\mathcal{W}_{c0}/2\pi)(2/\pi)\left\{\arccos(s/2) - (s/2)\left[1 - (s/2)^2\right]^{1/2}\right\}. \tag{5-90}$$

From the definition of the perfect OTF,

$$\hat{O}(s) = \hat{b}_0(s)/\hat{b}_0(0), \tag{5-91}$$

the expression in Eq. (5-90) gives

$$\hat{O}(s) = (2/\pi)\left\{\arccos(s/2) - (s/2)\left[1 - (s/2)^2\right]^{1/2}\right\}, \tag{5-92}$$

which is identical to Eq. (5-67) arrived at by a different approach.

In summary, Eq. (5-67) for the perfect OTF is reached by assuming an aberration-free pupil function and deriving, first, the spatial frequency function in the image plane and then the perfect OTF. On the other hand, Eq. (5-92) for the perfect OTF is based on the expression describing the flux-density distribution in the image plane resulting from an axial point object in a diffraction-limited system. To arrive at identical expressions for the perfect OTF by these two procedures requires that an equivalence be recognized between the space coordinates on the exit pupil reference sphere and the spatial frequencies in the image plane.

EFFECTS OF CERTAIN ABERRATIONS ON THE OPTICAL TRANSFER FUNCTION

In the previous two sections, the same expression for the perfect OTF, Eqs. (5-67) and (5-92), has been reached by two different approaches. Although a general OTF is a complex function consisting of the product of a real modulus, called the modulation transfer function (MTF), and a factor having an imaginary exponential term, called the phase transfer function (PTF), the perfect OTF is real, which means that the phase shift is zero for all spatial frequencies and that the OTF and MTF are identical. The OTF, MTF, and PTF are discussed earlier in this chapter in connection with Eq. (5-45) and associated equations.

It is of interest now to introduce aberrations into the optical system and to observe the consequent effects on the OTF by comparison with the perfect OTF.

Since our approach to aberrations has been to describe them as distortions of the wave emerging from the exit pupil, we return again to the pupil function stated in this chapter as Eqs. (5-2) and (5-3). Two variables, $G(x, y)$ and $W(x, y)$, each dependent as indicated on location in the exit pupil, occur in the pupil function. The perfect OTF has been derived by assuming $G(x, y)$ constant and $W(x, y)$ zero at all locations; so, to study departures from the perfect OTF, we have to allow either or both variables to take on different values as functions of

(x, y). The way to learn the specific effects of the variables, of course, is to allow only one to vary at a time. With reasonable assumptions, the variation of $W(x, y)$ has much the greater effect on the OTF. Our present discussion will be confined to studying the effects of this variable with $G(x, y)$ held constant. In a later section the reverse, called apodization, is undertaken where $W(x, y)$ is zero or constant and $G(x, y)$ is varied (which can be done experimentally by introducing into the optical system a filter with variable transmission across its face).

As indicated in Eq. (5-15), the transfer function $\hat{b}_0(s_S, s_T)$ is the autocorrelation function of the pupil function $\hat{G}(x, y)$. For a constant amplitude pupil wave, conveniently made unity, the pupil function of Eqs. (5-2) and (5-3) becomes

$$\hat{G}(x, y) = \exp[-i\mathit{k}nW(x, y)],$$

when $(x^2 + y^2) \leq 1$, that is, within a unit-radius circle centered at the origin, and

$$\hat{G}(x, y) = 0, \qquad \text{when } (x^2 + y^2) > 1. \tag{5-93}$$

The transfer function is then

$$\hat{b}_0(s_S, s_T) = \iint_{\alpha} \exp\left\{-i\mathit{k}n[W(x, y) - W(x - s_S, y - s_T)]\right\} dx\, dy, \tag{5-94}$$

where the only nonzero region of integration is the overlap area of the sheared circles of Fig. 5.2, as discussed following Eq. (5-15). To simplify the mathematics, the axes are rotated and the variables changed so that Eq. (5-94) becomes

$$\hat{b}_0(s) = \iint_{\alpha} \exp\left\{-i\mathit{k}n\left[W\left(x + \frac{s}{2}, y\right) - W\left(x - \frac{s}{2}, y\right)\right]\right\} dx\, dy. \tag{5-95}$$

The results of this manipulation are that the two unit-radius circles are centered on the x-axis at $x = -s/2$ and $x = +s/2$, and the component spatial frequencies s_S and s_T have been combined in the single frequency s.

Although the variables of integration (x, y) appear the same in both Eq.

(5-94) and Eq. (5-95), one set of coordinates is, in general, rotated and shifted relative to the other so that certain symmetries in W with reference to the coordinate axes no longer hold in Eq. (5-95). It is also important to note that the (x, y) coordinates *before* rotation are normalized or reduced coordinates defined in Eq. (3-27) (and later shorn of subscripts and primes) to transform all exit pupils from their generally elliptical form in real-space coordinates to circles in reduced coordinates. When it becomes necessary to examine real-space results that are calculated in terms of reduced coordinates, an inverse transformation has to be made. For instance, a real-space diffraction pattern for an off-axis point object is stretched in the y direction compared with its shape in reduced coordinates. Before the inverse transformation to real space is made, any axis rotations to write Eq. (5-95) must first be reversed.

A further simplification in the integrand can be achieved by defining an *aberration difference function* as

$$V(x, y; s) = (1/s)\left[W\left(x + \frac{s}{2}, y\right) - W\left(x - \frac{s}{2}, y\right)\right]. \qquad (5\text{-}96)$$

Then Eq. (5-95) can be written

$$\hat{b}_0(s) = \iint_{\mathfrak{a}} \exp[-iknsV(x, y; s)]\, dx\, dy. \qquad (5\text{-}97)$$

Some useful conclusions can be drawn concerning the *relative modulation* $M(s)$, which may be expressed from Eq. (5-54):

$$M(s) = |\hat{b}(s)|/|\hat{b}(s)|_{\text{perfect}}, \qquad (5\text{-}98)$$

where $|\hat{b}(s)|_{\text{perfect}}$ is the transfer function for a diffraction limited system ("perfect" OTF).

A well-known mathematical relation, Schwartz's inequality [1, p. 63; 4, p. 86], is expressed as

$$\left|\int_a^b f_1(t) f_2(t)\, dt\right|^2 \leq \left[\int_a^b |f_1(t)|^2\, dt\right]\left[\int_a^b |f_2(t)|^2\, dt\right]. \qquad (5\text{-}99)$$

When the statement is extended to two dimensions and the general functions $f_1(t)$ and $f_2(t)$ are replaced by $\hat{G}(x, y)$ of Eq. (5-93) and the shifted function $\hat{G}^*(x - s_S, y - s_T)$ under our simplifying assumptions:

$$\left| \iint_{\mathfrak{a}} [\hat{G}(x, y)] [\hat{G}^*(x - s_S, y - s_T)] \, dx \, dy \right|^2 \leq \iint_{\mathfrak{a}} dx \, dy, \qquad (5\text{-}100)$$

because $|\hat{G}|^2 = |\hat{G}^*|^2 = 1$. Comparison of the left side of Eq. (5-100) with Eq. (5-12) indicates that it represents $|\hat{b}_0(s_S, s_T)|^2$, the squared transfer function for a unit amplitude pupil function, *which may have aberrations*. Review of the previous section in this chapter on "The Perfect OTF" shows that the right side of Eq. (5-100) represents $|\hat{b}_0(s_S, s_T)|^2_{\text{perfect}}$, the squared transfer function for a unit amplitude pupil function *that has no aberrations*. Again, the coordinate axes can be rotated and shifted, as in Eq. (5-95), so that only one frequency s is involved without affecting the conclusions reached concerning Eq. (5-100). Then, according to the definition of Eq. (5-98),

$$M(s) \leq 1 \quad \text{or} \quad M(s_S, s_T) \leq 1. \qquad (5\text{-}101)$$

This inequality shows that with negligible amplitude variations in the pupil function ($G(x, y)$ constant), the effect of aberrations of any kind introduced by $W(x, y)$ is always to produce an additional loss of contrast at all spatial frequencies other than those very near $s = 0$ and $s = 2$. These extreme values can be investigated by applying Eq. (5-95). As s approaches zero, the two unit-radius circles that define the integration area become superimposed and the exponent approaches zero. Thus, the value of the integral approaches π, the area of a unit-radius circle, independently of the value of W, that is, independently of the aberrations in the system. At the other extreme, where s approaches 2, the two circles separate until there is no overlap and, consequently, there is no region of integration where the integrand is other than zero; so the integral becomes zero, again quite independently of the value of W. Summarizing these statements, for all aberrations expressed by a variation of W, we have

$$b_0(0) = \pi, \qquad (5\text{-}102)$$

$$M(0) = 1, \qquad (5\text{-}103)$$

$$b_0(2) = 0, \qquad (5\text{-}104)$$

where the numbers in parentheses are the limiting values of s. (Of course, $M(2)$, a ratio of zeros, is mathematically ambiguous at the limit.)

The value of $M(s)$ at $s = 0$, Eq. (5-103), shows why the gross features, corresponding to very low spatial frequencies, in an object seem not to suffer in poorly corrected systems, while the finer features become indiscernible. Test charts to evaluate the quality of an imaging system usually include a pattern consisting of alternating white and black bars arranged so that narrower and narrower bars can be observed through the system. The maximum number of bar pairs that can be distinguished in a unit distance (maximum spatial frequency) is frequently given as a significant figure of merit for the system. Included in such a numerical evaluation is the degree of the inequality expressed as Eq. (5-101). Though Eq. (5-104) indicates that the normalized frequency $s = 2$ is the cutoff frequency independent of aberrations, the dips in the MTF ($T(s)$ of Eq. (5-48)) curve for a poorly corrected system cause the apparent cutoff frequency to be considerably less than the theoretical value. When the MTF does become zero at some $s < 2$ and then negative, it will ultimately come back to zero at $s = 2$. The generally declining value of the MTF curve as the spatial frequency increases, especially for poorly corrected systems, accounts for the approximations in geometrical optics being most appropriate in the low-frequency (gross features) region of the spectrum.

Since the greatest discrepancy in performance between well-designed and poorly designed optical systems occurs at the high spatial frequencies, quality evaluation of a camera, for instance, often consists of close visual inspection of fine detail, particularly of low contrast, in the photograph.

From the relations given in Eqs. (3-52) and (3-53) between the normalized or reduced spatial frequencies and the real-space spatial frequencies, it is evident that the real space cutoff frequency can be raised by increasing the numerical aperture of the optical system. In Fig. 5.8, comparison is made between systems having cutoff frequencies ω_{c1} and ω_{c2}. Though we might expect better

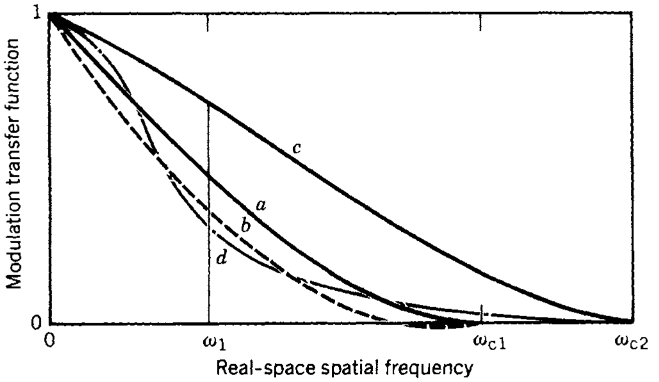


Figure 5.8. Hypothetical MTF curves illustrating (*a*) a perfect MTF with real space cutoff frequency ω_{c1}, (*b*) an MTF for small aberrations with the same cutoff frequency ω_{c1}, (*c*) a perfect MTF with cutoff at a significantly higher frequency ω_{c2}, and (*d*) an MTF curve having the same cutoff at ω_{c2} but having a large amount of aberration, which often results from simply increasing the numerical aperture.

performance from a highly corrected high-cutoff system (curve c) than from a corresponding low-cutoff system (curve a) at some frequency ω_1, we often discover that the situation shown by the broken curves b and d prevails: Increasing the numerical aperture inevitably brings other changes including increased aberrations. The result can be actually poorer performance by the high-cutoff system (curve d) than by the low-cutoff system (curve b).

Deterioration of optical performance caused by defocusing is shown in Fig. 5.9. The curves are for a circularly symmetrical optical system free of aberrations except for the stated maximum displacement of the exit pupil wave front from the reference sphere due to focus error. The number on each curve is n in the expression $n\lambda/\pi$, which gives the maximum distance between the surfaces; so the maximum phase distortion of the wave front is $2n$ radians.

A marked feature of the curves in Fig. 5.9 is the collapse, as s increases, of the response for spatial frequencies above approximately $s = 0.4$, only 20% of the cutoff frequency. Even at $n = 1$, which is only about a third of a wavelength of wave distortion, the response degrades significantly at the higher spatial frequencies.

In Fig. 5.9, there are dips in the curves well below zero, and a fair question is of what physical significance are negative values of $T(s)$. Before this can be answered, more must be known about the whole OTF function, which includes both the $T(s, \psi)$ and the phase transfer function $\phi(s, \psi)$ (see Eq. (5-48)).

A review of the aberration characteristics as described in Chapter 4 reveals that various symmetries occur in the wave aberration function that influence the outcomes of the integrating operations discussed in this chapter. Defocusing represented by Fig. 5.9 has circular symmetry so that variation along any line through the origin is an even function, which would also be true for spherical aberration. The result of the described symmetries for defocusing and spherical aberration (not represented here by a figure) is that each produces a real and even pupil function. By reasoning similar to the discussion of real and even functions in Appendix B, it is found that the autocorrelation of a real and even function produces a real and even function; so $\hat{O}(s, \psi)$, the OTF, must be real and even. This in turn requires that $\phi(s, \psi)$ be zero for all values of s and ψ, with a notable exception. From the theory of complex numbers, it is known that a negative $T(s, \psi)$ and a zero $\phi(s, \psi)$ are equivalent to a positive $T(s, \psi)$ of the same absolute value and a $\phi(s, \psi) = \pi$ radians or 180°. The latter interpretation turns out to be the useful one to explain the negative curve values in Fig. 5.9. As a $T(s, \psi)$ curve passes from positive to negative values, the sinusoidal phase of the spatial frequency reverses; that is, an alternating black and white object at this frequency suddenly reverses so that white appears in the image where black formerly occurred and vice versa. This effect is shown for a defocused sector grating image in Fig. 5.10.

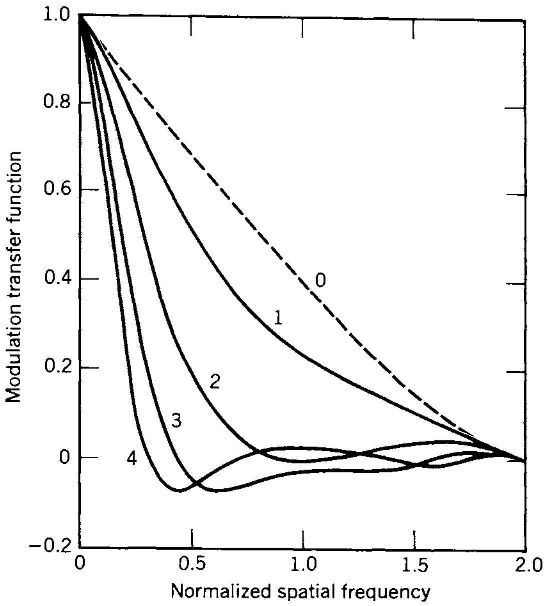


Figure 5.9. Calculated MTF curves for an optical system that is free of aberrations but has a defect of focus. The wave-front distortion at the edge of the pupil is $n\lambda/\pi$ in which n is the number of each curve [7].

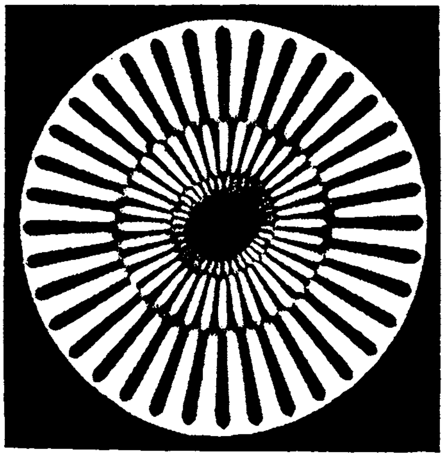


Figure 5.10. A photograph of spurious resolution. (From P. Lindberg, Measurement of Contrast Transmission Characteristics in Optical Image Formation. *Opt. Acta* **1**, 80 (1954).)

When a $T(s, \psi)$ curve dips into the negative region, even though it may emerge into the positive region for one or more intervals at higher frequencies, the resolution is said to be *spurious* at all frequencies above the first crossover. For instance, the curve for $n = 4$ in Fig. 5.9 crosses the $T(s, \psi) = 0$ value at s values of approximately 0.28, 0.65, 1.41, and 1.65, being negative for s between 0.28 and 0.65 and also for s between 1.41 and 1.65. The curve is positive for all other values; however, the resolution is said to be spurious for all frequencies above $s = 0.28$.

A study of astigmatism is given in Fig. 5.11 for $T(s, \pi/6)$, which, according to the nomenclature used in Eq. (5-53), is for a spatial frequency orientation of $\psi = \pi/6$ radians or 30°. (See Fig. 5.1 and Fig. 5.3 for the definition of the angle ψ.) The details of calculating the OTF shown in Fig. 5.11 are given in connection with Fig. 9.6.

Instead of the familiar MTF versus s presentation, Fig. 5.11 is an *Argand diagram* (named after a mathematician who pioneered in complex numbers). When the OTF, $\hat{O}$, is to be expressed as a complex quantity, it can be in either polar or rectangular form as indicated by the equation in Fig. 5.11. The curve is the locus of points on the complex plane (real numbers a on the abscissa axis, imaginary numbers b on the ordinate axis) for values of s from 0.0 to 0.7. The geometrical relation of the polar and rectangular forms is indicated for $s = 0.2$. The length of the arrow is the MTF and the angle (here negative) of the arrow with the a-axis is the PTF. The tip of the arrow is at the rectangular coordinates (a, b).

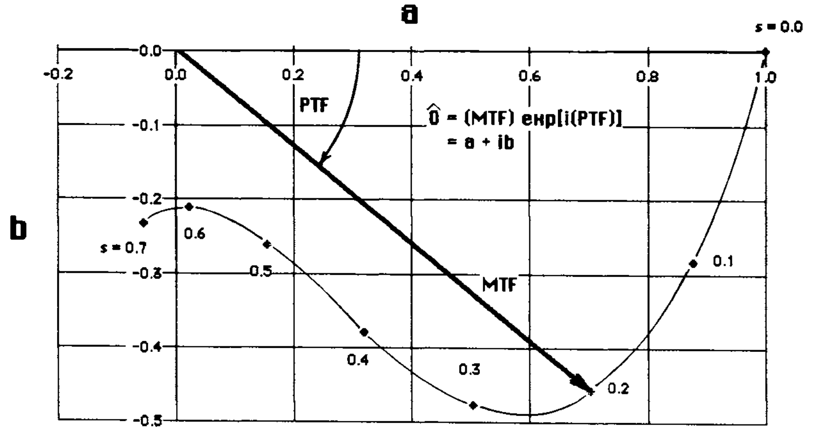


Figure 5.11. An optical transfer function plotted on a complex plane for astigmatism and a spatial frequency orientation of 30°. (See discussion of Fig. 9.6 in Chapter 9.)

APODIZATION

The shape of an MTF curve can be changed in certain preplanned ways, and the most usual method of doing so is to operate on a wave front as it passes through the exit pupil. Because the purpose of this process is to alter the shape of a response curve, the term *apodization* is used, although this usage departs from the word's etymology since the root, *apodize*, suggests to "remove the feet." The wave front is caused to pass through an optical filter that can be designed to act on the wave front in one or both of two ways: by absorbing energy (reducing amplitude) and by introducing a phase delay. By design these effects are made to vary across the face of the filter so that amplitude and phase become functions of the exit pupil coordinates.

In Fig. 5.12 a ray schematically represents a wave-front incident upon an optical filter. The particular ray indicated enters the surface at the general point (x_1, y_1) with a complex amplitude of $\hat{G}_1(x_1, y_1)$, and after losses caused by absorption and multiple internal reflections, the wave front emerges with an amplitude of $\hat{G}_2(x_1, y_1)$. The ratio of $\hat{G}_2(x_1, y_1)$ to $\hat{G}_1(x_1, y_1)$ is defined as the amplitude transmittance $\hat{\tau}(x_1, y_1)$ of the filter at (x_1, y_1). Expressed generally,

$$\begin{aligned}\hat{\tau}(x, y) &= \tau(x, y) \exp[i\alpha(x, y)] \\ &= \hat{G}_2(x, y)/\hat{G}_1(x, y) \\ &= G_2/G_1 \exp[i(\beta_2 - \beta_1)]. \qquad (5\text{-}105)\end{aligned}$$

In apodization design, results are achieved by manipulating the real amplitude of the wavefront rather than the phase shift or a combination of the two. To minimize the introduction of aberrations, in fact, the phase is held as constant as possible as a function of (x, y). With constant phase, $\alpha(x, y)$ is con-

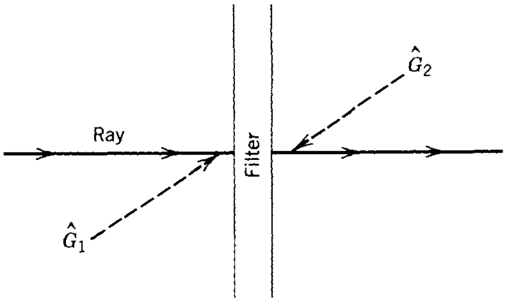


Figure 5.12. A schematic to illustrate absorption as a function of coordinates in a partially transparent slab.

stant as is the difference $\beta_2 - \beta_1$. (The magnitude of α, as long as it is held constant, is of no consequence in this analysis.) If it is assumed that the incident wave front $G_1(x, y)$ is constant, the generality of our development is not affected by assigning unity as its value. Then, from Eq. (5-105), we can write

$$\tau(x, y) = G_2(x, y). \tag{5-106}$$

As in most of the optics discussed in this book, radial symmetry is assumed, which makes polar coordinates particularly convenient:

$$x = \rho \sin \varphi, \qquad y = \rho \cos \varphi, \tag{5-107}$$

so that τ becomes a function of ρ:

$$\tau(\rho) = G_2(\rho). \tag{5-108}$$

Four hypothetical examples of apodization are illustrated in Figs. 5.13 and 5.14.

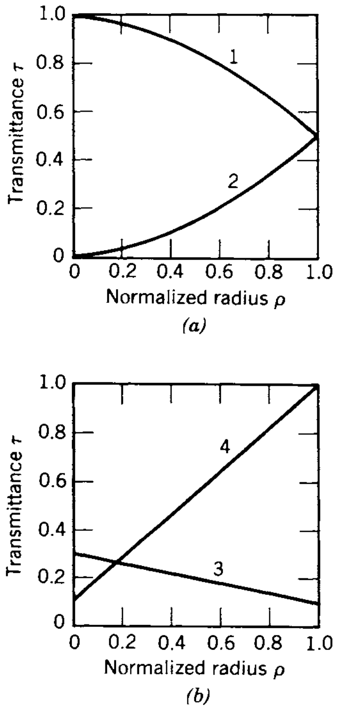


Figure 5.13. Apodization: Absorption by filters placed in the exit pupil according to the transmittance curves in (*a*) and (*b*).

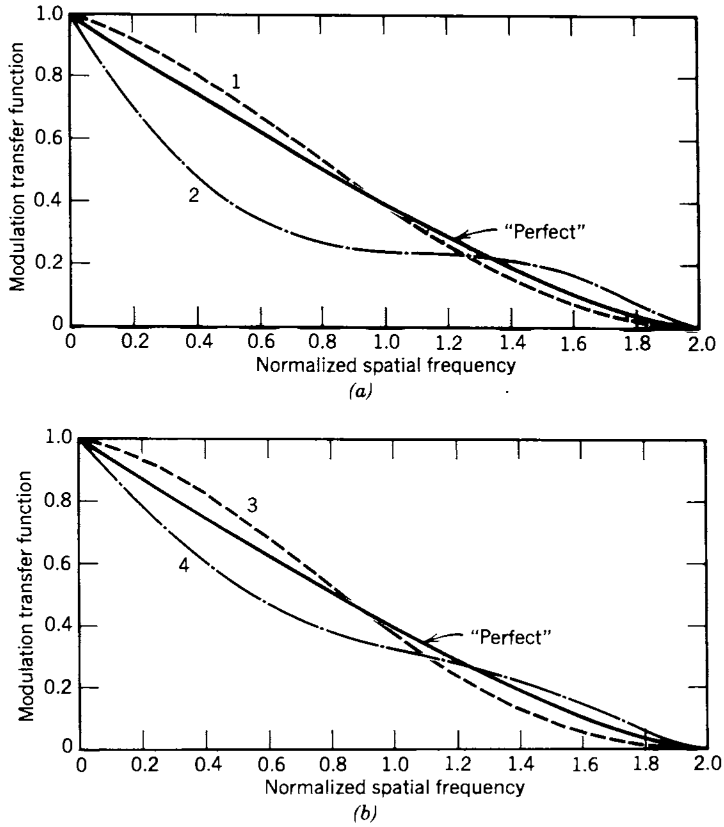


Figure 5.14. Apodization: MTF curves resulting from the transmittance characteristics of (*a*) Fig. 5.13*a* and (*b*) Fig. 5.13*b*.

Figure 5.13 shows the assumed filter transmittance for each example as a function of the normalized distance ρ from the optic axis, and Fig. 5.14 shows the calculated modulation transfer functions corresponding to the four transmittance functions. For reference, the perfect OTF, expressed in Eq. (5-67), is plotted as the solid curve in both parts of the figure. The four examples have transmittance functions as follows:

Example 1: $$\tau(\rho) = \cos^2(\pi\rho/4). \tag{5-109}$$

Example 2: $$\tau(\rho) = \sin^2(\pi\rho/4). \tag{5-110}$$

Example 3: $$\tau(\rho) = 0.3 - 0.2\rho. \tag{5-111}$$

Example 4: $$\tau(\rho) = 0.9\rho + 0.1. \tag{5-112}$$

The curves are numbered to correspond to the example numbers. The angles in the first two examples are in radians. As in previous MTF curves having the reduced frequency s as the independent variable, the range of frequency for the curves of Fig. 5.14 is

$$0 \leq s \leq 2. \tag{5-113}$$

Calculation of MTF curves from the exit pupil filter characteristics is based on the autocorrelation of the pupil function expressed as Eq. (5-15); the corresponding diagram of nonzero areas of integration is shown in Fig. 5.2. If the axes are rotated so that both unit-radius circles are centered on the x-axis and if the phase of $G(x, y)$ is constant (and, therefore, assumed zero), Eq. (5-15) becomes

$$b_1(s_n) = \iint_{\mathcal{A}} G(x, y)\, G(x - s_n, y)\, dx\, dy. \tag{5-114}$$

The pupil centers are thus separated by s_n, and s_n ranges from zero to 2. The area $\mathcal{A}$ is the common overlapping area of the two sheared circles.

For numerical calculation of $b_1(s_n)$, Eq. (5-114) can be modified, first, by shifting the x-axis for symmetry, and then, with the substitution permitted by Eq. (5-108), by writing an equivalent summation expression:

$$b_1(s_n) = \iint_{\mathcal{A}} G\left(x + \frac{s_n}{2}, y\right) G\left(x - \frac{s_n}{2}, y\right) dx\, dy, \tag{5-115}$$

$$b_n = 4 \sum_j \sum_i \tau(\rho_{1ij})\, \tau(\rho_{2ij})\, \Delta y\, \Delta x. \tag{5-116}$$

As indicated in Fig. 5.15, the numerical integration of Eq. (5-116) is performed in the first quadrant; then, by taking advantage of symmetry, the integral over the whole sheared area is obtained by multiplying the first quadrant integral by four. The range of x for the numerical integration is from the origin to $(1 - s_n/2)$; the range of y is from the origin to the positive ordinate of the left-hand circle, $[1 - (s_n/2 + x)^2]^{1/2}$. This integrating procedure can be used for the entire range of s_n including the most lengthy integration, for $s_n = 0$, where the two circles are superposed with centers at the origin. Time can be saved, however, for this one integration by taking advantage of the simplified geometry as indicated in Fig. 5.16. The summation expression is

$$\begin{aligned} b_n(0) &= 4 \sum_i \frac{\pi \rho_i}{2} \tau(\rho_i)\, \Delta\rho \\ &= \sum_i 2\pi\rho_i \tau(\rho_i)\, \Delta\rho. \end{aligned} \tag{5-117}$$

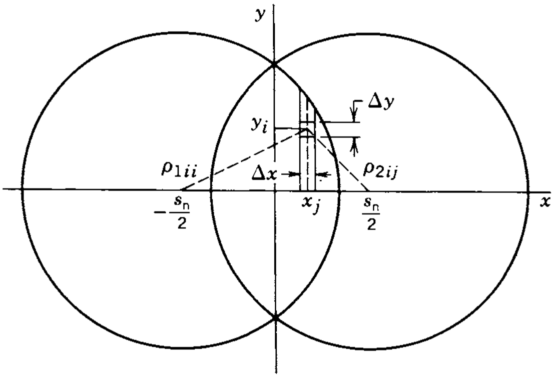


Figure 5.15. Numerical integration coordinates for calculating values on the MTF curves from transmittance characterstics.

Comparison of the filter characteristics of Fig. 5.13 with the corresponding MTF curves of Fig. 5.14 suggests that characteristics with negative slopes (examples 1 and 3) produce MTF curves that are above the "perfect" curve at lower spatial frequencies and lower than the "perfect" curve at higher frequencies. The reverse appears to be true of characteristics with positive slopes (examples 2 and 4). These observations lead to the question of how even more extreme filter characteristics would affect the MTF curves. The extremes selected to answer this question are shown in Fig. 5.17. In each of the two filters represented, the unit-radius circle is divided by a concentric boundary so that half the total area has 100% transmission and the other half is opaque. The first, example 5 of the present series, has an opaque outer ring and is described by

$$\text{Example 5:} \quad \tau(\rho) = 1, \quad 0 \le \rho < +\sqrt{0.5}.$$
$$\tau(\rho) = 0, \quad 0 > \rho \ge +\sqrt{0.5}. \qquad (5\text{-}118)$$

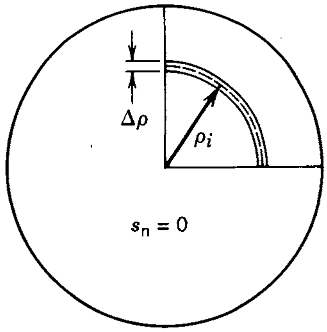


Figure 5.16. Simplified numerical integration coordinates for $s_n = 0$.

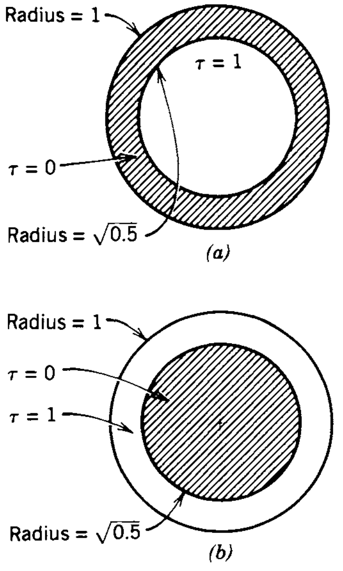


Figure 5.17. (*a*) Filter with an opaque ring around a zero absorption circle. (*b*) Filter with a zero absorption ring around an opaque circle.

The second, example 6, is described by

$$\text{Example 6:}\quad \begin{aligned} \tau(\rho) &= 0, & \rho &< +\sqrt{0.5}. \\ \tau(\rho) &= 1, & +\sqrt{0.5} &\leq \rho \leq 1. \end{aligned} \qquad (5\text{-}119)$$

Coincidentally, example 5 illustrates the effect of reducing the numerical aperture, and example 6 illustrates the effect of a central obscuration. The filter characteristic functions for examples 5 and 6 are plotted in Figs. 5.18*ab*, respectively. The resulting MTF curves are shown in Fig. 19. Although neither MTF curve is above the "perfect" curve, except for a short interval at high frequencies in example 6, the frequency discrimination tendencies of the earlier examples are borne out. Example 5, in which transmittance abruptly drops from unity to zero with increasing radius, has an MTF curve that drops to zero at about 71% of the cutoff frequency of the other curves. Example 6, in which the transmittance abruptly rises from zero to unity with increasing radius, has a relatively flat MTF from a normalized frequency of 0.4 to 1.6 instead of the steadily declining value of the "perfect" curve.

Because more than the overall slope of the transmittance curve is involved in determining the shape of the corresponding MTF curve, we have to limit our generalizations to noting that transmittance near the optic axis tends to favor

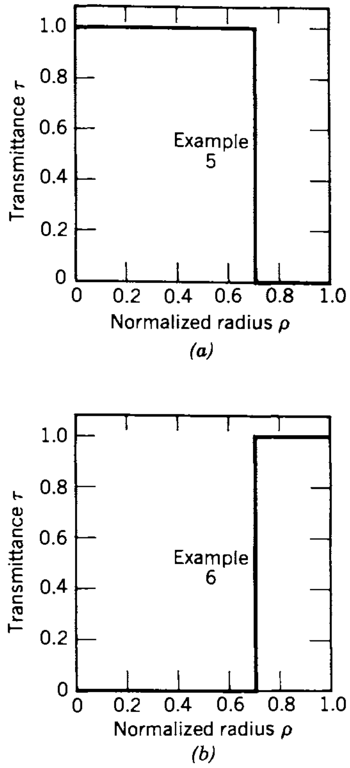


Figure 5.18. Transmittance characteristics for the filters of Fig. 5.17.

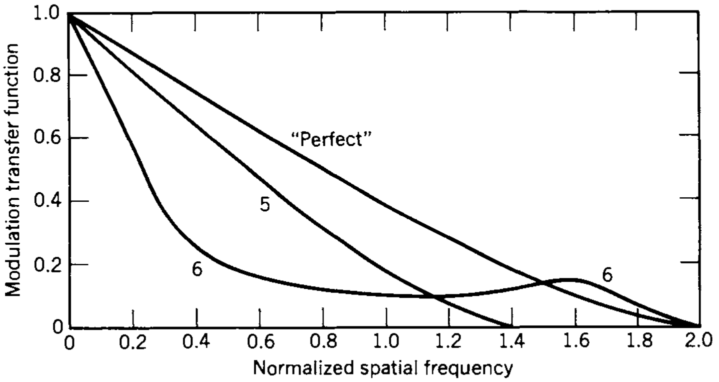


Figure 5.19. MTF curves for the filters of Fig. 5.17.

the lower frequencies and transmittance near the periphery emphasizes the higher frequencies.

In all applications of apodization filters, results are achieved by a subtractive process: Part of the input optical energy is being deliberately removed before it reaches the image plane. Compensating design changes, such as increasing the optical aperture, increasing the source strength, and increasing exposure time, are often required. It is recognized that some of the filter characteristics assumed in our six hypothetical examples might be difficult to realize with available techniques and materials, thus making certain compromises necessary to achieve similar MTF results.

THE GEOMETRICAL OPTICS OTF APPROXIMATION

A number of authors have discussed calculation of the OTF by a geometrical optics approximation [9–11] which, if valid, could save appreciable labor during early design stages. Briefly, the method consists of mapping rays from an element of area on the wave surface at the exit pupil onto the image plane by a transformation involving Jacobians. The result in the image plane is a ray density, which is interpreted as flux density. This is converted by a Fourier transform to a spatial frequency distribution. Typical results from OTF calculations by this approximation are given in Refs. 10–13.

In lens design or evaluation, certain merit functions based, for instance, on the mean square value of classical ray aberrations may ultimately be replaced by merit functions based on physical optics or the OTF. It is conceivable that the geometrical optics approach to approximating the OTF could smooth the transition by indicating some sort of equivalence between the two sets of merit functions. Some question remains, however, as to the validity of the assumptions in the geometrical approximation.

A transformation involving Jacobians is a standard procedure in certain physics problems dealing with irrotational flow of a perfect fluid through a narrow tube of changing cross section. Equipotential surfaces, which are analogous to wave fronts, are perpendicular to stream lines, analogous to rays. The analogy to electromagnetic energy flow from exit pupil to image plane breaks down for several reasons. In ray optics, rays must sometimes cross or intersect, which brings into question the one-to-one point correspondence that is basic to a coordinate transformation by Jacobians in a double integral, especially when aberrations are present. A given element of area on the image plane typically receives light from more than one element on the pupil. In some instances the theoretical ray density distribution in the image becomes infinite at certain points, such as at a focus, where the Jacobian becomes zero, and delta functions are required to handle the representation. Singularities in general invalidate the simple model. Finally, to be strictly correct, the pupil coordinates must be on the reference sphere rather than on the pupil plane as assumed in the transformation involving Jacobians.

At best, the indicated discrepancies make the geometrical optics OTF approximation suspect. Trying to make the approximation more accurate may well add complexities that would compromise its utility, the reason for its use in the first place. A complete changeover from ray optics and aberrations to physical optics and the OTF appears presently to be the preferred route.

THE POLYCHROMATIC OTF

Many discussions of the OTF go on to include what might be considered an OTF resulting when the light beam illuminating the optical system is white light. In some quarters the evaluation of optical systems by means of a polychromatic OTF is almost standard practice. But there are serious pitfalls associated with the polychromatic OTF because optical systems are not color-blind. The formula for the cutoff spatial frequency, Eq. (5-69), tells us that the cutoff for blue light is different from the cutoff for red light because of the dependence on wavelength. Of course, an optical system may be achromatized to place the focus for two or three wavelengths at the same focal point; and we might, therefore, conclude that the caustic of rays converging toward the focal plane would have the same size and shape. Nevertheless, at present we have no way to predict what the wave-front shape, the wave aberration function, and the OTF at these same wavelengths and at intermediate wavelengths would be. The three things might differ appreciably. An evaluation based on the polychromatic OTF of an optical system having these unknown characteristics would at least always be suspect. In fact, it has been shown that lenses with completely different chromatic aberrations can have the same polychromatic OTF. Some sort of average of the OTF at two or three different wavelengths cannot provide a reliable assessment criterion.

Several papers have been published to point out the danger of evaluating lenses on the bases of color-blind polychromatic OTFs. Barnden [17, 18] has given two important conditions that must be satisfied for the polychromatic OTF to be used:

1. The light beam from the object must have a constant spectral composition, both spatially and temporally, and the detector must have a uniform spectral sensitivity.

2. Variations of the local magnifications, m_S and m_T, with wavelength must be insignificant.

The necessity to meet these conditions adds another degree of complexity to the already difficult problem of designing high-quality optical systems. There are, therefore, difficult problems in the practical application of the polychromatic OTF to the specification and evaluation of optical systems.

Takeda [19] goes further into the difficulties involved with polychromatic OTFs. An analytic technique—chromatic balancing—is proposed that will ascertain the chromatic aberrations that are different and yet give the same polychromatic OTF for a specified spectral combination of the light source and the detector. A typical numerical example is given that supports the theory and the conclusions. Takeda emphasizes that more theoretical and experimental studies must be made before reasonable and generally acceptable use of the polychromatic OTF can be made.

Because there has been a tendency to consider and apply the polychromatic OTF too naively and because we believe that further study is sorely needed, the polychromatic OTF will not be considered further in this book.

REFERENCES

1. A. Papoulis, *The Fourier Integral and Its Application.* McGraw-Hill, New York, 1962.
2. E. C. Kintner and R. M. Sillitto, Edge Ringing in Partially Coherent Imaging. *Opt. Acta* **24**, 59 (1977).
3. W. T. Cathey, *Optical Information Processing and Holography.* Wiley, New York, 1974.
4. E. L. O'Neill, *Introduction to Statistical Optics.* Addison-Wesley, Reading, MA, 1963.
5. A. Papoulis, *Systems and Transforms with Applications in Optics.* McGraw-Hill, New York, 1968.
6. S. H. Lerman, Application of the Fast Fourier Transform to the Calculation of the Optical Transfer Function. *SPIE Proc.* **13,** 51 (1969). (Please see the note following Ref. 2 of Chapter 1.)
7. H. H. Hopkins, The Frequency Response of a Defocused Optical System. *Proc. R. Soc. London Ser. A* **231**, 91 (1955).
8. M. De, The Influence of Astigmatism on the Response Function of an Optical System. *Proc. R. Soc. London Ser. A* **233**, 91 (1955).
9. H. H. Hopkins, Geometrical-Optical Treatment of Frequency Response. *Proc. Phys. Soc. London Ser. B* **70**, 1162 (1957).

10. K. Miyamoto, On a Comparison between Wave Optics and Geometrical Optics by Using Fourier Analysis, I. General Theory. *J. Opt. Soc. Am.* **48**, 57 (1958).

11. K. Miyamoto, Wave Optics and Geometrical Optics in Optical Design. In *Progress in Optics*, Vol. 1, E. Wolf, (Ed.). North Holland, Amsterdam, 1961, p. 31.

12. N. S. Bromilow, Geometrical-Optical Calculation of Frequency Response for Systems with Spherical Aberration. *Proc. Phys. Soc. (London) Ser. B* **71**, 231 (1958).

13. A. S. Marathay, Geometrical Optical Calculation of Frequency Response for Systems with Coma. *Proc. Phys. Soc. (London)* **74**, 721 (1959).

14. Guide to Optical Transfer Function Measurement and Reporting. ANSI PH3.57-1978, American National Standards Institute, New York.

15. J. E. Wilkins, Jr., Apodization for Maximum Strehl Ratio and Specified Rayleigh Limit of Resolution, I. *J. Opt. Soc. Am.* **67**, 1027 (1977); II. *J. Opt. Soc. Am.* **69**, 1526 (1979).

16. M. J. Yzuel and F. Calvo, A Study of the Possibility of Image Optimization by Apodization Filters in Optical Systems with Residual Aberrations. *Opt. Acta* **26**, 1397 (1979).

17. R. Barnden, Calculation of Axial Polychromatic Optical Transfer Function. *Opt. Acta* **21**, 981 (1974).

18. R. Barnden, Extra-Axial Polychromatic Optical Transfer Function. *Opt. Acta* **23**, 1 (1976).

19. M. Takeda, Chromatic Aberration Matching of the Polychromatic Optical Transfer Function. *Appl. Opt.* **20**, 684 (1981).

6

Optical Design and Image Criteria

THE NATURE OF OPTICAL DESIGN

The OTF has appreciable potential for use in specifying, designing, and evaluating optical systems; this book, then, responds to an obligation by discussing such associated topics as *optical design, image quality, merit functions,* and *evaluation criteria*—especially as they relate to ways that the OTF may be applied. This chapter and the next are devoted to discussions of these topics. *Optical design*, the topic of this chapter, is the process of finding the descriptions of individual elements and their arrangement in a desired optical system. The descriptions provide the information as specifications necessary to manufacture the system.

The designer starts with an idea for a needed system that will solve a specific imaging problem. The required image implies certain imaging characteristics that have been variously specified using such terms as resolving power, acutance, MTF, permissible distortion, field of view, $f/$number or numerical aperture, encircled energy, and range of color. Also to be met are the physical requirements such as size of package, size of image, and object–image distance. The ultimate in convenience would be to go directly from the requirements to the specifications, but it does not work that way. Instead, designers must draw upon the optical-design literature and their own background and experience with other systems to make a guess as to the assembly of elements needed to produce the desired optical characteristics.

In the early days of optical design, designers could usually assemble a system from available, or easily shaped, elements, and could detect by visual examination of the image such imperfections as chromatic aberration, astigmatism, or coma. Not so today. Even if it were economically feasible to construct an optical system at intermediate stages of the design so that laboratory tests could guide further refinements, visual inspection methods could no longer achieve the correction required in present-day systems.

Nowadays, designers assume a system by listing a set of initial construction *parameters* as a first guess in describing the desired system. For each element the parameters typically are the index of refraction and dispersion of the material, the diameter and thickness of the element, and the curvatures of the two surfaces. The separation and order of the elements are also part of the design. For the system to be realizable, a suitable optical material must be available that has the necessary index of refraction and dispersion. The material also must be transparent over a specified wavelength range. Generally, the material will have to be selected from *available* glasses that are cataloged, with their characteristics, by the several glass suppliers.

Fortunately the lore of optical technology is well developed in many ways. Libraries of existing lenses, proved by use, are especially helpful. Three examples of patented lenses and some of their characteristics are shown in Figs. 6.1–6.3. Tables 6.I, 6.II, and 6.III give the construction data for these lenses. Cox [1], among sources of this kind of information, lists 240 patented lenses and gives the construction data for each one.

Calculated diffraction patterns for two systems having a wavelength of coma are shown in Fig. 6.4.

The designer will first trace a few key rays through the system using the assumed parameters, and then will note the intersection points of these rays with a chosen initial image plane. A small number of well-chosen rays suffice to guide reduction of the gross Seidel aberrations. A study of the transverse deviations of certain rays and the distribution of ray intersection points are useful in revealing types and amounts of aberrations. After these preliminary adjustments, optical design becomes more subtle. Changes of parameters to reduce aberrations further and to bring the system more nearly into its physical requirements require almost intuitive judgment on the part of the designer, more art than science. This ''feel'' usually has to be exercised to some degree before automatic lens design can be applied. We note for later discussions of the OTF that this same artistic skill based on ray analysis probably does not carry over to a corresponding skill based on OTF analysis for making a parameter improvement, like a small change in a surface curvature, from a calculated or measured OTF. In fact, R. Barakat [2] suggests that the fundamental problem of lens design is finding an analytic expression relating image quality criteria (like some characteristic of the OTF) to design parameters. Optical design will likely continue to be very much an art as well as a science in spite of the contributions of fast, large-capacity computers and ingenious software. Instances of a complete automatic lens design starting from an elementary system are rare. Each designer generally employs a kit of special-purpose computer programs as tools to fashion optical systems according to the designer's particular experience and ability. Our discussion, for tutorial reasons, deals with more general procedures.

To accomplish the more subtle adjustments, optical design can be turned over to a computer. The subsequent procedure, referred to in this book as *au-*

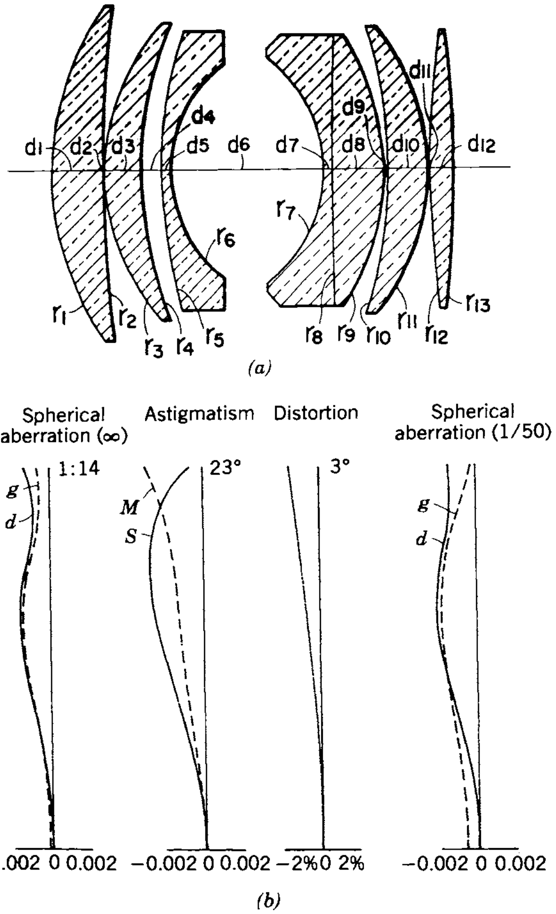
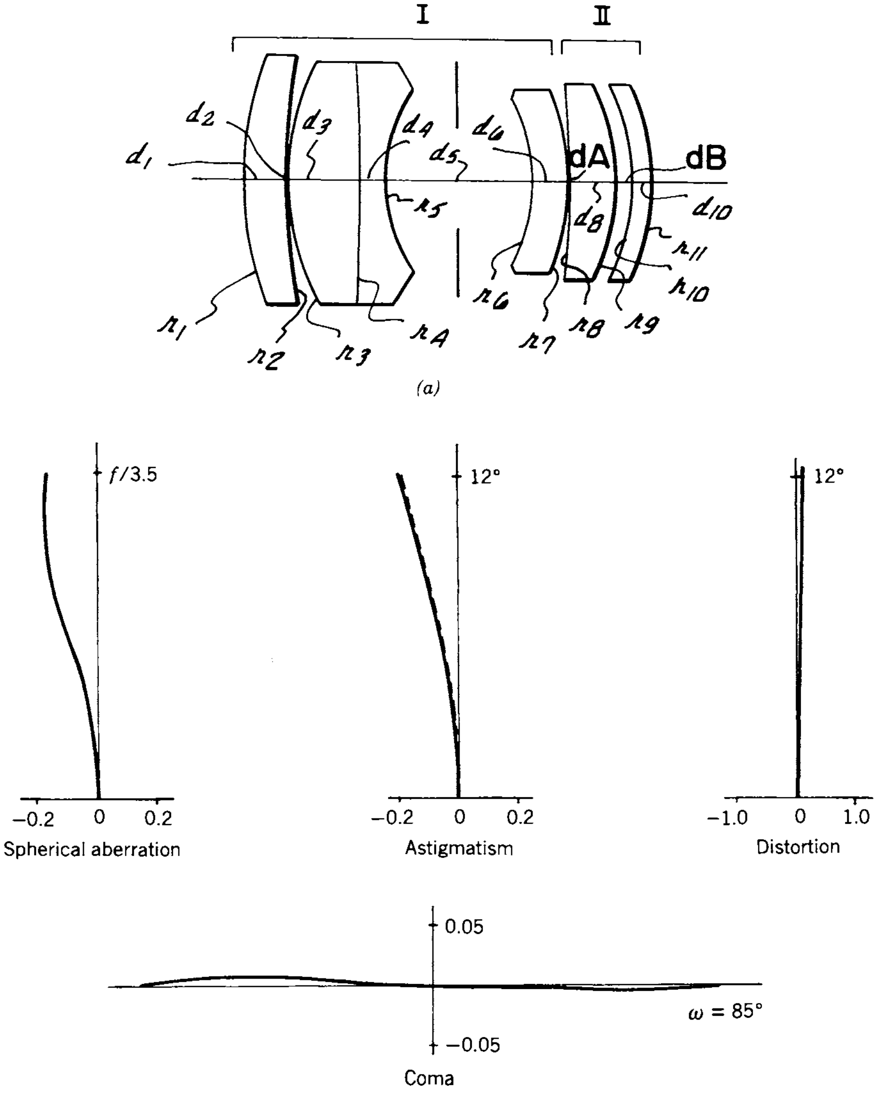


Figure 6.1. (*a*) The Gauss-type photographic lens of U. S. Patent Number 4,094,588. (*b*) Transverse aberrations.

tomatic lens design, is calculation by electronic computer according to programmed instructions. The program sequence typically makes small changes in one or more of the parameters and then tests the altered system to determine whether an improvement has been made. For testing image quality the computer calculates an appropriate *merit function* that gauges the image quality in a way preferred by the designer. When every possible further parameter change degrades the image quality, the automatic design procedure is programmed to stop. The design is then superior to any other design in which the value of each



(a)

(b)

Figure 6.2. (*a*) The objective lens of U. S. Patent Number 4,165,916 with close object focusing aberration correction. Axial spacings *dA* and *dB* are adjustable. (*b*) Transverse aberrations for the lens focused at infinity.

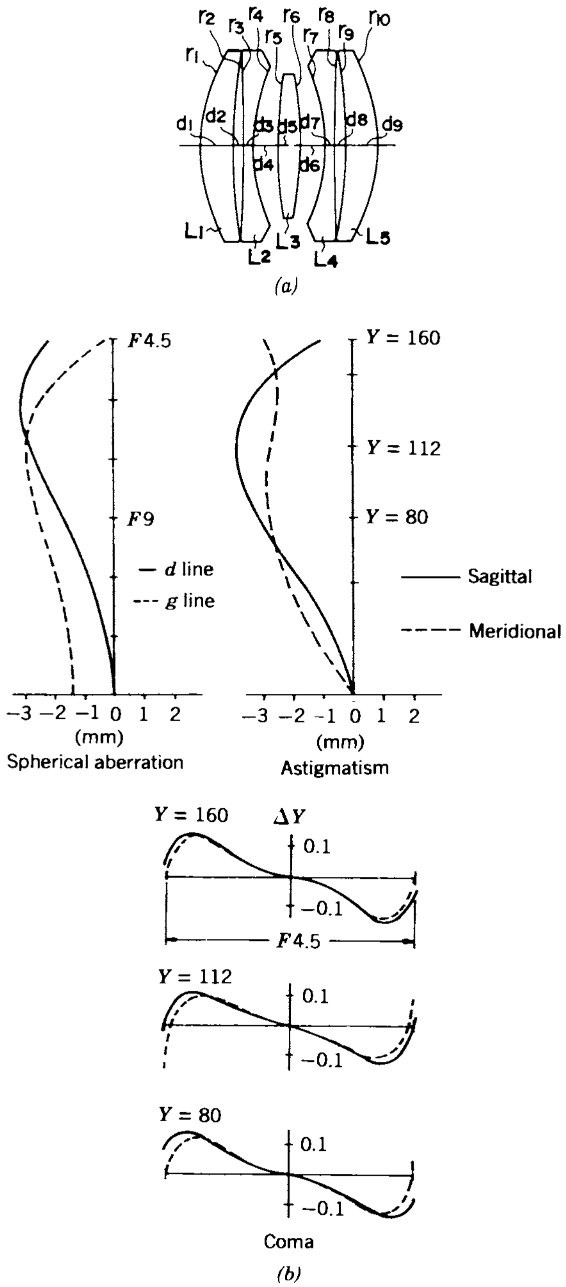


Figure 6.3. The small copy lens of U. S. Patent Number 4,173,396.

Table 6.I Construction Data for Lens of Fig. 6.1

Radius of Curvature	Axial Distance	Refractive Index	Abbe Number
$r_1 = 0.6147$			
	$d_1 = 0.1158$	$n_1 = 1.6204$	$\nu_1 = 60.3$
$r_2 = 2.9052$			
	$d_2 = 0.0019$		
$r_3 = 0.4738$			
	$d_3 = 0.0777$	$n_2 = 1.6935$	$\nu_2 = 50.8$
$r_4 = 0.8002$			
	$d_4 = 0.0468$		
$r_5 = 1.0617$			
	$d_5 = 0.0203$	$n_3 = 1.5814$	$\nu_3 = 40.8$
$r_6 = 0.2764$			
	$d_6 = 0.3357$		
$r_7 = -0.2984$			
	$d_7 = 0.0193$	$n_4 = 1.7552$	$\nu_4 = 27.5$
$r_8 = 15.8469$			
	$d_8 = 0.1160$	$n_5 = 1.6935$	$\nu_5 = 53.3$
$r_9 = -0.5113$			
	$d_9 = 0.0019$		
$r_{10} = -1.0617$			
	$d_{10} = 0.0890$	$n_6 = 1.8061$	$\nu_6 = 40.9$
$r_{11} = -0.4676$			
	$d_{11} = 0.0023$		
$r_{12} = 1.8784$			
	$d_{12} = 0.0570$	$n_7 = 1.6935$	$\nu_7 = 50.8$
$r_{13} = -2.2973$			
	$f = 1.0$	$f_B = 0.7430$	

parameter is in the neighborhood of the value pertaining when the design stopped. The system is then said to be optimized.

The design background for an optical system designer includes:

1. A knowledge of geometrical optics, which gives a picture of how a lens works.
2. A mental catalog of lenses and the general performance characteristics of the various types.
3. A general knowledge of the characteristics of the various kinds and orders of aberrations.

Table 6.II Construction Data for Lens of Fig. 6.2 when Lens is Focused at Infinity[a]

Radius of Curvature	Axial Distance	Refractive Index	Abbe Number
$r_1 = 0.452$			
	$d_1 = 0.048$	$n_1 = 1.7755$	$\nu_1 = 37.9$
$r_2 = 0.788$			
	$d_2 = 0.003$		
$r_3 = 0.293$			
	$d_3 = 0.082$	$n_2 = 1.6779$	$\nu_2 = 55.5$
$r_4 = -16.234$			
	$d_4 = 0.028$	$n_3 = 1.6545$	$\nu_3 = 33.9$
$r_5 = 0.201$			
	$d_5 = 0.174$		
$r_6 = -0.233$			
	$d_6 = 0.04$	$n_4 = 1.6815$	$\nu_4 = 36.8$
$r_7 = -0.348$			
	$d_7 = 0.003$	(dA: Variable)	
$r_8 = -1.916$			
	$d_8 = 0.05$	$n_5 = 1.6589$	$\nu_5 = 56.5$
$r_9 = -0.282$			
	$d_9 = 0.02$	(dB: Variable)	
$r_{10} = -0.271$			
	$d_{10} = 0.02$	$n_6 = 1.6583$	$\nu_6 = 57.4$
$r_{11} = -0.329$			

[a] $f = 1.0$; $F = 3.5$; $2\omega = 24°$; back focal length $S = 0.69$

4. An ability to reduce gross primary aberrations by relatively simple parameter changes.
5. A knowledge of how certain parameter changes affect particular higher order aberrations and image quality in specific systems.
6. A knowledge of the need for optimum balance between primary and higher order aberrations.

A general treatment of these and other background areas is beyond the scope of this book, but excellent material for background purposes can be found in Cox [1], Conrady [3], Smith [4, 5], Kingslake [6], Welford [7], and Feder [8–10]. Our particular purpose is to discuss, on an intermediate level, how the OTF can contribute to:

1. Specifications of image quality for optical systems.

Table 6.III Construction Data for Lens of Fig. 6.3[a, b]

Radius of Curvature	Axial Distance	Refractive Index	Abbe Number
$r_1 = 42.701$			
	$d_1 = 7.08$	$n_1 = 1.69680$	$\nu_1 = 55.5$
$r_2 = 102.239$			
	$d_2 = 2.50$		
$r_3 = -\ 421.315$			
	$d_3 = 2.00$	$n_2 = 1.59551$	$\nu_2 = 39.2$
$r_4 = 43.423$			
	$d_4 = 5.57$		
$r_5 = 93.836$			
	$d_5 = 5.11$	$n_3 = 1.74400$	$\nu_3 = 44.7$
$r_6 = -93.836$			
	$d_6 = 5.57$		
$r_7 = -43.423$			
	$d_7 = 2.00$	$n_4 = 1.59551$	$\nu_4 = 39.2$
$r_8 = 421.315$			
	$d_8 = 2.50$		
$r_9 = -\ 102.239$			
	$d_9 = 7.08$	$n_5 = 1.69680$	$\nu_5 = 55.5$
$r_{10} = -42.701$			

[a] $f = 150$; $F = 4.5$; $2\omega = 56°$; $Y = 160$.

[b] $f_B = 63.80$; total length of the lens: 39.41; value of the condition (1): 0.425; f value of the condition (2): 0.037; f value of the condition (3): 5.13, 3.19.

2. Merit functions in automatic design (see Chapter 7).
3. Calculation of an indicator of image quality from design data (see Chapter 9).
4. Measurement and evaluation of a manufactured lens (see Chapter 8).

AUTOMATIC LENS DESIGN

Automatic lens design is usually based on the solution of simultaneous linear equations [6, 11]. As our previous discussions have indicated, optical systems are inherently nonlinear; so incremental changes in parameters must be small enough that only the linear range (first term in the Taylor's series) need be considered in the relation between parameter change and a particular aberration variation. If we assume that an optical system has N parameters p_j, which in mathematics are the degrees of freedom, available for adjustment and the same

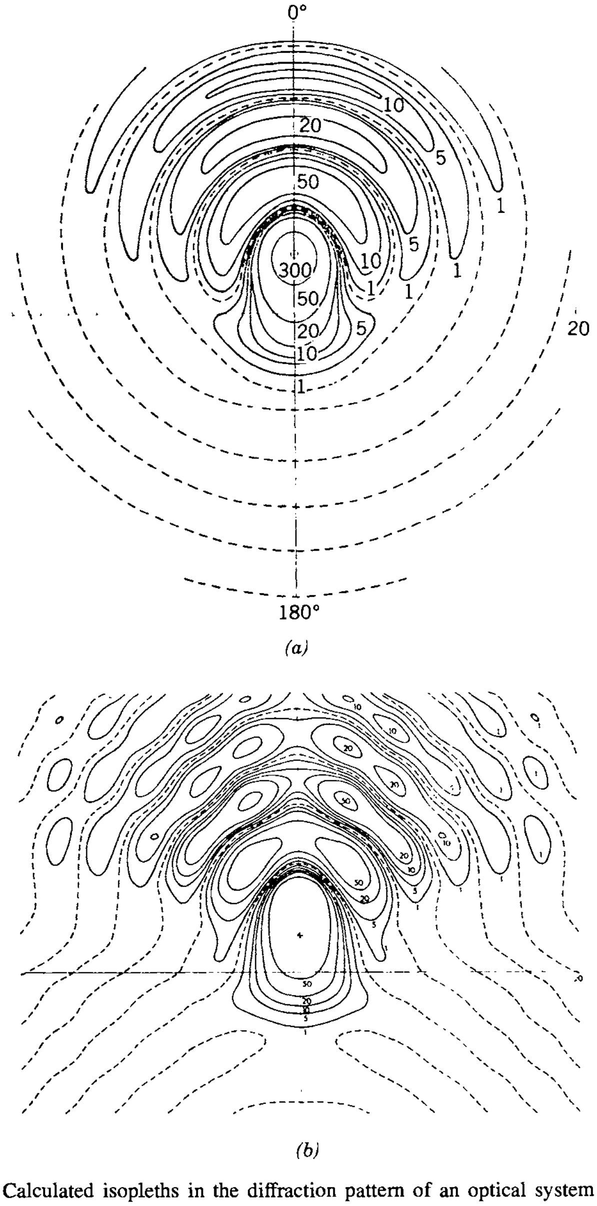


Figure 6.4. Calculated isopleths in the diffraction pattern of an optical system having a third-order coma coefficient of approximately one wavelength. The maximum of the pattern has moved upward on the image plane owing to the coma. (*a*) Circularly symmetric system. (*b*) A system with a square aperture stop [37].

number of aberrations A_i to observe, then

$$A_i = f(p_j), \tag{6-1}$$

where

$$j = 1, 2, 3, \ldots, N, \quad \text{and} \quad i = 1, 2, 3, \ldots, N.$$

Corresponding values in the functional relation of Eq. (6-1) are usually calculated by ray-tracing methods. For small changes Δp_j in the parameters, one can write the linear system of equations for the resulting changes ΔA_i in the aberrations:

$$\Delta A_i = \sum_j^N (\partial A_i/\partial p_j)\Delta p_j. \tag{6-2}$$

To simplify notation,

$$a_{ij} = \partial A_i/\partial p_j. \tag{6-3}$$

Then if a small change is introduced in each assumed parameter in turn and the resulting change in each aberration is calculated individually, the values of the coefficients a_{ij}, totaling N^2 in all, will have been determined; and the conventional computer solution methods for simultaneous linear equations can be applied—with appropriate attention to the extent of the linear region.

In the initial specifications for the optical system, the maximum allowable aberrations are given, implicitly or explicitly, so that the designer can derive desirable values of the A_i. In general, the calculated values of A_i for the initially assumed system will differ considerably from the desired values; so simply making each difference the ΔA_i in the system of equations, Eq. (6-2), is likely to violate the linearity requirement for the equations. Instead, smaller ΔA_i, say less than 30% of the ultimate reductions, can be set up in Eq. (6-2). From the computer-calculated Δp_j, a new set of parameters p_j results; and the designer repeats the previous steps: evaluating the a_{ij}, choosing a new set of ΔA_i, computing the Δp_j, and calculating the new parameters p_j. Eventually, by successive reductions ΔA_i, the residual A_i are within design specifications and the process is complete. In automatic lens design, as much of the described iterative procedure as possible is turned over to the computer. However, certain practical problems still challenge the designer's ingenuity in each conversion of a set of specifications into a manufacturable optical system. Some of these are discussed in the following paragraphs.

Nothing in the physics of optical systems produces the same number of adjustable parameters as there are types and orders of aberrations; in fact, there are usually fewer parameters than aberrations. So we have to modify the original assumption expressed in Eq. (6-1):

$$A_i = f(p_j), \tag{6-4}$$

where

$$j = 1, 2, 3, \ldots, N, \qquad i = 1, 2, 3, \ldots, M,$$

and

$$M > N.$$

This change in assumption carries over into our expressions for systems of linear equations. From the assumed definition of ΔA_i, Eq. (6-2) can be extended to

$$A_i' = A_i + \Delta A_i = A_i + \sum_j^N a_{ij}\,\Delta p_j. \tag{6-5}$$

As indicated, A_i' are the new values of the aberrations after starting with the aberrations A_i and executing one cycle of the iterative procedure already described. After each cycle, the A_i' become the A_i for the next cycle until the final set of A_i' meet specifications and are called the *residual aberrations* or just *residuals* for the optical system.

Because $M > N$ in Eq. (6-4), there is no longer a single solution for the system of equations. For this and other reasons, designers usually prefer to define a *merit function* to measure optical performance rather than set individual maxima for the various aberrations. A merit function commonly used in mathematics and engineering is the mean square of selected errors occurring in a system. In the present context, this merit function $\overline{A}^2$ is defined as

$$\begin{aligned}\overline{A}^2 &= (1/\mathrm{M})\left[(A_1')^2 + (A_2')^2 + \cdots + (A_M')^2\right]\\ &= (1/M)\sum_{i=1}^{M}(A_i')^2. \end{aligned}\tag{6-6}$$

This kind of averaging has at least two characteristics to recommend it. Errors tend to have both positive and negative values; but in most physical systems, it is not desirable for a positive error of one kind to cancel a negative error of another kind. Squaring, of course, makes all errors positive. Squaring also exaggerates the relative magnitudes of the errors: so any correcting program will realize an appreciable reduction in $\overline{A}^2$ by reducing one or more of the relatively large errors. Whatever the functional relations between errors, squaring tends to equalize the various errors when the objective is to minimize $\overline{A}^2$.

All errors or aberrations are not equally deleterious to the task at hand. In image-forming optics, considerable research has been done to determine what influences apparent image quality. When the results are interpreted in terms of the aberration system, the designer can weight the various aberrations according to their respective degrading effects. The weighting factors w_i are introduced into Eq. (6-6) as follows:

$$\overline{A}^2 = (1/M) \sum_{i=1}^{M} w_i(A_i')^2. \tag{6-7}$$

The net effect is that a computer program minimizing this mean square average will tend to reduce the more detrimental aberrations to smaller residuals.

Computer programs to minimize $\overline{A}^2$ are generally available [1, 12–14] and are based on a least-squares procedure invented by Legendre in 1805. Because in the first few cycles of an iterative $\overline{A}^2$ minimizing procedure the swings Δp_j in parameter values may be so large as to become oscillatory, the definition in Eq. (6-7) is often modified to include a damping term:

$$\overline{A}^2 = (1/M) \sum_{i=1}^{M} w_i(A_i')^2 + d \sum_{j=1}^{N} (\Delta p_j)^2, \tag{6-8}$$

where the damping coefficient d is relatively large at the beginning of the iterations and is reduced in each succeeding cycle until the damping term almost drops out at the end. Convergence of the computer sequence to the desired residual level is often tediously slow; so considerable effort has been put on techniques to hasten the process [12–16].

SELECTED FEATURES OF DESIGN PROGRAMS

The functions $\overline{A}^2$ have been discussed at some length as an example, and an A_i was identified only as "an aberration." In fact, any flaw of the imaging process could qualify as an aberration and be used in the design procedure provided each flaw is independent, can be identified and quantified, and have accord with other flaws sufficient to form a merit function. Coefficients of the terms in a power series or the polynomial coefficients in a Zernike (or another) polynomial series, which were discussed in Chapter 4, could be the aberrations A_i. Of course, the series would have to be calculated from ray-tracing data, a relatively large number of higher order terms would have to be included to ensure more aberrations than parameters, and the higher and lower orders would have to be kept in balance correctly.

A design program described by Friedman [17] is cited for an example that uses the Zernike circle polynomials. The program, entitled ZEST, illustrates a few of the complexities that are being built into modern lens design programs. Wave fronts are first determined from optical path difference (OPD) data; then the series is "fitted" to wave fronts for a nominal lens and for a variety of manufacturing perturbations to the lens, such as radius, thickness, asphericity, tilt, and irregularities of errors. The differences between the nominal and perturbed Zernike terms are calculated. These differences are treated as perturbation coefficients enabling the synthesis of wave front and MTF with no additional ray tracing. The set of Zernike coefficients, for the circle polynomials representing the wave front, consists of 25 values. These are the primary, secondary, tertiary, and quarternary terms involving up to the tenth power of ρ as illustrated in Table 4.II. The MTF is calculated by conventional integration techniques for the two directions, radial and tangential. The program is an MTF-based approach to tolerancing; it also facilitates changing the merit function when necessary as the design progresses. As each of 16 perturbations is separately applied to each lens surface, first in the positive and then in the negative sense, ZEST computes and prints image shifts, changes in each Zernike coefficient, and changes in MTF. This analysis is performed both axially and at one specified field angle. The effect of each perturbation is computed as if it were the only perturbation applied to an otherwise nominal lens. ZEST prints a diagnostic when the size of the perturbation is likely to destroy the linearity of the resulting Zernike coefficient. An AUXILIARY program enables several perturbations to be applied simultaneously. Wave-front and MTF values are computed for the image plane location specified by the user. Coefficient and derivative information determined by ZEST is used in the AUXILIARY program along with the same wave-front and MTF equations as are in ZEST. Combinations are linearly performed at the Zernike wave-front level and then MTF is computed. ZEST-supplied MTF change values for individual perturbations are never directly combined. The derivative of each Zernike coefficient with respect to image shift is also computed and stored.

Many design programs use points of ray intersection with a chosen or specified image plane because they are more direct and convenient; information readily available is transverse displacement of the principal ray (or the pupil ray) from the axis, displacement of selected rays from the principal ray intersection point, spot diagrams, and ray densities. Coefficients of the power series may be calculated for the information they provide, but the departure of each ray from its desired position is used more often for an aberration A_i than the coefficients.

Optical system-design programs have become almost unbelievably versatile. Such things as ray tracing, optimization, finding the pupil function from ray path lengths, and calculation of the OTF have become routine, at least as options for the designer. Computers now accomplish many tasks that once seemed the private domain of the optical designer. They do minor but significant design changes such as changing a single element to a doublet or triplet, shifting a stop position, eliminating weak elements, and even deliberately introducing tilts or decentricities. The designer is left with the difficult theoretical analyses involved in the design.

Computers are being used in the design of complex systems, for example:

1. Nonrotationally symmetric systems.
2. Systems using tilted and decentered elements.
3. Heads-up displays.
4. Multistage designs, as in the forward-looking infrared systems, that call for off-axis and oddly shaped elements.
5. Lens design using gradient index materials and plastics.

Desk-top calculators and microcomputers are being used more and more; when used by the optical designer, either type is normally dedicated to optical design, and many are dedicated to the design of specialized systems and programs. These small computers have a number of advantages, not the least of which is their constant availability and their very short turnaround time for program runs.

An optical-design computer and associated software will now usually have a graphics option. A few things routinely plotted for study by the designer are:

1. The MTF and MTFs at different wavelengths.
2. A map showing the wave-front shape in the exit pupil.
3. A map showing the difference between the wave front and the reference sphere.
4. A knife-edge trace.
5. A spot diagram and spot diagrams in different colors.
6. Through-focus spot diagrams.
7. Through-focus MTFs.
8. Geometrical and diffraction MTFs.
9. Transverse ray fans.
10. Point spread functions.

Graphical outputs are especially useful to the designer for study by providing a link between what can be observed and the desired optimum of a merit function. When presented with all of this information, in graphical form, it is not difficult to improve one's understanding of the relation between transverse ray aberrations, or wave-front aberrations, and the OTF. A plot of the MTF or a map of the wave-front shape should give the designer a more comfortable feel for the relation between wave-optics functions and wave-optics specifications of image quality.

It should be emphasized that graphical representations are not to be used as design criteria, but as guides to understanding. The spot diagram is a convenient representation of the geometrical point spread function, and also, from it the geometrical MTF can be calculated.

Two additional significant features are:

1. ''Global optimization,'' meaning that the program can pass through a local image-quality optimum to explore outside the neighborhood of the first optimum for a second, more favorable optimum; that is, the lowest minimum of the merit function can be found even though it is not the nearest minimum.

2. Many programs can choose a point on a glass chart where the type of glass nearest the point is the most favorable material to use for a particular element. A glass chart, or glass map, is a plot of Abbe number ν versus n_{D} where $\nu = (n_{\mathrm{D}} - 1)/(n_{\mathrm{F}} - n_{\mathrm{C}})$ and n_{D} is the refractive index near the mean index in the wavelength transparency band of most optical glasses. Values n_{F} and n_{C} are the indices near the edges of the wavelength band. Every common glass type is located on the glass chart. By this means the index of refraction and dispersion of the material are chosen by the computer program.

In addition to the references already cited, two timely and comprehensive sources of lens design information are the two *Proceedings of the SPIE* which are cited as Refs. 39 and 40 of this chapter.

MANUFACTURING TOLERANCES

The best design in the world is of no value if a shop or a factory cannot use the design to make a practical optical system. A certain practical precision must apply to every parameter that has to be controlled during the manufacturing processes: dimensions of the elements as they are being made and positions (including orientations as tilt and decentering) of the elements as they are assembled. As in the manufacture of other kinds of products, *tolerances* tell how much, plus or minus, a numerical value can differ from the design value and yet not compromise the overall performance of the optical system. Tolerances must also be within the capability of the shop to measure and control; in fact, a design is of little value if tolerances are so tight that the cost of scrapped parts becomes excessive. The process of determining tolerances is referred to as *tolerancing*.

Many design programs include tolerancing as an option (see [10, 12–14, 17]). The main feature of the program described by Friedman (already cited as Ref. 17) is tolerancing. In difficult designs the program can be stopped deliberately to allow the designer time to study the current status of the design and the tolerances that may be involved.

To set the tolerances, designers work with two sets of relations. The first set gives the sensitivity of the system performance to departures from design values, and the second set gives the manufacturing cost versus the allowed tolerances [18, 19]. Designers may make calculations with the extremes in the first set for "worst case" performance, or they may choose a statistical approach, often intuitive, and assume that all parameters do not reach tolerance extremes simultaneously. Before a large production run, a few trial systems are usually assembled and tested, and the design is fine-tuned to accommodate the realities of the shop. Thereafter, the production systems are tested according to contract or quality control requirements, which often are abbreviated procedures to assure compliance with minimum system performance specifications. These are frequently expressed as certain MTF values.

Besides designing an optical system to operate satisfactorily over a shop-dictated tolerance range for each parameter, the designer must also assure operation over certain tolerance ranges imposed by environmental and operational conditions. For instance, slight temperature changes can distort optical surfaces enough, especially in large telescopes, to introduce significant aberrations. And so, in addition to the usual optical considerations, the designer may have to work with insulating and other temperature-control techniques to complete an adequate system design. In cameras of large numerical aperture, as another example, practical methods of spherical aberration correction usually leave the inner zones undercorrected and the outer zones overcorrected. The application is further complicated by giving the user the option of aperture size over a wide range. Here the designer must exercise tolerance judgment so that the camera functions acceptably in all modes of operation. In a popular camera having zoom lenses designed to operate over a wide range of focal lengths as well as of apertures, the tolerance problem is compounded by having to keep residuals within bounds while various elements are moved smoothly along the optic axis.

ASSESSMENT OF IMAGE QUALITY

When the optical system has been designed and made, a question will be asked, "How well does the system perform the imaging task in its intended use?" If the system performs well, then a question could still be asked, "How should the degree of success be graded?" The designer, the manufacturer, the evaluater, and the user must all be satisfied with the same answers. When we ponder these questions, we realize that we are now in the business of evaluating the image produced by the system, that is, the *assessment of image quality*, because the degree of design success will be rated by quantified information extracted from the actual image.

The "intended use" of a system could refer to many and varied uses; that fact is perhaps the main obstacle to formulating a satisfactory set of criteria for assessment. The qualities required of an image will differ, or, at least, have different emphasis, in different applications; and a change in aberration-balancing that improves an optical system for one purpose might possibly make it worse for another. It seems hardly reasonable to expect a unique measure of image quality to be defined. Perhaps the most general measure of image quality is how effectively the image in some way makes information accessible about the object.

A certain specified set of OTFs of an optical system is increasingly being employed for the required criteria. A hypothetical problem can be used to demonstrate how an OTF might reveal the quality of a system and also to illustrate the limitations inherent within a system, whatever its intended use.

If we consider the image in terms of spatial frequency, the finest system that could be produced would have a relative modulation $M(s)$ of unity as calculated by Eq. (5-54):

$$M(s) = T(s)/T(s;\,0), \tag{6-9}$$

at every point of the image plane. Such an image would be produced by a diffraction-limited system, and its MTF $= T(s;\,0)$, given by Eq. (5-67), is tabulated in Table 5.I and plotted in Fig. 5.5. The real-space frequencies ω are given by Eqs. (3-52) and (3-53):

$$\omega = s(\rho/\lambda)(n \sin \alpha). \tag{6-10}$$

Then by choosing, arbitrarily, a minimum value of one-half for the MTF at a certain frequency s_1, or the corresponding ω_1, so that $T(s_1) = T(s_1;\,0) = 0.5$ we can find the contrast in the image as given by Eq. (2-16) to be one-half the contrast in the object. In fact, this is the meaning of MTF.

A value of $T(s_1;\,\mathit{0}) = 0.5$ occurs almost at 0.8 on the s scale. But since cutoff in terms of s occurs at $2s$, $T(s_1;\,0) = 0.5$ occurs at 0.4 ω_c, where ω_c is the real-space cutoff frequency. An optical system designer could hardly do better than this because it is the diffraction-limited case.

When the cutoff frequency, for example, is 400 lines/mm, then the MTF for a perfect system at 160 lines/mm is very close to 0.5. A "good eye" could begin to notice the degradation of contrast, especially if object and image can be observed side by side, and fine detail in the image begins to be lost. Some designers routinely ignore the performance above $0.2\omega_c$ until the system in design appears to be close to the desired system.

When all parties concerned agree that a certain relative modulation at a given spatial frequency ω_1 must be achieved, the designer has two ways to accomplish the goal: work on the design until the relative modulation at the chosen frequency is unity, or very nearly so; or enlarge the numerical aperture ($n \sin \alpha$). Enlarging the aperture raises the cutoff frequency so that ω_1 is a smaller fraction of ω_c. But enlarging the aperture also increases the design difficulties.

Any criterion for image quality should fulfill the following requirements:

1. For the optical designer, the criterion should be so expressed quantitatively that the system parameters can be adjusted toward an optimum design.
2. For the evaluator of the manufactured system, the criterion should be so expressed that the quality of the system can be rated from optical measurements.
3. For the evaluator and the user, the criterion should allow a practical, unambiguous specification of the quality to be expressed in its terms.
4. The nature of the criterion should allow a statement of quality to be expressible numerically on a continuous scale from bad to excellent. Just a "go" or "no go" discrimination does not suffice.
5. The same quantitative rating of quality for an optical system in terms of the criterion should be determinable by calculating both from the design data and from experimental measurements on the manufactured system.
6. Quality ratings should be independent of the kind of object being imaged and independent of the degree of coherence of the light illuminating the object.
7. The limits of expressing quality in terms of the criterion should be recognizable.

The OTF is increasingly recognized as an excellent criterion for rating image quality. A comparison of its properties and the requirement list above indicates that the OTF has at least the potential for satisfying each of the qualifications.

Our discussions of optical systems have been limited to systems starting with the object and ending with the image on a plane. Actually, such optical systems are usually followed by receiving systems that process the image configuration before the information is utilized. Examples of system extensions are the eye and its associated nervous system, a photographic film and the processing steps that follow, and a television camera tube followed by its electronics and presentation device. A great deal of effort has been expended in trying to incorporate the post-image-plane apparatus into the system to be evaluated. Except for a discussion in this chapter of some of the problems encountered beyond the image plane, we confine our efforts to just the system from object to image plane.

When images are viewed—by observing the image directly, by examining a photograph, or by seeing a television display—the human visual system contributes its properties to the overall image-transfer process. The problem thus becomes psychophysical rather than physical and requires a wholly new approach to quality evaluation. The psychophysical aspects include the effects of experience and can include aesthetic and other subjectives factors, which are far more difficult to quantize than physical measurements. The armed forces have studied at length such visual tasks as the detection, recognition, and identification of military targets and have tried to correlate the results with the outcome of bar chart tests or with some property of the frequency response function [20, 21].

The following section reviews some aspects of human perception of images and illustrates some problems that occur in the selection of criteria.

RESOLVING POWER VERSUS ACUTANCE

Extensive research by engineers and scientists at the Eastman Kodak Company and other organizations has been concerned with evaluation of the images produced by optical systems [22–34]. The Kodak investigation of *acutance* (defined later) in particular has dispelled some commonly held misconceptions of what constitutes quality in a photograph [27, 32]. Simply using some number loosely defined as *resolution* to measure quality is demonstrated to be futile.

The fundamental problem in evaluating the quality of a photograph is to define a subjective property, commonly referred to as *definition*, that measures the observable clarity by which detail is reproduced. So far no single quantity has been found to do this satisfactorily.

In the Kodak experiments, a number of photographs were made under controlled conditions, some having poor resolution but good definition and others having good resolution but poor definition. Typical results are illustrated in Figs. 6.5 and 6.6. The photograph in Fig. 6.5 was printed in such a way as to give good resolving power; from the same negative, the photograph in Fig. 6.6 was printed so that the specially defined property of *acutance* would be high. When the original photographs are compared, Fig. 6.5 shows detailed structure in the shrubs and tiled roof; so the resolution is good. But observers generally agree that the other picture, Fig. 6.6, has better *definition*. (The differences

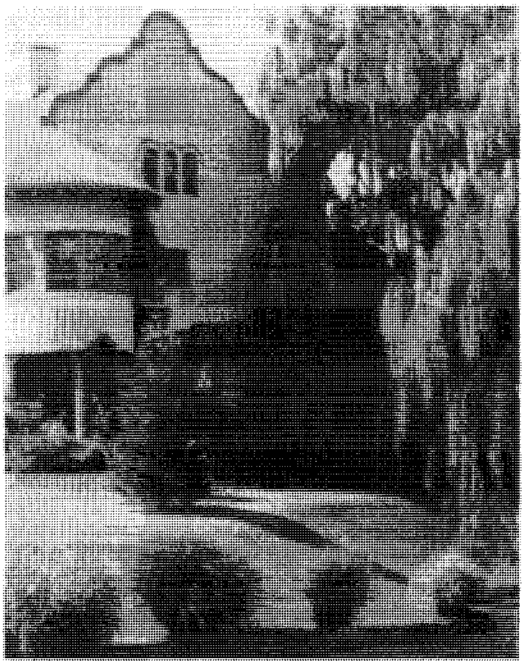


Figure 6.5. A photograph exposed for high resolution [27, 32]. (Reprinted by permission of SPSE, The Society for Imaging Science and Technology, sole copyright owners of *Photographic Science and Engineering*. Permission to reprint was also granted by the authors of Ref. 27 and by the authors and publisher of Ref. 32.).

between the photographs are admittedly difficult to see in the halftone reproductions.)

Wolf and his associates [33, 34] discovered poor correlation between resolving power and definition. In seeking the features that distinguished the high-definition pictures, they concluded that the manner in which edges are reproduced had a lot to do with the property of definition. This points to the OTF as a criterion of definition because a spatial frequency spectrum is directly derivable from an edge trace. Along with the specially shaped edge trace, of course, will always be the requirement of reasonably high resolution—particularly when the picture includes repetitive detail in the nature of a bar chart.

Acutance, which is a numerical property of an edge trace, can be defined by reference to Fig. 6.7, where $\log(1/t)$, the logarithm of the inverse of film transmittance, is plotted as a function of x, the distance along the film surface perpendicular to the indicated knife edge. The choice of ordinate variable re-

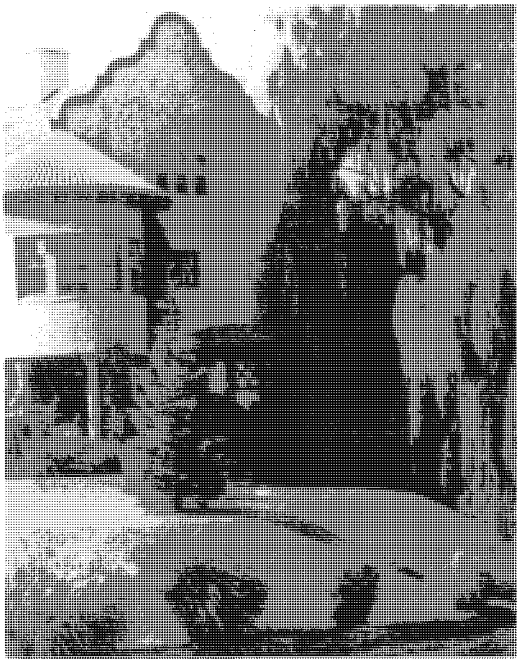


Figure 6.6. A photograph exposed for high acutance [27, 32]. (Reprinted by permission of SPSE, The Society for Imaging Science and Technology, sole copyright owners of *Photographic Science and Engineering*. Permission to reprint was also granted by the authors of Ref. 27 and by the authors and publisher of Ref. 32.).

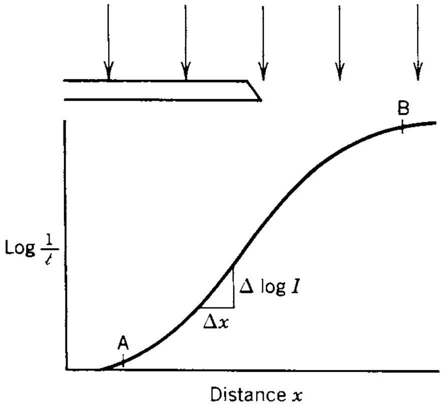


Figure 6.7. Edge trace of the image of a knife edge showing analysis for computing acutance.

sults from the relation between transmittance and film density D:

$$D = \log(1/t) = -\log t. \tag{6-11}$$

The toe and knee points A and B on the curve are at some predetermined low value of the slope or gradient. The interval between A and B is divided into N equal increments of x, and the mean square gradient of density is determined as follows:

$$\begin{aligned}\overline{G_x^2} &= (1/N)\sum_N (\Delta D/\Delta x)^2 \\ &= (1/N)\sum_N \left\{[\Delta \log(1/t)]/\Delta x\right\}^2 \\ &= (1/N)\sum_N [(\Delta \log t)/\Delta x]^2\end{aligned} \tag{6-12}$$

Acutance is then defined:

$$\begin{aligned}\text{acutance} &= \overline{G_x^2}/(D_\mathrm{B} - D_\mathrm{A}) \\ &= \overline{G_x^2}/[(\log 1/t)_\mathrm{B} - (\log 1/t)_\mathrm{A}] \\ &= \overline{G_x^2}/[(\log t)_\mathrm{A} - (\log t)_\mathrm{B}].\end{aligned} \tag{6-13}$$

The differences in the edge traces for the lens–film systems producing Figs. 6.5 and 6.6 correspond to the differences in their line spread functions shown in Fig. 6.8. Calculated from these spread functions are the MTF curves of Fig. 6.9. Curve A indicates high resolution; but curve B, which is for the system of higher acutance though poorer resolution, was judged as having the better definition of Fig. 6.6 over Fig. 6.5.

Comparison of the characteristics of the two systems that produced the photographs of Figs. 6.5 and 6.6 gives a clue as to what might serve as a merit function for evaluation or automatic design. The sharp peak of the A spread function is associated with the higher MTF values, meaning better contrast, at high spatial frequencies, which is interpreted as higher resolution. On the other hand, the higher acutance, and thus better definition, in Fig. 6.6 is associated with the higher MTF values (though the margin is only about 15%) below the crossover of the two curves. The reason for the significance of the midcurve MTF values may be understood from a hypothetical comparison of an MTF curve and a curve of an observer's perception of detail under specific test conditions of lighting, distance, and so on. In Fig. 6.10, curve a is a plot of the assumed MTF values for a system, and curve b is the significant part of the

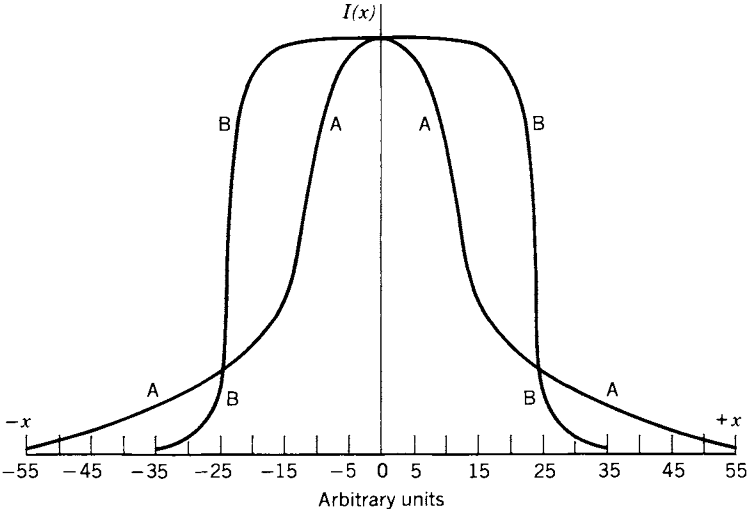


Figure 6.8. Line spread functions corresponding to Figs. 6.5 (A) and 6.6 (B).

observer's threshold curve, below which the spatial frequency variations in contrast cannot be seen. The shaded area is the margin by which the MTF curve exceeds the threshold over the frequency interval from ω_1 to ω_2. Any superiority of one MTF curve over another in the region below the threshold curve has no significance. Because MTF curves usually decline in value with increasing fre-

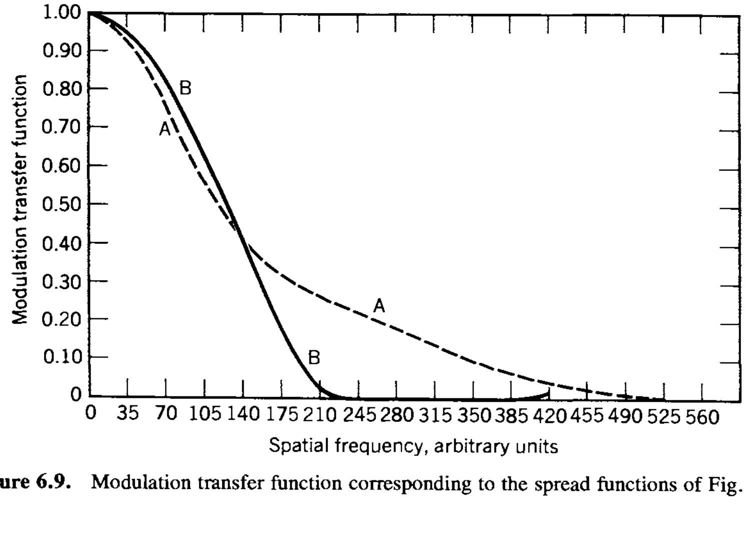


Figure 6.9. Modulation transfer function corresponding to the spread functions of Fig. 6.8.

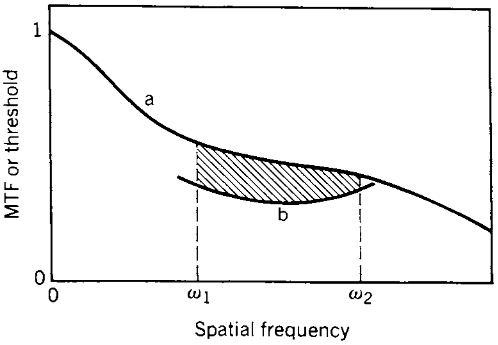


Figure 6.10. Forming a merit function by using the area (shaded) between MTF curve a and eye modulation threshold b.

quency (increasing "fineness of detail"), the region of no consequence is likely to be the upper end of the spectrum. Hence, to rate a system for definition, one must first determine where the MTF and threshold curves are likely to cross and then, somewhat arbitrarily, select a band of frequencies, as in Fig. 6.10, just under the crossover and evaluate, as a merit function, the area bounded by the two curves and the frequencies ω_1 and ω_2.

THE PHASE TRANSFER FUNCTION

In Chapter 5, the OTF, $\hat{O}(s_S, s_T)$, is written

$$\hat{O}(s_S, s_T) = T(s_S, s_T) \exp[i\phi(s_S, s_T)]. \tag{6-14}$$

where $T(s_S, s_T)$ is identified as the MTF, and $\phi(s_S, s_T)$ is the PTF or phase transfer function in terms of reduced spatial frequencies. When only a single frequency s is involved, these expressions become, respectively, $\hat{O}(s)$, $T(s)$, and $\phi(s)$ for the OTF, MTF, and PTF. In discussions of the OTF in the optical literature, emphasis is on the nature of the MTF; very little is said about the PTF. It is our purpose here to suggest why this is so and to indicate which of the OTF properties are contributed by the PTF [28, 35, 36].

The mathematical form of Eq. (6-14) indicates that the MTF is the absolute value of the OTF, and PTF gives the phase shift in radians of the OTF. In terms of the image, the MTF describes the contrast in the object. The PTF tells how much the detail at each spatial frequency is shifted in position on the image plane relative to that detail on the object plane. The numerical value of $\phi(s)$ is in terms of the spatial wavelength at each frequency. For example, if the frequency s_3 is three times s_1 and, thus, the wavelength of s_1 is three times the wavelength of s_3, a value of one radian for $\phi(s_3)$ would indicate a shift on the image plane one-third as far as one radian for $\phi(s_1)$. This means that if a combination of different frequency details on the object is to be shifted so that details maintain their relative positions to each other, their PTF values must be proportional to their frequencies; or, in other words, the $\phi(s)$ curve must be linear and must pass through the origin.

In a nonsinusoidal extended object, the variation of flux across the pattern on the object plane can be broken down, as demonstrated in Chapter 2, into sinusoidal components by Fourier methods. To preserve the exact appearance of the pattern, one requirement would be to keep the sinusoidal components in their original relative positions, which, as indicated above, requires a linear PTF curve.

An example of what a nonlinear PTF can do to a pattern can be demonstrated by starting with a bar pattern (discussed in Chapter 2) in the object plane. The variation of flux along a line perpendicular to the bars is given in Eq. (2-1), and when this variation is transformed to a Fourier series, Eq. (2-4) results. If the crenalate or square wave shape (Figs. 2.2 and 2.3) is approximated by writing just the first three terms of this series, the equation becomes

$$\begin{aligned} f(x) &= 0.5 + (4/\pi)\left[\cos(2\pi\omega x) - (1/3)\cos(6\pi\omega x)\right. \\ &\quad \left. + (1/5)\cos(10\pi\omega x)\right]. \end{aligned} \tag{6-15}$$

where the spatial frequency ω is equal to $1/(2x_1)$ in the Chapter 2 development. It is evident that the three terms of the series in Eq. (6-15) consist of the fundamental, third harmonic, and fifth harmonic of the frequency set by the bar pattern. (The coefficients of the even harmonics are all zero because of our choice of pattern.) If the fundamental frequency ω is assumed to be 150 cycles/mm and the pattern is to be transmitted through an optical system having the pass band shown by curve b of Fig. 2.23, it is evident that only the fundamental and the third harmonic will reach the image plane. Curve set A of Fig. 6.11 shows how one lobe of the square wave would look if these two components were transmitted perfectly (PTF $= 0$ for all frequencies). Curve a of the set is the fundamental, curve b is the third harmonic, and curve c is their sum, which is the plot for the image of the square wave bar pattern. (All curves in this figure have arbitrarily been normalized to the amplitude of the fundamental.) Though symmetrical, the corners of the square image are rounded, and there is a sag in the top of the lobe instead of a straight line. However, if the PTF curve were zero at the fundamental frequency and showed a phase shift of $-\pi/2$ radians at the third harmonic, the asymmetry of curve set B in Fig. 6.11

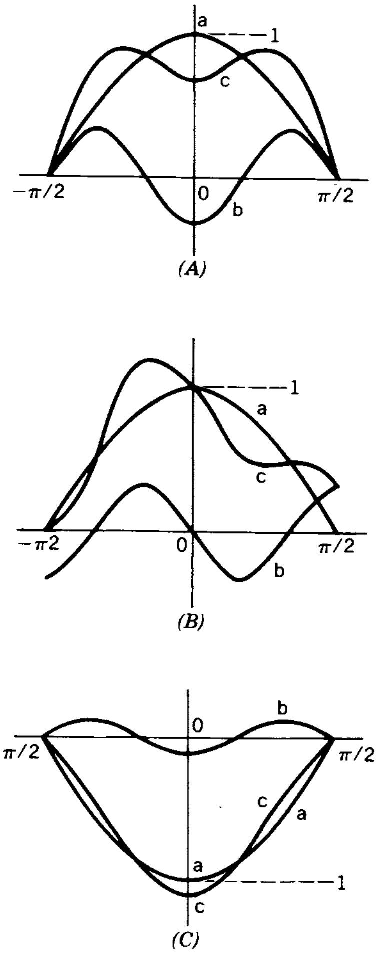


Figure 6.11. Synthesis of a fundamental plus a third harmonic for several phase relations.

results in the image, the difference between sets A and B being due entirely to the assumed difference in the PTF for the two sets. Comparison of curves c of sets A and B shows a marked deterioration of the square wave shape resulting from the assumed departure of the PTF from the ideal.

If the square wave pattern were transmitted through the defocused system represented by the MTF curve of Fig. 2.23 (PTF assumed zero at all frequencies), the negative MTF at the fundamental frequency and the small positive MTF at the third harmonic would produce the curve set C of Fig. 6.11. The lobe (curve c) is now negative and slightly more peaked than a sinusoidal lobe. The reversal of sign indicates that in the vicinity of the fundamental, the black and white bars of the bar pattern object are interchanged in the image. Though the result shown in curve set C is sometimes referred to as a "π-radian shift," it is usually regarded as an MTF rather than a PTF phenomenon and is indicated by a negative value on the MTF curve.

From our discussion of the PTF characteristic necessary to keep different frequency components of a photographic detail in correct relative position on the image plane, certain observations and conclusions may be made. The example of asymmetry in a square wave resulting from a drift of the harmonics relative to the fundamental can be extended to images of a point (point spread function) and a line (line spread function): a nonlinear PTF produces asymmetric point and line spread functions. Also, as with the square wave, a linear PTF curve that passes through the origin displaces a spread function from its correct position on the image plane, but the spread function is symmetrical. The slope of the PTF curve determines the amount of displacement. Therefore, if the slope varies as a function of the distance from the optic axis, one can expect to see pincushion or barrel distortion in the overall image. Ordinarily distortion represented by a straight-line PTF curve of modest slope is not serious except in photogrammetry, where a purpose of the photographic image is to present details in their exact positions relative to their positions in the object plane. Examples are aerial photographs taken in geographical surveys and aerial photographs to establish locations of military targets.

When point and line spread functions become asymmetrical due, for instance, to coma, the corresponding PTF will be nonlinear. Such asymmetry can coexist with photogrammetric distortion, which is the same as saying that the PTF has both linear and nonlinear components. To analyze the nonlinear component, one has to subtract the linear component from the total curve, which is equivalent to changing the coordinate origin relative to the image.

If the image coordinate axis is the same as the axis of symmetry of a symmetrical line spread function, the real and even mathematical properties of this function cause it to transform into a *real* and even spatial frequency function, that is, a function having a zero PTF at all frequencies. (See the discussion following Eq. (B-14) in Appendix B.)

As a practical matter, the PTF nonlinearity associated with an asymmetrical spread function is usually of little consequence. Most MTF curves indicate significant attenuation of the high-frequency harmonics; so the distortion caused by a shift in their phases is muted—especially in photogrammetry where the lens OTF combines with the emulsion OTF. The emulsion spread function is inherently symmetrical. When detail in the image is so fine that the fundamental is shifted by the nonlinear PTF, the combined effects of blurring by a low MTF and the reduced actual shift in millimeters due to short wavelength (see earlier discussion on the declining scale of ϕ with increasing frequency) make the PTF characteristic unimportant.

K.-J. Rosenbruch [38] thoroughly studied the role played by the PTF and has found that the PTF never exceeds $\pi/4$ radians as long as the MTF is higher than 0.2. Since many optical designers plan their systems to have sufficient contrast in the image over only the lower fourth of the spatial frequencies below cutoff, they can neglect the PTF for evaluating image quality when the OTF is part of their criteria.

As our discussion in the preceding paragraphs has suggested, no simple general statement can be made concerning the degradation of image quality caused by a large or erratic PTF. For certain special instances, the PTF may supplement the information given by the MTF, but the PTF alone usually conveys little information. Besides being more difficult to interpret than the MTF, the PTF is also more difficult to measure; so common practice is to report only the MTF part of the OTF.

REFERENCES

1. A. Cox, *A System of Optical Design.* Focal Press, London, 1964.
2. R. Barakat, Computation of the Transfer Function of an Optical System from the Design Data for Rotationally Symmetrical Aberrations, I. Theory. *J. Opt. Soc. Am.* **52,** 985 (1962).
3. A. E. Conrady, *Applied Optics and Optical Design.* Oxford Univ. Press, 1929. (Reprinted by Dover, New York, 1957.) A. E. Conrady and R. Kingslake, *Applied Optics and Optical Design*, Part 2. Dover, New York, 1960. (This is an excellent work on optical design. Unfortunately, both books are out of print. Conrady's computational procedures need to be modified and applied to the hand-held calculator.)
4. W. J. Smith, Image Formation: Geometrical and Physical Optics. In *Handbook of Optics*, W. G. Driscoll (Ed.), McGraw-Hill, New York, 1978.
5. W. J. Smith, *Modern Optical Engineering: The Design of Optical Systems.* McGraw-Hill, New York, 1966.
6. R. Kingslake, *Lens Design Fundamentals.* Academic, New York, 1978; *Optical System Design.* Academic, New York, 1983.
7. W. T. Welford, *Aberrations of the Symmetrical Optical System.* Academic, New York, 1974.
8. D. P. Feder, Automatic Lens Design Methods. *J. Opt. Soc. Am.* **47,** 902 (1957).
9. D. P. Feder, Calculation of an Optical Merit Function and Its Derivatives with Respect to the System Parameters. *J. Opt. Soc. Am.* **47,** 913 (1957).
10. D. P. Feder, Automatic Lens Design with a High Speed Computer. *J. Opt. Soc. Am.* **52,** 177 (1962).
11. H. H. Hopkins, The Use of Diffraction-Based Criteria of Image Quality in Automatic Optical Design. *Opt. Acta* **13,** 343 (1966).

12. J. Meiron, Damped Least-Squares Method for Automatic Lens Design. *J. Opt. Soc. Am.* **55,** 1105 (1965).

13. P. N. Robb, Accelerating Convergence in Automatic Lens Design. *Appl. Opt.* **15,** 4191 (1979).

14. D. P. Feder, Automatic Optical Design. *Appl. Opt.* **2,** 1209 (1963).

15. P. M. J. H. Wormell, Version 14, a Program for the Optimization of Lens Designs. *Opt. Acta* **25,** 637 (1978).

16. N. v. d. W. Lessing, Method for the Automatic Correction of the Aberrations of Optical Systems. *Appl. Opt.* **19,** 487 (1980).

17. D. Friedman, Application of Zernike Polynomial Lens Sensitivity Program ZEST to Optimizing a Lens Design. *SPIE Proc.* **237,** 99 (1980). (Please see the note following Ref. 2, Chapter 1.)

18. J. P. Starke and C. M. Wise, Modulation-Transfer-Function-Based Optical Sensitivity and Tolerancing Programs. *Appl. Opt.* **19,** 1768, (1980).

19. S. Rosin, Merit Function as an Aid in Optical Tolerancing. *Appl. Opt.* **15,** 2301 (1976).

20. L. M. Biberman (Ed.), *Perception of Displayed Information.* Plenum, New York, 1973.

21. A. van Meeteren, Prediction of Realistic Visual Tasks from Image Quality Data. *SPIE Proc.* **98,** 58 (1976). (Please see the note following Ref. 2 of Chapter 1.)

22. J. H. Altman, Image-Quality Criteria for Data Recording and Storage. *J. SMPTE* **76,** 629 (1967).

23. L. R. Baker and T. Moss, Electro-Optical Methods of Image Evaluation. *SPIE J.* **8,** 213 (1970).

24. K. G. Birch, A Survey of OTF Based Criteria Used in the Specification of Image Quality. National Physical Laboratory, Op. Met. 5, April 1969 (Teddington, Middlesex, England).

25. D. Dutton (Ed.), *Image Assessment and Specification. SPIE Proc.* **46** (1974). (Please see the note following Ref. 2 of Chapter 1.)

26. P. B. Fellgett and E. H. Linfoot, On the Assessment of Optical Images. *Trans. R. Soc. London* **247,** 369 (1955).

27. G. C. Higgins and F. H. Perrin, The Evaluation of Optical Images. *Phot. Sci. Eng.* **2,** 66 (1958).

28. C. L. Norton, Optical and Modulation Transfer Functions (Photogrammetry). *Photogramm. Eng. Remote Sensing (USA)* **43,** 613 (1977).

29. F. H. Perrin, Methods of Appraising Photographic Systems. *J. SMPTE* **69,** 151, 239 (1960).

30. K.-J. Rosenbruch, Comparison of Different Methods of Assessing the Performance of Lenses. In *Optical Instruments and Techniques*, J. Home Dickson (Ed.), Oriel Press, Stocksfield, Northumberland, U.K. (Please see Ref. 18 of Chapter 2 for more information.)

31. R. R. Shannon, Some Recent Advances in the Specification and Assessment of

Optical Images. In *Optical Instruments and Techniques*, J. Home Dickson (Ed.), Oriel Press, Stocksfield, Northumberland, UK. (Please see Ref. 18 of Chapter 2 for more information.)

32. G. C. Higgins and L. A. Jones, The Nature and Evaluation of the Sharpness of Photographic Images. *J. SMPTE* **58,** 277 (1952).

33. R. N. Wolf and F. C. Eisen, Psychrometric Evaluation of the Sharpness of Photographic Reproductions. *J. Opt. Soc. Am.* **43,** 914 (1953).

34. G. C. Higgins, R. N. Wolf, and R. L. Lamberts, Relationship between Definition and Resolving Power with Test Objects Deficient in Contrast. *J. Opt. Soc. Am.* **46,** 752 (1956).

35. R. E. Hufnagel, Significance of Phase of Optical Transfer Functions. *J. Opt. Soc. Am.* **58,** 1505 (1968).

36. R. V. Shack, On the Significance of the Phase Transfer Function. *SPIE Proc.* **46,** 39 (1974). (Please see the note following Ref. 2 of Chapter 1.)

37. R. Barakat and A. Houston, Diffraction Effects in Coma. *J. Opt. Soc. Am.* **54,** 1084 (1964).

38. K.-J. Rosenbruch, The Meaning of the Phase Transfer Function and the Modular Transfer Function in Using OTF as a Criterion for Image Quality. *Optik* **38,** 173 (1980). (In German)

39. R. E. Fischer (Ed.), *1980 International Lens Design Conference. SPIE Proc.* **237** (1980). (Please see the note following Ref. 2 of Chapter 1.)

40. W. H. Taylor and D. T. Moore (Eds.), *1985 International Lens Design Conference. SPIE Proc.* **554** (1986). (Please see the note following Ref. 2 of Chapter 1.)

7

Merit Functions and Aberration Balancing

INTRODUCTION

In Chapter 6 merit functions are discussed as a means of optimizing the design of an optical system. During the long history of optical design, perhaps the most familiar approach to utilizing merit functions has been to reduce the classical Seidel aberration coefficients to a common scale and then minimize a function of the mean square (possibly weighted) of the aberrations. This procedure is described in some detail in Chapter 6. Other merit functions based on ray optics in the vicinity of the image plane have also proved effective. Two that correlate well with each other are the size of the ray-trace spot diagram and the optical path difference (OPD) [1]. In the first, the design program aims to reduce the area on the image plane covered by the rays coming from a common point object. In the second, the objective is to make the optical path length for all rays the same from the point object to the image plane. (Conrady [2, p. 585] sets a goal of a small fraction of a wavelength for the greatest difference between optical path lengths.) A perfectly spherical wave front at the exit pupil leads geometrically to a zero OPD, both conditions indicating the absence of aberrations.

When an optical system has been corrected to a Strehl ratio of 0.8 or to the Rayleigh quarter-wavelength criterion, which are roughly equivalent, further improvement has been found more responsive to correction of wave aberrations than to reduction of spot diagrams or ray aberrations (like the transverse Seidel aberrations).

As we have suggested earlier in this book, there is a growing acceptance of the OTF, either as a graph or a table of data, as the most complete presentation of imaging information; so one could expect that superior merit functions would involve OTFs. However, the user must observe certain precautions. Like many other kinds of optical system performance data, each OTF gives information about imaging only one point in the object plane; furthermore, an OTF usually holds for only one orientation of the spatial wave pattern at that point. So, to have a comprehensive set of data about a system, OTFs must be taken for multiple points in the object plane and for at least two orientations at each point.

Also, all of these observations are functions of the wavelength of light passing through the system; so complete sets of OTFs should be taken at selected wavelengths. Once it is appreciated what constitutes a complete set of OTF data, it is usually wise to limit actual observations to just critical values. Instead of a complete OTF from zero to the cutoff frequency, the application may be adequately served by recording only a limited range of spatial frequencies; in fact, in some instances the MTF at two or three frequencies may suffice.

Designers striving to improve the performance of an optical system would be helped in their decisions by knowing the ''observer's'' sensitivity to changes of the MTF. The human observer rarely observes an optical image directly with the eyes; the image that is seen is produced by some intervening system, for example, by a ground glass plate, an exposed and processed photographic film, or a television-type picture produced by one of several possible methods. Therefore, no general sensitivity threshold, that will serve all applications, can be set for the primary optical image. However, when some property of the OTF is used as a merit function, an accepted rule of thumb is that a human observer will notice an improvement of image quality only if the MTF is increased by at least 0.1 within the range of spatial frequencies common to the frequency distributions of the object and the observer. Any degradation that might be caused by the intervening systems must be accounted for when a minimum for the merit function is specified.

Because some aberrations inevitably remain after all practical correction measures have been taken, the designer has to decide what the desired ratios should be between the various residuals. Procedures to determine these ratios differ according to the circumstances of the design; so our purpose in a later section of this chapter is to select arbitrarily an example for detailed solution to illustrate the philosophy of residual balancing rather than to set up a pat formula for general use. To treat the residual balancing problem, both the power series and the Zernike polynomial representations of the aberration function are considered.

As indicated in Chapter 6, the effects on optical system performance of specific parameter changes are useful in setting manufacturing tolerances. In a later section of this chapter, the effects of parameter changes on the OTF are studied in some detail. This choice of topic is consistent with our general purpose to emphasize the OTF significance in optics rather than to attempt a balanced presentation of all prevailing merit function practice and residual balancing.

A successful merit function depends in large part on a careful choice of the criterion for optical quality [3], which is based on what is desired in the application. At the outset, it must be appreciated that the position of the imaging plane is one of the factors that determine quality. Along with other parameters of the system, positioning of the plane can be conveniently included in the optimizing procedure. So also must be the location of what is regarded as the ideal image point, which in turn establishes the reference sphere for the wave front at the exit pupil. Choice of image plane position and image point location is treated in some detail in the residual balancing example employing the power series representation of the aberration function.

As in all calculations where exact integration processes give way to corresponding numerical methods appropriate for digital computers, summation expressions have to be written with short enough intervals in the independent variable(s) to attain desired accuracy.

SINGLE MTF VALUES AND CERTAIN GRAPHICAL AREAS AS CRITERIA OF PERFORMANCE

In general, when evaluating a choice of parameters, a designer of an optical system is interested in the entire MTF curve (probably for multiple object points and at various spatial frequency orientations), as well as other data. However, after manufacturing procedures are under control, acceptance testing, for instance, can be safely reduced to establishing a single point or, at most, a few points on the MTF curve.

If a merit function is set up to maximize the area under the MTF curve between spatial frequencies ω_1 and ω_2, as in Fig. 7.1, measuring the MTF value at some ω_i intermediate to ω_1 and ω_2 may be all that is needed to assure specification compliance of a manufactured system; however, confidence in establishing the MTF curve would be greater if, instead, the values at ω_1 and ω_2

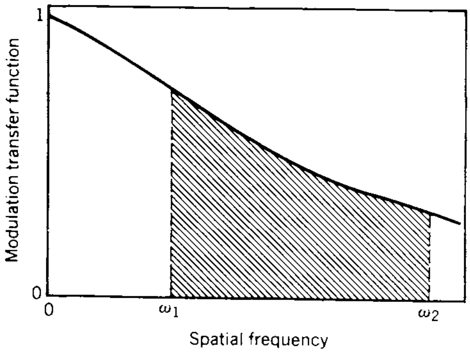


Figure 7.1. Area under the MTF curve as a merit function.

were measured, that is, two observations rather than one. This is particularly true if the manufacturing process allows marked variations in the shape of the curve in the interval of interest. Ordinarily, test measurements of the MTF are not made at low spatial frequencies because differentiation between the curves of high-performance and low-performance systems is difficult in this region. On the other hand, the high frequencies just under cutoff are also avoided because even for well-designed systems the transfer function is always small so that erratic behavior in this region is usually of little consequence. One reason for ignoring the MTF characteristics at high frequencies is indicated, for example, in the photogrammetry curves of Fig. 7.2. (See also the discussion of Fig. 6.10 in the previous chapter.) Here Fig. 7.1 has been modified by adding an emulsion curve representing a modulation detectability or minimum resolvable modulation in an aerial image. The crossover at ω_T is called the *lens/film resolution limit*. The characteristics of the MTF at higher frequencies have no significance. The merit function for the photogrammetry system could be the area between the two curves either down to a selected frequency ω_1, as in Fig. 7.1, or all the way down to zero. The enclosed area is called the *modulation transfer function area criterion*, MTFA. Since the film threshold curve of Fig. 7.2 is fixed during the optimization of the optical design, maximizing the area between ω_1 and ω_2 (where $\omega_2 = \omega_T$) attains the same results as maximizing the simpler cross-hatched area in Fig. 7.1. The principal purpose of the threshold curve is to establish ω_T. Still another criterion used in connection with the conditions represented by Fig. 7.2 is the length *AB*, the height of the MTF curve over the film threshold at the chosen frequency ω_1.

How the MTF at spatial frequency ω_1 relates to the ratio of output to input signals of a scanning optical system gives a laboratory significance to this value.

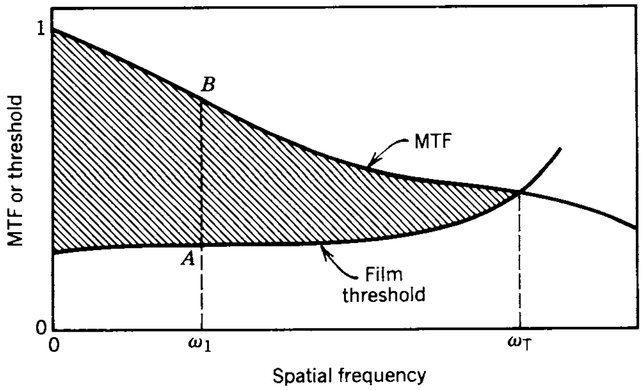


Figure 7.2. Area between the MTF and emulsion curves for defining merit functions.

The ideal object for such a test, of course, is a ripple pattern with a sinusoidal distribution of radiant power along the line that is scanned by the system. However, a review of Fourier analysis principles indicates that a pattern less precisely described as having a structure of alternating high and low flux-density regions serves almost as well for a test object. The predominant frequency ω_1 in such a structure has a half period equal to the average length of the high and low intervals in the direction of scan. If the object flux-density ratio between high and low regions, as measured by the scanning detector, is R_{ob}, the image ratio R_{im} is

$$\begin{aligned} R_{\text{im}} &= [T(s_1)]\, R_{\text{ob}} \\ &= [M(s_1)]\, [T(s_1;\, 0)]\, R_{\text{ob}}, \end{aligned} \tag{7-1}$$

where s_1 is the reduced spatial frequency (sometimes referred to as the *normalized* frequency) corresponding to ω_1 (see Eqs. (3-52) and (3-53)). As indicated by the two expressions in Eq. (7-1), the MTF value at s_1, $T(s_1)$, can also be given in terms of the *relative modulation* $M(s_1)$ (see Eq. (5-54)). The expression $T(s;\, 0)$, as in Chapter 5, stands for the "perfect" MTF whose values are given in Table 5.I.

A minimum relative modulation at an arbitrarily selected spatial frequency of ω_1 has been frequently suggested as a merit function. This application is closely equivalent to using resolving power as a merit function.

In electronic equipment, where performance versus a time frequency rather than a spatial frequency is considered, the passband or the bandwidth is an often used merit function. The two limits of the band are typically defined as the frequencies where the performance value has fallen to a given fraction of the intermediate "flat" value. These limits are applied, of course, so that the band includes only the frequency region where significant (arbitrarily defined) performance prevails. However, the typical MTF curve in optics differs markedly from the corresponding performance curve in electronics. Instead of a mesalike shape, the MTF characteristic starts at a high value at zero frequency and generally declines until a cutoff frequency is reached. To specify the bandwidth as the total span from zero to cutoff is often regarded as excessive because the low performance values near the cutoff frequency are insignificant. High values can be emphasized here, as we have already indicated in other contexts, by squaring all values and then making comparisons. Following this weighting procedure, optical workers define an *equivalent passband* N_{e} as

$$N_{\text{e}} = \int_0^{+\infty} [T(\omega)]^2\, d\omega, \tag{7-2}$$

where $T(\omega)$ is the MTF as a function of the actual frequency ω. (No attempt is usually made either to apply a normalizing coefficient to the integral or to take its square root as in rms calculations.)

Besides the equivalent passband significance of Eq. (7-2), the expression is similar to Parseval's formula in Appendix B, Eq. (B-18). If $T(\omega)$ is assumed an even function,

$$2N_e = \int_{-\infty}^{+\infty} [T(\omega)]^2 \, d\omega = \int_{-\infty}^{+\infty} [f(x)]^2 \, dx, \tag{7-3}$$

where

$$T(\omega) \leftrightarrow f(x), \tag{7-4}$$

that is, the two functions are transforms of each other and both are real and even functions according to the discussion of Fourier transforms in Appendix B. A review of earlier discussions of spread functions identifies $f(x)$ as the line spread function; so the equivalent passband can be evaluated, by Eq. (7-3), from the area under the curve of squared flux-density distribution in the image of a line source, as well as from the area under the squared MTF curve.

A MERIT FUNCTION BASED ON THE LOW-FREQUENCY END OF THE MTF

In a general consideration of the various frequency sections of the MTF as indicators of optical performance, we have already dismissed the low extreme as varying too little with merit changes and the high extreme as being too erratic to tell a good story. However, under certain conditions—particularly that of starting with a fairly well-corrected system—observing the slope and related defined graphical areas at the low-frequency end looks promising for merit evaluation.

When the wave aberration, of any type, approaches 1.25 wavelengths, the MTF curve drops rapidly with increasing frequency; and slightly above $s = 0.3$ the MTF is in the vicinity of zero, as indicated by curve d in Fig. 7.3. Reference to Appendix A indicates that MTF curves of this type become erratic above the zero or minimum near $s = 0.3$; so the higher frequency region of the MTF cannot provide a merit function for further improvement of lens quality. However, comparison of the 1.25-wavelength curve with curves of lesser aberration shows that the first dip toward zero occurs at higher and higher frequencies as the aberration is reduced. To convert this wheeling behavior about

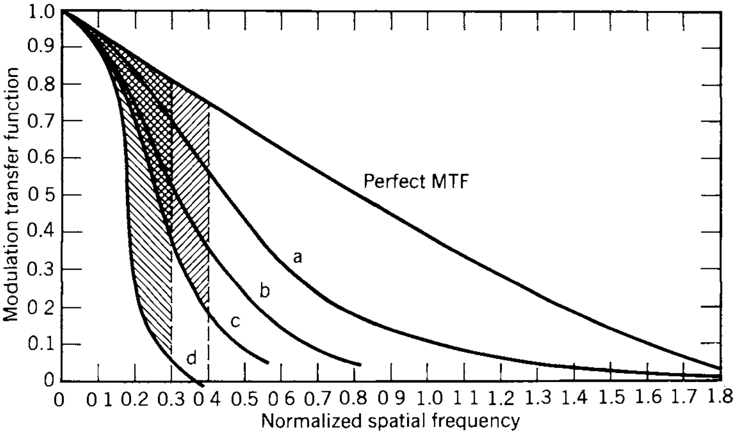


Figure 7.3. Characteristics of the MTF curve at low frequencies for defining merit functions.

the point (0, 1) into a numerical value, an area can be defined within boundaries formed by the curve under consideration (say curve *d* in Fig. 7.3), the perfect MTF, and the line $s = 0.3$. Quality is improved by minimizing this area. As this process becomes insensitive, a new area can be defined with the boundary $s = 0.4$, or any greater value below the erratic region (see curve *c*, Fig. 7.3). The minimizing procedure can be repeated with the new area, and so on until the whole area between the MTF under consideration and the perfect MTF is the value to be minimized. In the actual computational mechanics, only the successive MTF values bounding the area to be minimized are calculated. In particular, no time has to be wasted on the erratic high-frequency end of the curve.

OTHER OTF-RELATED MERIT FUNCTIONS

Granger [4] and his associates at the Eastman Kodak Company have developed an image quality merit function, or image quality reference standard, related to the OTF. Because they find that it correlates linearly with subjective quality judgments, they call the function the *subjective quality factor* (SQF). It is said to give significant evaluations over a span from the unusable to the best reasonable reproduction of a test scene. Birch [5] also discusses a number of other criteria based on the OTF.

The question of which merit function or merit criterion to choose for evaluating a given optical combination cannot be answered within the scope of this text. The choice depends partly on the physical characteristics of optical systems, devices, or materials (for instance, the eye, detector, or photographic film) used with the optical combination under test. However, the least tangible factor in the decision is based on psychometric tests. These involve subjective opinions by a number of observers as to what constitutes quality and then correlation analyses against the various merit functions to see which one best tracks the accepted subjective criteria.

MERIT EVALUATIONS BASED ON THE ABERRATION FUNCTION

As already indicated in previous chapters, the aberration function describes the aberration characteristics of an optical system in terms of wave-front distortion at the exit pupil. The OTF is derivable from the expression for the wave front (pupil function) and, therefore, from the aberration function. So, instead of basing merit functions on the OTF, it appears consistent to base the merit function on the aberration function itself. This, in fact, is close to early practice—especially if one allows the equivalence between the traditional Seidel aberrations and certain terms in the series describing the shape of the aberration function. To illustrate the relation between the aberration function and the OTF, a particular expression is assumed in Chapter 9 for the aberration function and the OTF is calculated from it.

Three general types of merit functions based on the aberration function seem to prevail in practice: (1) the mean square value of the aberration function, (2) the variance of the aberration function, and (3) the variance of the aberration difference function [3, 6–12]. These are discussed in the following sections.

MEAN SQUARE VALUE OF THE ABERRATION FUNCTION AS A MERIT FUNCTION

In the polar coordinate system used in Chapter 4 for the exit pupil, the mean square value of the aberration function may be expressed as

$$\overline{W^2} = (1/\mathcal{A}) \iint_{\mathcal{A}} \left[W(\rho, \varphi)\right]^2 \rho \, d\rho \, d\varphi, \tag{7-5}$$

where $\mathcal{A}$ is the area or the region of the wave front at the exit pupil. Before this expression can be written for a specific optical system, the position of the image plane has to be arbitrarily chosen, which is usually undesirable because optimizing this position (focusing) is often part of the procedure for reducing aberrations.

VARIANCE OF THE ABERRATION FUNCTION AS A MERIT FUNCTION

Early work on the variance of the aberration function as a merit function was done by Maréchal [13].

As a reference to calculate the variance, the mean $\overline{W}$ of $W(\rho, \varphi)$ must be established:

$$\overline{W} = (1/\mathfrak{A}) \iint_{\mathfrak{A}} W(\rho, \varphi)\rho \, d\rho \, d\varphi. \tag{7-6}$$

Then the point-by-point difference ΔW between the mean and the aberration function is

$$\Delta W = W - \overline{W}. \tag{7-7}$$

The variance $\mathcal{E}$ is defined as the mean square of the difference ΔW:

$$\begin{aligned} \mathcal{E} = \overline{(\Delta W)^2} &= (1/\mathfrak{A}) \iint_{\mathfrak{A}} (\Delta W)^2 \rho \, d\rho \, d\varphi \\ &= (1/\mathfrak{A}) \iint_{\mathfrak{A}} (W - \overline{W})^2 \rho \, d\rho \, d\varphi. \end{aligned} \tag{7-8}$$

In Fig. 7.4, which illustrates the various quantities involved in this development for an assumed wave front, the pupil ray for the particular off-axis object point that generates the wave front goes from the pupil point $\overline{E}'$ to $\overline{Q}'$; $\overline{Q}'$ is the initial image point, which is the center of the reference sphere through $\overline{E}'$ and of radius $\overline{R}$. Because of aberrations, the actual wave front does not coincide with the reference sphere; so a ray through some point P on the wave front strikes the image plane at Q' instead of $\overline{Q}'$. For purposes of illustration, the distortion of the wave front has been grossly exaggerated. In proper scale, the angle $\angle \overline{Q}'PQ'$ would be extremely small making the differences $\overline{R} - R_0$ and $W_1 - W_2$ negligible.

Expansion of the squared binomial in Eq. (7-8) leads to considerable simplification in the expression for the variance $\mathcal{E}$:

$$\mathcal{E} = (1/\mathfrak{A}) \iint_{\mathfrak{A}} (W^2 - 2W\overline{W} + \overline{W}^2)\rho \, d\rho \, d\varphi. \tag{7-9}$$

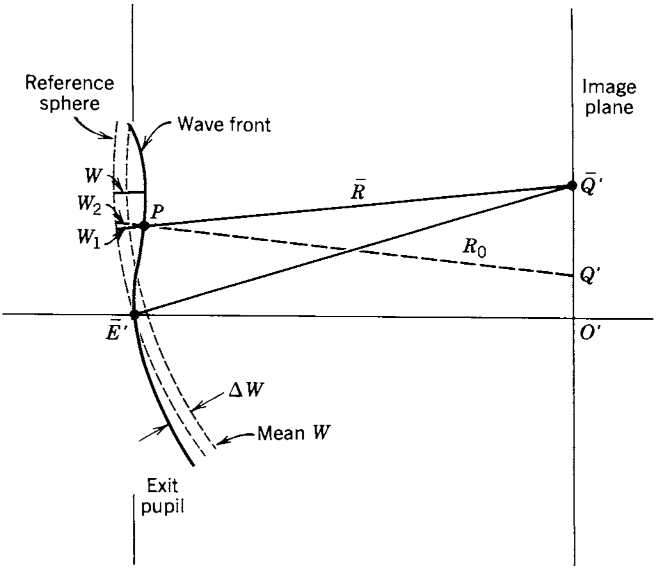


Figure 7.4. Image space geometry and symbols.

Each of the terms of the trinomial can be treated separately in the integrating procedure and the three integrals added to get the final expression for $\mathcal{E}$. Integrating the first term, as in Eq. (7-5), gives the mean square value of W. Recognizing the mean $\overline{W}$ as a constant, we find that the second term integrates to minus two times the mean value squared. Finally, the third term turns out to be just the mean value squared. Adding the three integrals yields

$$\mathcal{E} = \overline{W^2} - 2(\overline{W})^2 + (\overline{W})^2 = \overline{W^2} - (\overline{W})^2. \tag{7-10}$$

By keeping in mind the relations in Fig. 7.4, one can visualize the effect of focusing, that is, moving the image plane to the right or left along the optic axis. Moving to the left, for instance, would shorten the reference sphere radius $\bar{R}$, which would pull the ends of the broken-line arc representing the sphere toward the image plane. For the particular wave front assumed in Fig. 7.4, this change would reduce $\overline{W}$, the mean W, until at some position of the image plane, $\overline{W}$ would become zero. We assume that this could be done for any practical wave front that might be encountered. Then Eq. (7-10) becomes

$$\mathcal{E}_s = \overline{W^2}. \tag{7-11}$$

Returning to the expression for $\mathcal{E}$ before focus correction, Eq. (7-10), and noting the definitions of $\overline{W^2}$ and $\overline{W}$ in Eqs. (7-5) and (7-6), we can rewrite Eq.

(7-10) with the definitions written in terms of the approximate discrete sums:

$$\mathcal{E} = \frac{1}{N}\sum_{n=1}^{N} W_n^2 - \left[\frac{1}{N}\sum_{n=1}^{N} W_n\right]^2. \tag{7-12}$$

Maréchal, working in geometrical optics, developed a relation between the variance $\mathcal{E}$ and the Strehl ratio R_S [13, 14]:

$$R_S = \left[1 - 2(\pi/\lambda)^2\,\mathcal{E}\right]^2. \tag{7-13}$$

The Strehl ratio is discussed in Chapter 4 where $0.8 \leq R_S \leq 1$ is given as Maréchal's estimate of the useful range. However, other workers [8] are willing to extend the range to $0.7 \leq R_S \leq 1$, which in the parlance of the art includes systems at the lower end that are not quite "fairly well corrected."

Maximizing R_S as a merit function, which is the same as minimizing $\mathcal{E}$, is interesting because it treats the wave front as a whole and does not involve the corrections of terms in a series with the consequent question of how to balance the various corrections to achieve the greatest good.

VARIANCE OF THE ABERRATION DIFFERENCE FUNCTION AS A MERIT FUNCTION

Subsequent to Maréchal's relating the Strehl ratio to the variance of the aberration function, Hopkins [15] also studied the gray region of systems not quite "fairly well corrected" and came up with a relation similar in form to Eq. (7-13) but written in terms of the relative modulation and the variance of the aberration difference function.

The relative modulation is defined by Eq. (5-54):

$$M(s, \psi) = T(s, \psi)/T(s, \psi; 0), \tag{7-14}$$

which is recognized as the ratio, at frequency s and orientation ψ, of a general MTF to the MTF limited only by diffraction.

The aberration difference function has been defined by Eq. (5-96) as

$$V(x, y; s) = \frac{1}{s}\left[W\left(x + \frac{s}{2}, y\right) - W\left(x - \frac{s}{2}, y\right)\right]. \tag{7-15}$$

The transfer function $\hat{b}_0(s)$, as indicated by Eq. (5-97), can be written in terms of $V(x, y; s)$:

$$\hat{b}_0(s) = \iint_{\mathcal{A}} \exp\left[-i k n s\, V(x, y; s)\right] dx\, dy, \tag{7-16}$$

where $\mathcal{A}$ is the overlapping region in the sheared circles discussed following Eq. (5-15). If we define the mean of $V(x, y; s)$,

$$\bar{V}(s) = (1/\mathcal{A}) \iint_{\mathcal{A}} V(x, y; s)\, dx\, dy, \tag{7-17}$$

which is a constant as far as x and y are concerned, we can introduce it into the integrand of Eq. (7-16) and then compensate by writing a reciprocal of this introduced constant as a coefficient of the integral:

$$\hat{b}_0(s) = \exp\left[-ikns\,\bar{V}(s)\right] \iint_{\mathcal{A}} \exp\left\{-ikns\left[V(x, y; s) - \bar{V}(s)\right]\right\} dx\, dy. \tag{7-18}$$

In the above expressions, it is obvious for convenience that the orientation ψ and the object coordinate r, though they are among the variables on which the functions are dependent, are not always explicitly indicated in the parentheses following the function symbol.

According to Eq. (5-49),

$$T(s, \psi) = \left|\hat{b}_0(s, \psi)/\hat{b}_0(0, \psi)\right|, \tag{7-19}$$

and

$$T(s, \psi; 0) = \left|\hat{b}_0(s, \psi; 0)/\hat{b}_0(0, \psi; 0)\right|. \tag{7-20}$$

As indicated, $s = 0$ in both denominator transfer functions of Eq. (7-19) and Eq. (7-20); so in both instances Eq. (7-18) becomes

$$\hat{b}_0 = \iint_{\mathcal{A}} dx\, dy = \mathcal{A}. \tag{7-21}$$

In the discussion following Eq. (5-15), from which Eq. (7-18) is derived, the area $\mathcal{A}$ is identified as the area of the overlapping region of two unit-radius circles whose centers are separated by the value of the frequency $s = (s_S^2 + s_T^2)^{1/2}$. Therefore, when $s = 0$ as in Eq. (7-21), the area $\mathcal{A}$ is the area of a unit-radius circle or π. Then Eq. (7-14) may be written

$$M(s, \psi) = \left|\hat{b}_0(s, \psi)\right| / \left|\hat{b}_0(s, \psi; 0)\right|. \tag{7-22}$$

Since by the definition of "diffraction-limited" we mean that $W(x, y; s) = 0$ and, therefore, $V(x, y; s) = 0$ by Eq. (7-15), Eq. (7-18) becomes

$$\hat{b}_0(s, \psi; 0) = \iint_{\mathcal{A}} dx\, dy = \mathcal{A}, \tag{7-23}$$

as in Eq. (7-21) except that $\mathcal{A}$ now takes on all the values from zero to π according to the value of s. From Eqs. (7-18), (7-22), and (7-23),

$$M(s, \psi) = 1/\mathcal{A} \Big| \exp[-i\mathscr{k}ns\, \bar{V}(s)] \cdot \iint_{\mathcal{A}} \exp\{-i\mathscr{k}ns[V(x, y; s) - \bar{V}(s)]\}\, dx\, dy \Big|. \tag{7-24}$$

From the exponential identity,

$$e^{-x} = 1 - x + x^2/2! - x^3/3! + \cdots, \tag{7-25}$$

the integrand of Eq. (7-24) can be written

$$\begin{aligned} I &= 1 - i\mathscr{k}s(V - \bar{V}) + \tfrac{1}{2}(i\mathscr{k}s)^2\, (V - \bar{V})^2 \\ &= 1 - i\mathscr{k}sV + i\mathscr{k}s\bar{V} + \tfrac{1}{2}(i\mathscr{k}s)^2\, V^2 - (i\mathscr{k}s)^2\, V\bar{V} + \tfrac{1}{2}(i\mathscr{k}s)^2\, \bar{V}^2, \end{aligned} \tag{7-26}$$

where (1) the dependent variables for V and $\bar{V}$ are not indicated, (2) n is assumed unity (for air), and (3) terms beyond the third in the series have been dropped. Integrating term by term and collecting gives

$$\begin{aligned} \iint_{\mathcal{A}} I\, dx\, dy &= \mathcal{A}\left\{1 + \left[(i\mathscr{k}s)^2/(2\mathcal{A}) \iint_{\mathcal{A}} V^2\, dx\, dy\right] - \left[\frac{(i\mathscr{k}s)^2 \bar{V}^2}{2}\right]\right\} \\ &= \mathcal{A}\{1 - \tfrac{1}{2}\mathscr{k}^2 s^2 \overline{V^2} + \tfrac{1}{2}\, \mathscr{k}^2 s^2 \bar{V}^2\} \\ &= \mathcal{A}\{1 - \tfrac{1}{2}\mathscr{k}^2 s^2 (\overline{V^2} - \bar{V}^2)\}. \end{aligned} \tag{7-27}$$

The binomial in parentheses has the same form in V as the expression for $\mathcal{E}$ has in W, Eq. (7-10); so it is consistent to define the variance $\mathcal{E}_v$ of the aberration difference function as

$$\mathcal{E}_v = \overline{V^2} - \bar{V}^2. \tag{7-28}$$

Substituting from Eqs. (7-27) and (7-28) in Eq. (7-24) yields

$$M(s, \psi) = \Big|\{\exp[-i\mathscr{k}s\, \bar{V}(s)]\}\, [1 - \tfrac{1}{2}\mathscr{k}^2 s^2 \mathcal{E}_v]\Big|. \tag{7-29}$$

Since the binomial in the right square brackets consists of real quantities, it is the modulus or absolute value of the expression within the absolute value signs of Eq. (7-29). Hence,

$$M(s, \psi) = 1 - \tfrac{1}{2}k^2 s^2 \mathcal{E}_v. \tag{7-30}$$

By backtracking through the derivation of Eq. (7-29), one finds that the value of the exponent in the expression within braces is the phase transfer function of the OTF:

$$\text{PTF} = -ks\,\overline{V}(s). \tag{7-31}$$

Because of the assumptions made in the derivation of Eq. (7-29), the expressions in Eqs. (7-30) and (7-31) are only approximate; but Hopkins [15] shows that the errors are small, even for systems having the Strehl ratio somewhat less than 0.8. As a reference value, setting $M(s, \psi) = 0.8$ in Eq. (7-30) gives

$$\mathcal{E}_v = (0.1\lambda^2)/(\pi^2 s^2). \tag{7-32}$$

ABERRATION BALANCING BASED ON THE POWER SERIES EXPANSION OF THE WAVE ABERRATION FUNCTION

In Chapter 4 the wave aberration function is expanded in a power series, Eq. (4-33), in terms of the radial displacement r of the object point in the object plane, the radial coordinate ρ of the wave front at the exit pupil, and the angular coordinate φ of the wave front at the exit pupil.

For an example of aberration balancing, we assume that the aberration function consists of only the following terms:

$$\begin{aligned} W(r, \rho, \cos\varphi) = {}& {}_0C_{20}\,\rho^2 + {}_0C_{40}\,\rho^4 + {}_0C_{60}\,\rho^6 + {}_1C_{11} r\rho\cos\varphi \\ & + {}_1C_{31} r\rho^3\cos\varphi + {}_1C_{51} r\rho^5 \cos\varphi. \end{aligned} \tag{7-33}$$

In this selection, besides omitting higher order terms of spherical aberration and coma, we have assumed no astigmatism, Petzval curvature, or distortion. Further, we have retained the two focus terms ${}_0C_{20}\rho^2$ and ${}_1C_{11}r\rho\cos\varphi$, which are often dropped from the conventional power series for the aberration function.

The object for which the aberration terms are to be balanced is a fixed point source; so r becomes a constant that can be incorporated into the coefficient. Distortion and Petzval curvature have significance only if a point image position is compared with some other position, which cannot be done in this analysis since only a single point object is involved.

With r considered a constant, Eq. (7-33) becomes

$$
\begin{aligned}
W(\rho, \cos\varphi) = {} & C_{20}\rho^2 + C_{40}\rho^4 + C_{60}\rho^6 + C_{11}\rho\cos\varphi \\
& + C_{31}\rho^3\cos\varphi + C_{51}\rho^5\cos\varphi. \qquad (7\text{-}34)
\end{aligned}
$$

During the changes in parameters in the design optimization procedure, the primary coefficients C_{40} and C_{31} vary more rapidly than the secondary coefficients C_{60} and C_{51}, respectively; so initial optimization efforts tend to concentrate on the primary coefficients.

In the present task of balancing aberrations, we choose the variance $\mathcal{E}$ as the merit function to minimize in arriving at optimum balance. With reference to Eqs. (7-6), (7-9), and (7-10) and to the discussion related to these equations, the variance can be written

$$
\begin{aligned}
\mathcal{E} = {} & (1/\mathcal{A}) \iint_{\mathcal{A}} \mathrm{W}^2(\rho, \cos\varphi)\rho\, d\rho\, d\varphi \\
& - \left[(1/\mathcal{A}) \iint_{\mathcal{A}} W(\rho, \cos\varphi)\rho\, d\rho\, d\varphi \right]^2, \qquad (7\text{-}35)
\end{aligned}
$$

where $\mathcal{A}$ is the region within the unit-radius circle at the exit pupil (area $= \pi$); and $\cos\varphi$, instead of φ, is shown as the second independent variable for W because φ occurs only in the $\cos\varphi$ function in the power series representation of W. This limitation on φ is consistent with the assumed optical symmetry about the tangential plane. If Eq. (7-35) is rewritten with the properties of $\mathcal{A}$ more explicitly expressed, then

$$
\mathcal{E} = (1/\pi) \int_0^{2\pi} \int_1^1 W^2 \rho\, d\rho\, d\varphi - \left[(1/\pi) \int_0^{2\pi} \int_0^1 W \rho\, d\rho\, d\varphi \right]^2. \qquad (7\text{-}36)
$$

As the integral expressions are compared with the W^2 and W expressions, it is at once apparent that all integrand terms with the first power of $\cos\varphi$ as a factor make no contribution to either integral and can be ignored in the determination of $\mathcal{E}$:

$$
\begin{aligned}
W^2 = {} & C_{20}^2\rho^4 + C_{40}^2\rho^8 + C_{60}^2\rho^{12} + 2C_{20}C_{40}\rho^6 + 2C_{20}C_{60}\rho^8 + 2C_{40}C_{60}\rho^{10} \\
& \cancel{+ 2(C_{11}C_{20}\rho^3 + \cdots + C_{51}C_{60}\rho^{11}) \cos\varphi} \\
& + (C_{11}^2\rho^2 + C_{31}^2\rho^6 + C_{51}^2\rho^{10} + 2C_{11}C_{31}\rho^4 + 2C_{11}C_{51}\rho^6 \\
& + 2C_{31}C_{51}\rho^8) \cos^2\varphi. \qquad (7\text{-}37)
\end{aligned}
$$

$$W = C_{20}\rho^2 + C_{40}\rho^4 + C_{60}\rho^6 + (C_{11}\rho + C_{31}\rho^3 + C_{51}\rho^5)\cos\varphi. \qquad (7\text{-}38)$$

In the power series representation of W as discussed in Chapter 4, the terms with even powers of ρ are associated with spherical aberration and those with odd powers are associated with coma. (As pointed out earlier, the first term of each grouping relates to focusing.) Because the cross-products of W^2 do not contribute to $\mathcal{E}$ as indicated in Eq. (7-37),

$$\mathcal{E} = \mathcal{E}_{\text{even}} + \mathcal{E}_{\text{odd}}, \qquad (7\text{-}39)$$

$$\text{where } \mathcal{E}_{\text{even}} = (1/\pi)\int_0^{2\pi}\int_0^1 W_{\text{even}}^2\,\rho\,d\rho\,d\varphi - \left[(1/\pi)\int_0^{2\pi}\int_0^1 W_{\text{even}}\,\rho\,d\rho\,d\varphi\right]^2, \qquad (7\text{-}40)$$

$$\mathcal{E}_{\text{odd}} = (1/\pi)\int_0^{2\pi}\int_0^1 W_{\text{odd}}^2\,\rho\,d\rho\,d\varphi - \left[(1/\pi)\int_0^{2\pi}\int_0^1 W_{\text{odd}}\,\rho\,d\rho\,d\varphi\right]^2, \qquad (7\text{-}41)$$

$$W_{\text{even}} = C_{20}\rho^2 + C_{40}\rho^4 + C_{60}\rho^6, \qquad (7\text{-}42)$$

$$W_{\text{odd}} = C_{11}\rho\cos\varphi + C_{31}\rho^3\cos\varphi + C_{51}\rho^5\cos\varphi. \qquad (7\text{-}43)$$

Rather than go through all the details of evaluating the above integrals, we will show only examples of parts of the polynomials involved and then present the results obtained by proceeding similarly with all parts. If we first consider the parts of W_{even}^2 and W_{even} that contribute C_{20}^2 terms to the expression for $\mathcal{E}_{\text{even}}$, Eq. (7-40), we have

$$I_1 = (1/\pi)\int_0^{2\pi}\int_0^1 C_{20}^2\rho^5\,d\rho\,d\varphi - \left[(1/\pi)\int_0^{2\pi}\int_0^1 C_{20}\,\rho^3\,d\rho\,d\varphi\right]^2$$
$$= C_{20}^2/3 - C_{20}^2/4 = C_{20}^2/12. \qquad (7\text{-}44)$$

Collecting all such parts for $\mathcal{E}_{\text{even}}$ yields

$$\mathcal{E}_{\text{even}} = C_{20}^2/12 + 4C_{40}^2/45 + 9C_{60}^2/112 + C_{20}\,C_{40}/6 + 3C_{20}\,C_{60}/20 + C_{40}\,C_{60}/6. \qquad (7\text{-}45)$$

Because of the rapidly convergent nature of the series for W_{even}, Eq. (7-42), the most significant variable for determining a minimum for $\mathcal{E}_{\text{even}}$ is C_{20}; so a good approximation of C_{20}, called $(C_{20})_{\text{m}}$, for such a minimum can be calculated by setting $\partial\mathcal{E}_{\text{even}}/\partial C_{20} = 0$:

$$\partial\mathcal{E}_{\text{even}}/\partial C_{20} = C_{20}/6 + C_{40}/6 + 3C_{60}/20 = 0. \tag{7-46}$$

$$(C_{20})_{\text{m}} = -(C_{40} + 9C_{60}/10). \tag{7-47}$$

Because C_{20} is the coefficient of the focusing term in the expression for W, $(C_{20})_{\text{m}}$ establishes the optimum focal surface in terms of C_{40} and C_{60}. When $(C_{20})_{\text{m}}$ is substituted in the expression for $\mathcal{E}_{\text{even}}$, Eq. (7-45), $(\mathcal{E}_{\text{even}})_{\text{min}}$ in terms of C_{40} and C_{60} results. After algebraic simplification,

$$(\mathcal{E}_{\text{even}})_{\text{min}} = \left[(C_{40}/C_{60})^2 + 3(C_{40}/C_{60}) + (81/35)\right](C_{60}^2/180). \tag{7-48}$$

To continue the minimizing procedure, it is convenient to assign a symbol to the ratio of the primary to the secondary coefficient, C_{40}/C_{60}:

$$\beta_{46} = -C_{40}/C_{60} \tag{7-49}$$

By adding and subtracting the fraction 9/4 in the brackets of Eq. (7-48), a perfect binomial square is formed, and Eq. (7-48) can be written

$$(\mathcal{E}_{\text{even}})_{\text{min}} = \left[(\beta_{46} - 3/2)^2 + 9/140\right](C_{60}^2/180). \tag{7-50}$$

Returning to the expression for W_{odd}, Eq. (7-43), we can substitute in the integrals for $\mathcal{E}_{\text{odd}}$, Eq. (7-41), and repeat calculations similar to those demonstrated for W_{even} and $\mathcal{E}_{\text{even}}$:

$$\mathcal{E}_{\text{odd}} = C_{11}^2/4 + C_{31}^2/8 + C_{51}^2/12 + C_{11}C_{31}/3 + C_{11}C_{51}/4 + C_{31}C_{51}/5. \tag{7-51}$$

By finding the C_{11} that satisfies $\partial\mathcal{E}_{\text{odd}}/\partial C_{11} = 0$, we find

$$(C_{11})_{\text{m}} = -(2C_{31}/3 + C_{51}/2). \tag{7-52}$$

If we again define a β as the ratio between primary and secondary coefficients,

$$\beta_{35} = -C_{31}/C_{51}, \tag{7-53}$$

then

$$(C_{11})_{\text{m}} = (2\beta_{35}/3 - 1/2)\,C_{51}. \tag{7-54}$$

Substituting this value in Eq. (7-51) yields

$$(\mathcal{E}_{\text{odd}})_{\text{min}} = \left[(\beta_{35} - 1.2)^2 + 3/50\right] C_{51}^2/72. \tag{7-55}$$

At this point of our development, in Eqs. (7-50) and (7-55), we have expressions for the two parts of the variance $\mathcal{E}$, $\mathcal{E}_{\text{even}}$ (spherical aberration) and $\mathcal{E}_{\text{odd}}$ (coma), for the optimum (signified by "min") location of the image. Each expression is in terms of the ratio of the primary to the secondary coefficient of the respective aberration and of the secondary coefficient. Now we propose to fix each variance and get the corresponding relation between each ratio and its secondary coefficient.

Marėchal [13] has suggested the following range of values, also adopted by others, for $\mathcal{E}$:

$$0 \leq \mathcal{E} \leq \lambda^2/180, \tag{7-56}$$

which, by substitution in Eq. (7-13), corresponds to the following range for the Strehl ratio:

$$0.7927 \leq R_S \leq 1. \tag{7-57}$$

This is consistent with Marėchal's well-known choice of 0.8 for an R_S lower limit. If we substitute the upper limit of $\mathcal{E}$, Eq. (7-56), in both Eq. (7-50) and Eq. (7-55), then

$$C_{60}/\lambda = \left[(\beta_{46} - 3/2)^2 + 9/140\right]^{-1/2}, \tag{7-58}$$

$$C_{51}/\lambda = \left[(\beta_{35} - 1.2)^2 + 3/50\right]^{-1/2} (2/5)^{1/2}. \tag{7-59}$$

These two equations are plotted in Fig. 7.5. (Usually the ranges expressed in Eqs. (7-56) and (7-57) are for the total aberrations in the system, whereas in Eqs. (7-58) and (7-59) we have chosen to introduce the extreme value of $\mathcal{E}$, $\lambda^2/180$, for each of the component aberrations, spherical aberration and coma. This means that the total $\mathcal{E}$, Eq. (7-39), would be $\lambda^2/90$ and the corresponding R_S would be 0.6095. Our reason for taking this liberty is to track closely with an example in our references [3].) Each of the equations gives the relation between the secondary coefficient and the primary-to-secondary ćoefficient ratio at the maximum $\mathcal{E}$ (equivalent to the minimum R_S) for a fairly well-corrected system. Each peak in the two curves of Fig. 7.5, therefore, gives the greatest value of the secondary coefficient allowable for the somewhat arbitrarily defined "fairly well-corrected system." The peak value of C_{60}/λ is 3.944 and occurs at a β_{46} value of 1.5; the peak value of C_{51}/λ is 2.5820 at a β_{35} of 1.2.

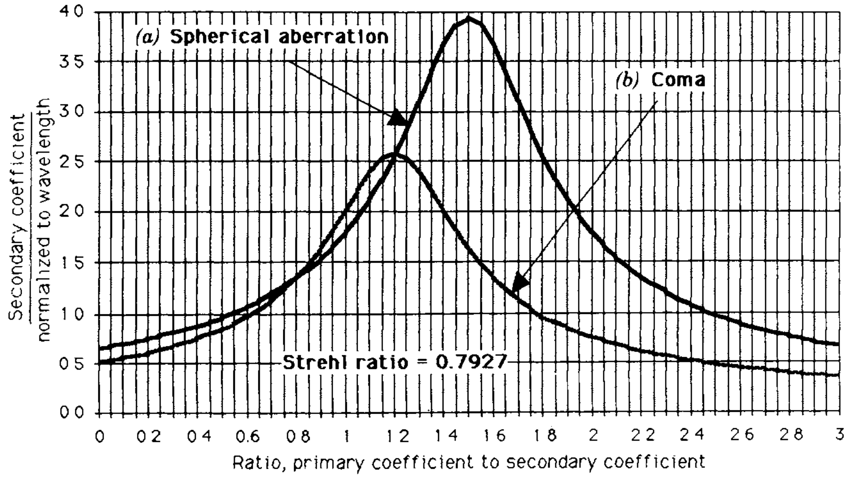


Figure 7.5. For a Strehl ratio of 0.7927 ($\mathcal{E} = \lambda^2/180$): (*a*) Variation of the secondary coefficient of spherical aberration as a function of the primary-to-secondary ratio of coefficients; (*b*) variation of the secondary coefficient of coma as a function of the primary-to-secondary ratio of coefficients [3].

The value for each β to achieve the maximum secondary coefficient for other given variances remains the same as for the particular variance value, $\lambda^2/180$, discussed above. This can be seen by considering the form of Eqs. (7-50) and (7-55). For example, solving for C_{60}^2 in Eq. (7-50) gives

$$C_{60}^2 = \frac{180\ (\mathcal{E}_{\text{even}})_{\text{min}}}{(\beta_{46} - 3/2)^2 + 9/140}, \qquad (7\text{-}60)$$

and it is at once evident by inspection that for any chosen value of $(\mathcal{E}_{\text{even}})_{\text{min}}$, C_{60}^2 (and, therefore, C_{60}) will be a maximum for $\beta_{46} = 3/2$. Similarly in Eq. (7-55), for any chosen value of $(\mathcal{E}_{\text{odd}})_{\text{min}}$, C_{51} will be a maximum for $\beta_{35} = 1.2$. For these maximizing values of the β's, Eqs. (7-50) and (7-55) become, respectively,

$$\mathcal{E}_{\text{spherical}} = (9/140)(C_{60}^2/180) = 9C_{60}^2/(140 \times 180), \qquad (7\text{-}61)$$

$$\mathcal{E}_{\text{coma}} = (3/50)(C_{51}^2/72) = 3C_{51}^2/(50 \times 72), \qquad (7\text{-}62)$$

where the $\mathcal{E}$ subscripts, *spherical* for *even* and *coma* for *odd*, more explicitly identify the aberrations involved. If the equivalent expressions in terms of the Strehl ratio instead of the variance are desired, the respective expressions for $\mathcal{E}$

can be substituted in Eq. (7-13) to get

$$(R_S)_{\text{spherical}} = [1 - 2(\pi/\lambda)^2(9C_{60}^2)/(140 \times 180)]^2$$
$$= \left[1 - \frac{2 \times 9\pi^2}{140 \times 180}(C_{60}/\lambda)^2\right]^2, \tag{7-63}$$

and

$$(R_S)_{\text{coma}} = [1 - 2(\pi/\lambda)^2(3C_{51}^2/(50 \times 72)]^2$$
$$= \left[1 - \frac{2 \times 3\pi^2}{50 \times 72}(C_{51}/\lambda)^2\right]^2. \tag{7-64}$$

These equations are plotted as the lower curves in Fig. 7.6.

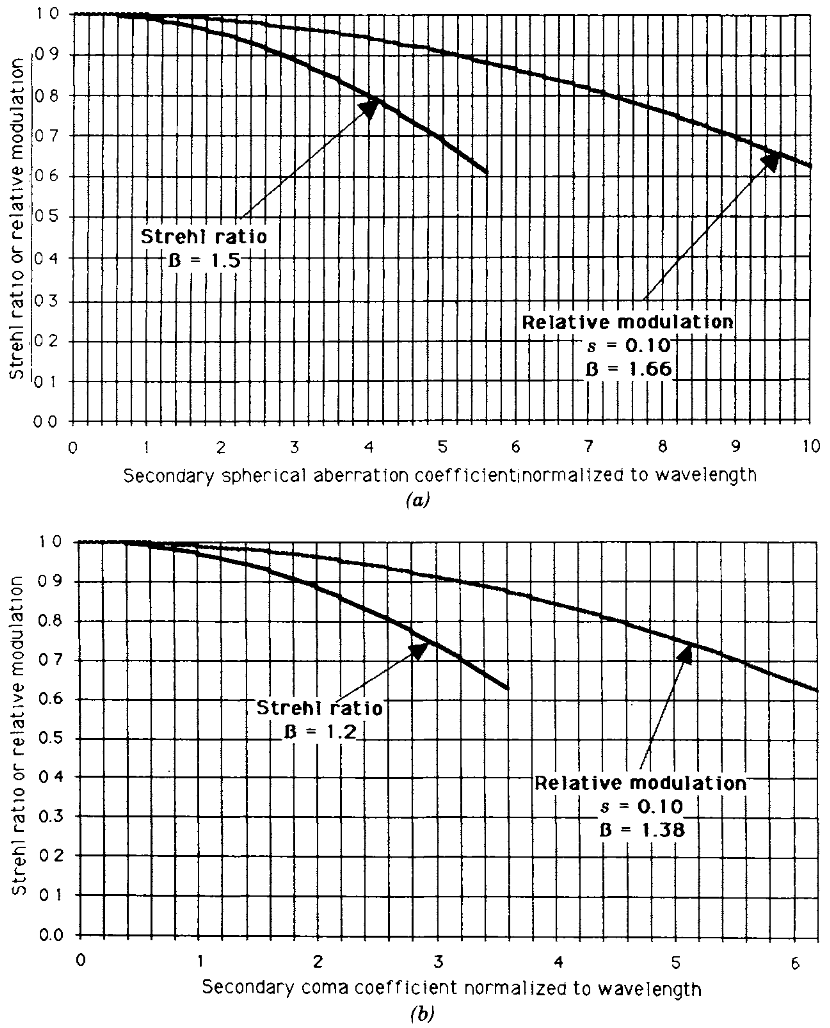
The significance of aberration *balancing* (that is, the optimum choice of β), as compared with aberration *reduction*, can be illustrated by two examples taken from the curves in Figs. 7.5 and 7.6.

All points on both curves in Fig. 7.5 are for a Strehl ratio R_S equal to 0.7927. (See the discussion preceding Eq. (7-58).) Suppose that the coefficient ratio is arbitrarily set at unity, $\beta_{46} = 1$, in Fig. 7.5*a*, that is, equal primary and secondary aberration coefficients but of opposite sign. The corresponding value for the secondary coefficient is $C_{60}/\lambda = 1.784$. If this same secondary coefficient value is combined with the optimum $\beta_{46} = 3/2$ (instead of unity), Fig. 7.6*a*, or Eq. (7-63), indicates a Strehl ratio of $R_S = 0.9556$. So by *increasing* the primary spherical aberration 50% with the same secondary coefficient, the Strehl ratio is improved from 0.7927 to 0.9556.

As a second example, we choose $\beta_{35} = 1$ in Fig. 7.5*b*, which results in $C_{51}/\lambda = 2$ for $R_S = 0.7927$. If the same value of the secondary coefficient, $C_{51}/\lambda = 2$, is located on the lower curve of Fig. 7.6*b* (or is substituted in Eq. (7-64)), meaning that β_{35} is increased to the optimum 1.2 by increasing the absolute value of the primary coefficient by 20%, R_S is improved from 0.7927 to 0.8727.

In each of the examples, despite increasing the primary aberration coefficient considerably (to reach the optimum primary-to-secondary ratio), the overall aberration is appreciably reduced, as indicated by the increased Strehl ratio.

Because of the minimum permissible value of R_S, Fig. 7.6 indicates that C_{60}/λ cannot exceed a value slightly above 4 and C_{51}/λ cannot exceed a value slightly above 3. If greater values of these coefficients (equivalent to smaller values of the Strehl ratio) are to be handled in optimum balancing, some other criterion, such as relative modulation $M(s, \psi)$, has to be considered instead of the Strehl ratio R_S. The upper curves in Fig. 7.6 show the results of pursuing a development for relative modulation corresponding to the described one for the Strehl ratio.



(a)

(b)

Figure 7.6. (*a*) Relative modulation and Strehl ratio as functions of the secondary coefficient for spherical aberration. (*b*) Relative modulation and Strehl ratio as functions of the secondary coefficient for coma [3].

As indicated by Eq. (7-14), the relative modulation is a function of both the frequency s and the orientation ψ, whereas the Strehl ratio is expressed for a combination of all frequencies and orientation in the optical system. So a meaningful balancing analysis in terms of the relative modulation would probably require a number of parallel calculations for various frequencies and orientations. Because of the shape of a typical MTF curve, significant results tend to be at the lower frequencies (small values of s).

As in the balancing development resulting in the R_S curves of Fig. 7.6, the development leading to the $M(s, \psi)$ curves starts with Eqs. (7-42) and (7-43) as expressions of the aberration functions W_{even} and W_{odd} for spherical aberration (even) and coma (odd). From these expressions, with appropriate attention to rectangular–polar coordinate transformation, the corresponding expressions for the aberration difference functions, $[V(x, y; s)]_{\text{even}}$ and $[V(x, y; s)]_{\text{odd}}$, are derived. The mean aberration difference function $\overline{V}(s)$ for each is found by Eq. (7-17) and the mean squared aberration function $\overline{V^2}(s)$ by

$$\overline{V^2}(s) = (1/\mathcal{A}) \iint_{\mathcal{A}} [V(x, y; s]^2 \, dx \, dy. \tag{7-65}$$

From the two means, Eq. (7-28) gives the variances $(\mathcal{E}_v)_{\text{even}}$ and $(\mathcal{E}_v)_{\text{odd}}$ of the aberration difference functions. From each variance, the corresponding relative modulation $M(s, \psi)$ can be found by Eq. (7-30).

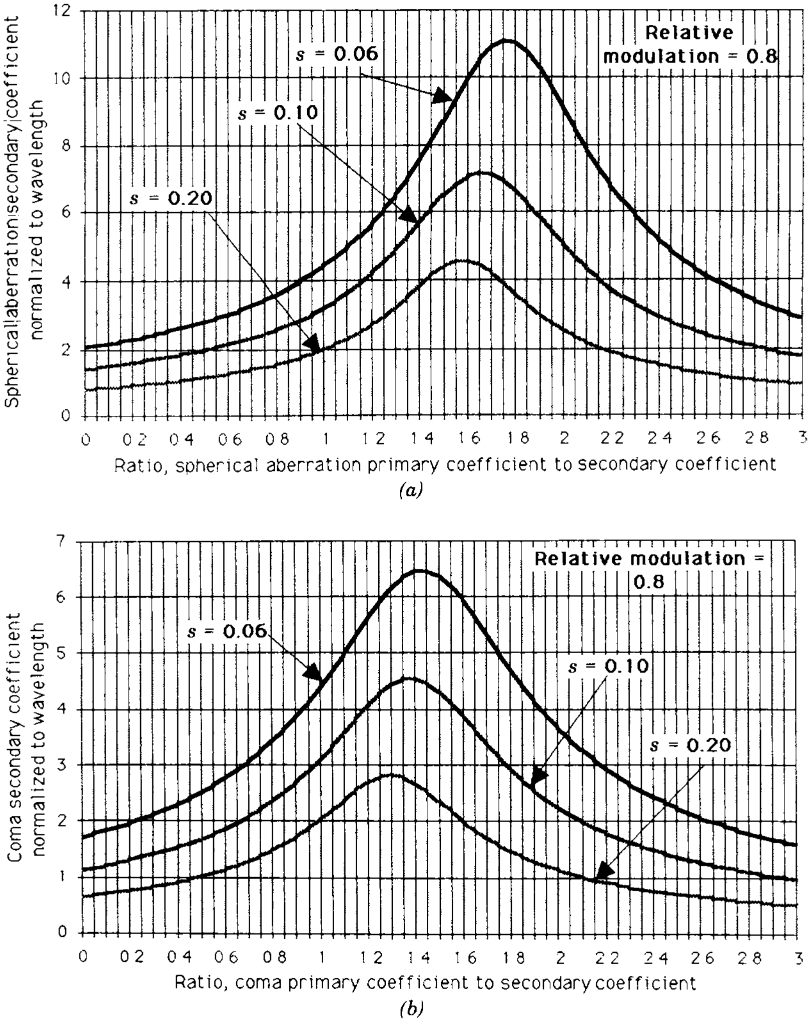
Hopkins [3] gives the variances as

$$\mathcal{E}_{\text{vs}} = A_{22}C_{20}^2 + A_{44}C_{40}^2 + A_{66}C_{60}^2 + A_{24}C_{20}C_{40} + A_{26}C_{20}C_{60} + A_{46}C_{40}C_{60}, \tag{7-66}$$

$$\mathcal{E}_{\text{vc}} = A_{33}C_{31}^2 + A_{55}C_{51}^2 + A_{35}C_{31}C_{51}, \tag{7-67}$$

where the subscript s (for *spherical*) has the same significance as the subscript *even*, and the subscript c (for *coma*) has the same significance as the subscript *odd*.

By pursuing minimizing steps as have been carried out for the Strehl ratio and by defining the coefficient ratios β_{46} and β_{35} as before, curves for C_{60}/λ and for C_{51}/λ as functions of the β's can be calculated for the arbitrary value $M(s, \psi) = 0.8$ (Fig. 7.7). Unlike the Strehl ratio development, each $M(s, \psi)$ curve is for a specific frequency s as indicated. At $s = 0.10$, the peak value of C_{60}/λ occurs at $\beta_{46} = 1.66$; the peak value of C_{51}/λ at $\beta_{35} = 1.38$. When these β values are substituted in the relative modulation equations corresponding to Eqs. (7-50) and (7-55) of the Strehl development and then $M(s, \psi)$ is substituted for the two expressions for $\mathcal{E}_v$ according to Eq. (7-30), the relations shown by the upper curves in Fig. 7.6 result. As indicated in the parentheses following the M's, the spatial frequency s for both relative modulation curves is 0.10; the angle ψ for spherical aberration is zero, and the angle for coma is $\pi/2$ radians. Obviously an infinity of curves could be drawn for all possible choices of s and ψ.



(*a*)

(*b*)

Figure 7.7. For a relative modulation of 0.8 at the indicated frequencies: (*a*) Variation of the secondary coefficient of spherical aberration as a function of the primary-to-secondary ratio of coefficients; (*b*) variation of the secondary coefficient of coma as a function of the primary-to-secondary ratio of coefficients [3].

If, instead of the β's for the maximum values of C_{60}/λ and C_{51}/λ, a certain design requires $\beta_{46} = 1.25$ and $\beta_{35} = 0.95$ at $s = 0.10$, then, for a relative modulation of 0.8 (for which all the curves of Fig. 7.7 are drawn), $C_{60} = 4.5\lambda$ and $C_{51} = 2.9\lambda$. However, the curves of Fig. 7.6 show that these values of C_{60} and C_{51} would result in relative modulation values of 0.92 and 0.91, respectively, for spherical aberration and coma if the optimum β's, 1.66 and 1.38, were applied. Again, as in the Strehl examples, an improvement is realized by *increasing* the primary aberration to attain optimum balance.

ABERRATION BALANCING WITH ZERNIKE POLYNOMIALS

In Chapter 4 the wave aberration function is expanded in two different series. The first is the traditional power series illustrated in Eq. (4-33) and in Table 4.I where the terms are sometimes referred to as the *classical aberrations*, some of which correspond to the historical Seidel aberrations. Balancing these classical aberrations to optimize the wave aberration function according to some merit function is the concern of the previous section.

The second series discussed in Chapter 4 is the one involving Zernike polynomials as indicated in Eq. (4-47). A significant difference in the two series is the way the polar coordinate angle φ occurs in the terms of each series. In the power series, $\cos^m \varphi$ is the form of the angular function; in the Zernike polynomials, $\cos m\varphi$ occurs as a factor. As suggested by Eq. (4-48), trigonometric identities relating $\cos^m \varphi$ and $\cos m\varphi$ functions allow the expression of a term in one series as a combination of terms in the other series.

In the previous section, the process of determining how large the primary aberration should be to balance the higher order terms proves rather tedious. On the other hand, when each Zernike term is converted to its equivalent classical combination of aberrations, a "built-in" property produces classical aberrations that are already balanced. Demonstrating this property is the purpose of the present section.

In Eq. (7-13) the Strehl ratio is given as the source of a binomial. When this binomial is expanded, the following expression in terms of the variance $\mathcal{E}$ results:

$$R_S = 1 - (4\pi^2/\lambda^2)\mathcal{E} + (4\pi^4/\lambda^4)\mathcal{E}^2. \qquad (7\text{-}68)$$

For R_S values close to unity, the third term on the right side can be dropped:

$$R_S \simeq 1 - (4\pi^2/\lambda^2)\mathcal{E}. \qquad (7\text{-}69)$$

(For convenience, since approximations based on restricted independent variable ranges are commonly expressed as equalities in optics, the "approximately" symbol in Eq. (7-69) will be dropped in derived relations.) When the expression for $\mathcal{E}$ in terms of the wave aberration function W, Eq. (7-10), is substituted in Eq. (7-69):

$$R_S = 1 - \left[(2\pi)^2/\lambda^2\right]\left[\overline{W^2} - (\overline{W})^2\right]. \tag{7-70}$$

The average of the squared wave aberration function $\overline{W^2}$ is defined in Eq. (7-5). Because the indicated integration is over a unit-radius circle, the expression can be written more explicitly as

$$\overline{W^2} = (1/\pi)\int_0^{2\pi}\int_0^1 W^2\rho\, d\rho\, d\varphi. \tag{7-71}$$

The average of the wave aberration function, defined in Eq. (7-6), can also be more explicitly written:

$$\overline{W} = (1/\pi)\int_0^{2\pi}\int_0^1 W\,\rho\, d\rho\, d\varphi. \tag{7-72}$$

One form of W expressed in a series of Zernike polynomials is given as Eq. (4-47). By revising the definition of A_{n0} as done by Bezdid'ko [16] and others, the series can be more compactly written:

$$W = \sum_{nm} A_{nm} R_n^m(\rho)\cos m\varphi. \tag{7-73}$$

Substitution of this series in Eq. (7-70) according to Eqs. (7-71) and (7-72) looks extremely complicated. However, because of the orthogonal property of Zernike polynomials, as discussed in Chapter 4, the cross-product terms (that is, where the factors have different m's) in the squared series integrate to zero:

$$\overline{W^2} = A_{00}^2 + (1/2)\sum_{n=m}^{\infty}\sum_{n=0}^{\infty} A_{nm}^2/(n+1), \tag{7-74}$$

$$\overline{W} = A_{00}, \tag{7-75}$$

so

$$R_S = 1 - \left[(2\pi)^2/\lambda^2\right]\sum_{nm} A_{nm}^2/(n+1). \tag{7-76}$$

Besides being surprisingly simple in form, this expression for R_S indicates that each individual Zernike aberration reduces the Strehl ratio independently of the others.

As a demonstration of the inherent balance in each Zernike aberration, the following single term will be explored:

$$W = A_{44}\, R_4^4\,(\rho)\cos 4\varphi. \tag{7-77}$$

From Eq. (7-76),

$$R_S = 1 - \left[(2\pi)^2/\lambda^2\right]A_{44}^2/5. \tag{7-78}$$

To transform the Zernike term into the equivalent combination of classical aberrations, the third identity of Eq. (4-48) is applied:

$$\cos 4\varphi = 8\cos^4\varphi - 8\cos^2\varphi + 1. \tag{7-79}$$

From Table 4.II, the expression for the Zernike radial polynomial can be found:

$$R_4^4(\rho) = \rho^4. \tag{7-80}$$

So the complete classical expression for the Zernike wave aberration function assumed in Eq. (7-77) is

$$W = 8A_{44}\rho^4\cos^4\varphi - 8A_{44}\rho^4\cos^2\varphi + A_{44}\rho^4. \tag{7-81}$$

To check whether this combination is balanced, an arbitrary third-order (Seidel) spherical aberration, ${}_0C_{40}\rho^4$, is added to W:

$$W' = W + {}_0C_{40}\rho^4 = 8A_{44}\rho^4\cos^4\varphi - 8A_{44}\rho^4\cos^2\varphi + A_{44}\rho^4 + {}_0C_{40}\rho^4. \tag{7-82}$$

The coefficient ${}_0C_{40}$ can be any nonzero real value, positive or negative. To show that the aberrations for W, Eqs. (7-77) and (7-81), are balanced, we have to demonstrate that any R'_S for W' must be less than the R_S for W, Eq. (7-78). To do this, the expression for W', Eq. (7-82), has to be written in terms of Zernike polynomials so that we can utilize the formula for R_S, Eq. (7-76). Since the first three terms on the right side of Eq. (7-82) constitute the Zernike polynomial for W in Eq. (7-77), we have only to determine the Zernike equivalent of the remaining term, ${}_0C_{40}\rho^4$. This can be done by finding what combination of the $m = 0$ Zernike radial polynomials in Table 4.II will sum up to ${}_0C_{40}\rho^4$. By inspection, it is apparent that the following, with appropriate coefficients, will do the job:

$$R_0^0 = 1, \qquad R_2^0 = 2\,\rho^2 - 1, \qquad \text{and} \quad R_4^0 = 6\,\rho^4 - 6\,\rho^2 + 1. \tag{7-83}$$

Forming the sum gives

$$R_4^0 + 3R_2^0 + 2R_0^0 = (6\rho^4 - 6\rho^2 + 1) + (6\rho^2 - 3) + 2 = 6\rho^4, \quad (7\text{-}84)$$

so

$${}_0C_{40}\,\rho^4 = ({}_0C_{40}/6)R_4^0 + ({}_0C_{40}/2)R_2^0 + ({}_0C_{40}/3)R_0^0, \quad (7\text{-}85)$$

and, therefore, the Zernike coefficients are

$$A_{40} = {}_0C_{40}/6, \qquad A_{20} = {}_0C_{40}/2, \qquad \text{and} \quad A_{00} = {}_0C_{40}/3. \quad (7\text{-}86)$$

When the Zernike equivalent of ${}_0C_{40}\rho^4$ in Eq. (7-85) is substituted in Eq. (7-82) with the coefficients of Eq. (7-86), the following results:

$$W' = A_{44}\,R_4^4(\rho)\cos 4\varphi + A_{40}\,R_0^4(\rho) + A_{20}\,R_0^2(\rho) + A_{00}. \quad (7\text{-}87)$$

Then, applying the formula for the Strehl ratio, Eq. (7-76), we have

$$R_S' = 1 - \left[(2\pi)^2/\lambda^2\right]\left[(A_{44}^2/5) + (A_{40}^2/5) + (A_{20}^2/3)\right]. \quad (7\text{-}88)$$

Comparing Eq. (7-76) with Eq. (7-78), we find

$$R_S' = R_S - \left[(2\pi)^2/\lambda^2\right]\left[(A_{40}^2/5) + (A_{20}^2/3)\right], \quad (7\text{-}89)$$

so

$$R_S' < R_S \quad (7\text{-}90)$$

for all real values of ${}_0C_{40}$ other than zero, which indicates that the classical aberrations equivalent to the single Zernike aberration are balanced to produce the maximum possible value of the Strehl ratio.

COMPARISONS OF OPTIMIZING AND BALANCING PROCEDURES

An optical designer has available a number of different optimizing and balancing procedures, some of which have been discussed in the earlier pages of this chapter. Although further comparison and appraisal of these procedures are beyond the scope of this book, the reader may benefit by knowing the nature of some authoritative discussions of the procedures. Prominent among these works are papers by Hopkins [3], Rosenbruch [17], and Rosenhauer et al. [18].

Hopkins compares three different procedures for optimizing and balancing using the following criteria or merit functions:

1. Variance of wave aberration $\mathcal{E}$, Eq. (7-8).
2. Mean square of wave aberration $\overline{W^2}$, Eq. (7-5).
3. Mean square of transverse aberration $\overline{\epsilon^2}$.

The first and second merit functions are defined by the indicated equations. The third requires some extension of definitions in Chapter 3.

Image plane canonical coordinates G'_S and H'_T are defined by Eqs. (3-33) and (3-34). Aberrations produce increments $\delta G'_S$ and $\delta H'_T$ in these coordinates, and the increments are related to partial derivatives of the wave aberration W according to Eq. (3-37). The transverse aberration ϵ is defined by

$$\epsilon^2 = (\delta G'_S)^2 + (\delta H'_T)^2, \tag{7-91}$$

and the mean square value is given by

$$\overline{\epsilon^2} = (1/\mathcal{A}) \int\int_{\mathcal{A}} \epsilon^2 \, d\mathcal{A}. \tag{7-92}$$

The wave aberration power series is broken into two polynomials, as in previous developments:

$$W_{\text{even}} = C_{20}\,\rho^2 + C_{40}\,\rho^4 + C_{60}\,\rho^6 \quad \text{spherical aberration,} \tag{7-93}$$

$$W_{\text{odd}} = C_{11}\rho \cos\varphi + C_{31}\rho^3 \cos\varphi + C_{51}\rho^5 \cos\varphi \quad \text{coma.} \tag{7-94}$$

When these polynomials are written in rectangular coordinates and $\delta G'$s and $\delta H'_T$ are evaluated by the partial derivatives of W_{even} and W_{odd}:

$$\overline{\epsilon^2_{\text{even}}} = 2C^2_{20} + 4C^2_{40} + 6C^2_{60} + \frac{16C_{20}C_{40}}{3} + \frac{48C_{40}C_{60}}{5} + 6C_{60}C_{20}, \tag{7-95}$$

for spherical aberration, and

$$\overline{\epsilon^2_{\text{odd}}} = C^2_{11} + \frac{5C^2_{31}}{3} + \frac{13\,C^2_{51}}{5} + 2C_{11}C_{31} + 4C_{31}C_{51} + 2C_{51}C_{11}, \tag{7-96}$$

for coma. These equations correspond to the equations for ξ_{even}, Eq. (7-45), and ξ_{odd}, Eq. (7-51); optimum primary-to-secondary coefficient ratios ($\beta_{46} = -C_{40}/C_{60}$ for spherical aberration, $\beta_{35} = -C_{31}/C_{51}$ for coma) can be determined for ϵ^2_{even} and ϵ^2_{odd} in a manner similar to finding the optimum ratios for $\mathcal{E}$.

Table 7.I Optimum Primary-to-Secondary Coefficient Ratios[a]

Merit Function	$\mathcal{E}$	$\mathcal{E}_0$	$\overline{W^2}$	$\overline{W_0^2}$	$\overline{\epsilon^2}$	$\overline{\epsilon_0^2}$
		Spherical Aberration β_{46}				
By integration	1.50	0.94	1.33	0.83	1.30	1.20
By discrete sum	—	1.17	1.50	0.97	2.25	1.42
		Coma β_{35}				
By integration	1.20	0.80	1.20	0.80	1.5	1.20
By discrete sum	—	1.27	1.50	0.94	2.5	1.50

[a] From Hopkins.

Instead of evaluating just three optimum ratios each for spherical aberration and for coma, Hopkins introduced two kinds of variations that increased the evaluations to 11 for each aberration. Wherever an integration was indicated, he would make a parallel evaluation by a discrete sum based on a marginal ray ($\rho = 1$) and a zonal ray ($\rho = 0.707$). Also, besides evaluations for optimum focus, he repeated these evaluations for the paraxial focus ($C_{20} = 0$, $C_{11} = 0$). His results are given in Table 7.I. The subscript 0 indicates criteria for paraxial focus.

The wide ranges of the values in the two parts of the table suggest that there is more to an optimum balance than simply calculating a coefficient ratio. The

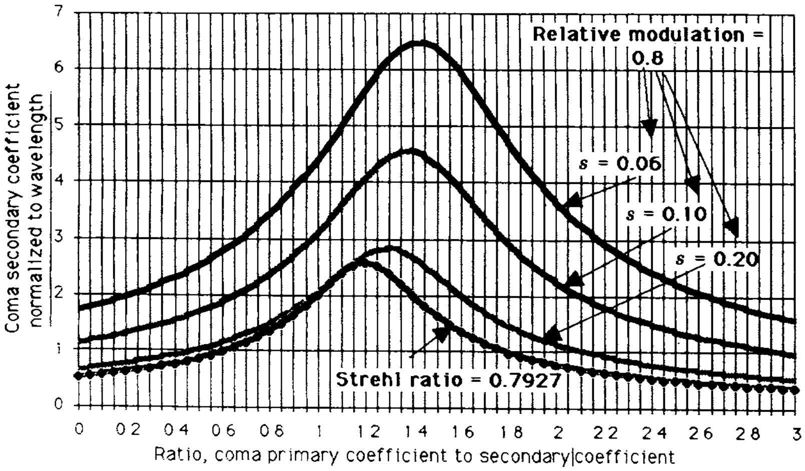


Figure 7.8. Comparison of secondary coefficient maxima for coma [16].

designer must study the application of the optical system to decide which merit function most closely coincides with the purposes of the system. Then, if discrete sums are used to evaluate integrals, the selection of which rays and how many of them (with possible weighting) will obviously affect the results.

As a further demonstration of how balance is dependent on which criterion is applied, Rosenbruch [17] adds the coma curve of Fig. 7.5 to the curves of Fig. 7.7*b* to produce Fig. 7.8 where the added curve is labeled "Strehl ratio = 0.7927." It is apparent that the maximum value of the coma secondary coefficient occurs at a lower value of the primary-to-secondary ratio for a fixed Strehl value than for any of the indicated relative modulation curves. Again, the designer has to make a decision: Does the Strehl ratio provide a better criterion than the OTF-based relative modulation at a particular frequency?

THE EFFECT OF OPTICAL PARAMETER VARIATIONS ON THE OPTICAL TRANSFER FUNCTION

By changing optical construction parameters and calculating the resulting changes in the MTF (the absolute value of the OTF), Rosenbruch [19] studied the influence of parameter inaccuracy on system performance. He worked with an $f/5$ system with a focal length of 300 mm. Seven optical surface radii were involved; he changed each in succession by 1% and found that the region occupied by the various resulting MTF curves was as shown by the cross-hatched area in Fig. 7.9. The total variation, as indicated, was about 0.1 of the MTF.

To make a similar study, King [10] chose to differentiate the OTF with respect to the design merit function, which was the squared OTF with certain weightings. This criterion allows a wide range in OTF values, and neither the spatial frequency nor the aberration function need be limited to small values. The mathematics of King's approach is outlined in the following paragraphs.

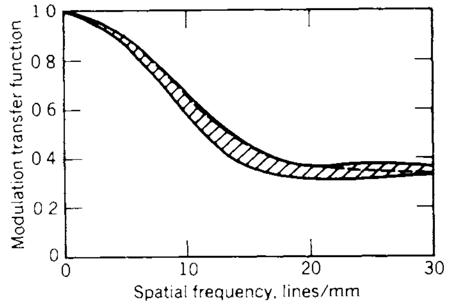


Figure 7.9. Region of variation of the MTF curve for 1% parameter variations [19].

The wave aberration $W(x, y)$ is assumed to be a function of N construction parameters p_n where

$$n = 1, 2, 3, \ldots, N. \tag{7-97}$$

If changes Δp_n are kept sufficiently small, the following truncated expansion of a Taylor's series is a valid approximation of $W(x, y)$ at parameter values p_n:

$$W(x, y; p_n) = W_0(x, y) + \sum_{n=1}^{N} \frac{\partial W}{\partial p_n} \Delta p_n, \tag{7-98}$$

where W_0 is the value of W for initial p_n values of p_{n0}. The transfer function $\hat{b}_0$ (s) for this $W(x, y; p_n)$ can be expressed by substitution in Eq. (5-95):

$$\begin{aligned} \hat{b}_0(s) = \int\int_{\mathfrak{A}} \exp\Bigg[-ik \Bigg\langle \Bigg[W_0\left(x + \frac{s}{2}, y\right)\Bigg] \\ + \sum_{n=1}^{N} \left\{\left[\partial W_0\left(x + \frac{s}{2}, y\right)\right] / \partial p_n \right\} \Delta p_n \\ - \left[W_0\left(x - \frac{s}{2}, y\right)\right] - \sum_{n=1}^{N} \left\{ \cdot \left[\partial W_0\left(x - \frac{s}{2}, y\right)\right] / \partial p_n \right\} \Delta p_n \Bigg\rangle \Bigg] \\ \cdot\, dx\, dy. \end{aligned} \tag{7-99}$$

The index of refraction n is assumed to be unity. Rearrangement of exponential terms gives

$$\begin{aligned} \hat{b}_0(s) = \int\int_{\mathfrak{A}} \exp\Bigg[-ik \Bigg\langle \left[W_0\left(x + \frac{s}{2}, y\right)\right] - \left[W_0\left(x - \frac{s}{2}, y\right)\right] \\ + \sum_{n=1}^{N} \Bigg(\left\{\left[\partial W_0\left(x + \frac{s}{2}, y\right)\right] / \partial p_n \right\} \\ - \left\{\left[\partial W_0\left(x - \frac{s}{2}, y\right)\right] / \partial p_n \right\} \Bigg) \Delta p_n \Bigg\rangle \Bigg]\, dx\, dy. \end{aligned} \tag{7-100}$$

For convenience, groups of terms will be represented by single symbols as

$$\hat{\mu} = \hat{\mu}(x, y; s) = \exp\left[-ik \left(\left[W_0\left(x + \frac{s}{2}, y\right)\right] - \left[W_0\left(x - \frac{s}{2}, y\right)\right]\right)\right], \tag{7-101}$$

$$\sigma_n = \sigma_n(x, y; s) = \left\{\left[\partial W_0\left(x + \frac{s}{2}, y\right)\right]/\partial p_n\right\} - \left\{\left[\partial W_0(x - \frac{s}{2}, y\Big)\right]/\partial p_n\right\}. \tag{7-102}$$

Making these substitutions in Eq. (7-100) yields

$$\hat{b}_0(s) = \int\int_{\mathcal{A}} \hat{\mu} \exp\left(-i\mathscr{k} \sum_{n=1}^{N} \sigma_n \,\Delta p_n\right) dy\, dy. \tag{7-103}$$

The exponential factor in the integrand can be expanded into the power series whose form is

$$\exp \theta = e^{\theta} = 1 + \theta + \theta^2/2! + \theta^3/3! + \cdots. \tag{7-104}$$

If Δp_n is sufficiently small, the infinite series can be adequately approximated by the first three terms (within braces):

$$\hat{b}_0(s) = \int\int_{\mathcal{A}} \hat{\mu}\left\{1 - i\mathscr{k} \sum_{n=1}^{N} \sigma_n \,\Delta p_n + (\mathscr{k}^2/2)\left[\sum_{n=1}^{N} \sigma_n \,\Delta p_n\right]^2\right\} dx\, dy. \tag{7-105}$$

If only one parameter p_m, where $1 \leq m < N$, is changed, the corresponding change in $\hat{b}_0(s)$ can be reached by taking its partial derivative with respect to Δp_m:

$$\frac{\partial[\hat{b}_0(s)]}{\partial[\Delta p_m]} = \int\int_{\mathcal{A}} \hat{\mu}\left\{-i\mathscr{k}\, \sigma_m + \mathscr{k}^2\left[\sum_{n=1}^{N} \sigma_n \Delta p_n\right] \sigma_m\right\} dx\, dy. \tag{7-106}$$

For further development of this equation, it is convenient to express each term of the complex integrand in rectangular form, explicitly defining for each a real (Re) and an imaginary (Im) part,

$$\mathrm{Re}(m, n) + i\mathrm{Im}(m, n) = \mathscr{k}^2 \int\int (\hat{\mu}\sigma_n\sigma_m \,dx\, dy)\, \Delta p_n, \tag{7-107}$$

where $n \neq 0$, and

$$\mathrm{Re}(m, 0) + i\mathrm{Im}(m, 0) = -\mathscr{k}\int\int_{\mathcal{A}} \hat{\mu}\sigma_m \,dx\, dy, \tag{7-108}$$

where $n = 0$. Substituting these defined expressions in Eq. (7-106) yields

$$\frac{\partial[\hat{b}_0(s)]}{\partial[\Delta p_m]} = \mathrm{Re}(m, 0 + i\mathrm{Im}(m, 0) - \sum_{n=1}^{N} [\mathrm{Re}(m, n) + i\mathrm{Im}(m, n)]\, \Delta p_n. \tag{7-109}$$

We also define a rectangular form for $\hat{b}_0(s)$,

$$\mathrm{Re}_1(s) + i\mathrm{Im}_1(s) = \hat{b}_0(s), \tag{7-110}$$

which, when differentiated with respect to Δp_m, becomes

$$\frac{\partial[\hat{b}_0(s)]}{\partial[\Delta p_m]} = \frac{\partial[\mathrm{Re}_1(s)]}{\partial[\Delta p_m]} + i\frac{\partial[\mathrm{Im}_1(s)]}{\partial[\Delta p_m]}. \tag{7-111}$$

Since both are the same partial derivative, the right sides of Eq. (7-109) and (7-111) are equal. As in all complex equations, the real parts on the two sides of the resulting equation are equal; the imaginary parts on the two sides are also equal:

$$\frac{\partial[\mathrm{Re}_1(s)]}{\partial[\Delta p_m]} = \mathrm{Re}(m, 0) - \sum_{n=1}^{N} [\mathrm{Re}(m, n)]\, \Delta p_n, \tag{7-112}$$

$$\frac{\partial[\mathrm{Im}_1(s)]}{\partial[\Delta p_n]} = \mathrm{Im}(m, 0) - \sum_{n=1}^{N} [\mathrm{Im}(m, n)]\, \Delta p_n. \tag{7-113}$$

King chooses a merit function having several terms, two of which are directly related to the OTF. One of these is

$$Q \sum_{n=1}^{N} q_n^2(\Delta p_n)^2, \tag{7-114}$$

where Q is a positive damping factor and the q_n's are damping coefficients. The other OTF-related term is

$$\phi_{\mathrm{D}} = 1 - \sum [\eta_a \mathrm{Re}_1^2(s) + \eta_b \mathrm{Im}_1^2(s)], \tag{7-115}$$

where the η's are positive weighting factors and the summation is over a set of off-axis image points. Since Re_1 and Im_1 are defined by Eq. (7-110), the quantity in the brackets of Eq. (7-115) is the squared value of the transfer function $\hat{b}_0(s)$ when $\eta_a = \eta_b = 1$.

Equation (7-115) could represent alternative merit function terms where the summation is over different spatial frequencies, different azimuths ψ, or different wavelengths λ. The weighting factors must be controlled so that ϕ_D remains positive during an optimization procedure. Minimizing the merit function term ϕ_D involves the partial derivative,

$$\frac{\partial\phi_D}{\partial(\Delta p_m)} = -2\sum\left\{\eta_a \mathrm{Re}_1(s)\frac{\partial \mathrm{Re}_1(s)}{\partial(\Delta p_m)} + \eta_b \mathrm{Im}_1(s)\frac{\partial \mathrm{Im}_1(s)}{\partial(\Delta p_m)}\right\}. \quad (7\text{-}116)$$

The real part of $b_0(s)$, which is defined in Eq. (7-110), is the real part of Eq. (7-105); the partial derivative of $\mathrm{Re}_1(s)$ is given by Eq. (7-112). Similarly, the term in Eq. (7-116) involving the imaginary part of $\hat{b}_0(s)$ can be obtained from Eqs. (7-105) and (7-113).

REFERENCES

1. B. Brixner, Lens Design Merit Functions: RMS Image Spot Size and RMS Optical Path Difference. *Appl. Opt.* **17,** 715 (1978).
2. A. E. Conrady, *Applied Optics and Optical Design*, Vol. 1. (Please see the reference to Conrady's book at the end of Chapter 6.)
3. H. H. Hopkins, The Use of Diffraction-Based Criteria of Image Quality in Automatic Optical Design. *Opt. Acta* **13,** 343 (1966).
4. E. M. Granger, Subjective Assessment and Specification of Color Image Quality. *SPIE Proc.* **46,** 86 (1974). (Please see the note following Ref. 2 of Chapter 1.)
5. K. G. Birch, A Survey of OTF Based Criteria Used in the Specification of Image Quality, National Physical Laboratory, Op. Met. 5, April 1969 (Teddington, Middlesex, England).
6. J. Meiron, The Use of Merit Functions Based on Wavefront Aberrations in Automatic Lens Design. *Appl. Opt.* **7,** 667 (1968).
7. R. W. Gostick, OTF-Based Optimization Criteria for Automatic Optical Design. *Opt. Quantum Electron.* **8,** 31 (1976).
8. W. B. King, Dependence of the Strehl Ratio on the Magnitude of the Variance of the Wave Aberration. *J. Opt. Soc. Am.* **58,** 655 (1968).
9. W. B. King, A Direct Approach to the Evaluation of the Variance of the Wave Aberration. *Appl. Opt.* **7,** 489 (1968).
10. W. B. King, Use of the Modulation-Transfer Function (MTF) as an Aberration-Balancing Merit Function in Automatic Lens Design. *J. Opt. Soc. Am.* **59,** 1155 (1969).

11. W. B. King, Correlation between the Relative Modulation Function and the Magnitude of the Variance of the Wave-Aberration Difference Function. *J. Opt. Soc. Am.* **59,** 692 (1969).

12. W. B. King, The Use of the Modulation-Transfer Function (MTF) as an Aberration-Balancing Merit Function in Automatic Lens Design. In *Optical Instruments and Techniques*, J. Home Dickson (Ed.), Oriel Press, Stocksfield, Northumberland, UK. (Please see Ref. 18 of Chapter 2 for more information.)

13. A. Maréchal, Study of the Combined Effect of Diffraction and Geometrical Aberrations on the Image of a Luminous Point. *Rev. d'Opt.* **26,** 257 (1947).

14. M. Born and E. Wolf, *Principles of Optics*. Pergamon, London, 1965, p. 469.

15. H. H. Hopkins, The Aberration Permissible in Optical Systems. *Proc. Phys. Soc. London Ser. B* **70,** 449 (1957).

16. S. N. Bezdid'ko, The Use of Zernike Polynomials in Optics. *Sov. J. Opt. Technol.* **41,** 425 (1974).

17. K.-J. Rosenbruch, Use of OTF-Based Criteria in Automatic Optical Design. *Opt. Acta* **22,** 291 (1975).

18. K. Rosenhauer, K.-J. Rosenbruch, and F.-A. Sunder-Plassman, The Relations between the Axial Aberrations of Photographic Lenses and Their Optical Transfer Functions. *Appl. Opt.* **5,** 415 (1966).

19. K.-J. Rosenbruch, Comparison of Different Methods of Assessing the Performance of Lenses. In *Optical Instruments and Techniques*, J. Home Dickson (Ed.), Oriel Press, Stocksfield, Northumberland, UK. (Please see Ref. 18 of Chapter 2 for more information.)

8

Measurement

INTRODUCTION

The ability to measure an optical transfer function (OTF) with sufficient accuracy and speed at a practicable cost is a goal of high significance in the art of optical instrumentation. Just the confidence among designers and users that a measurement of the OTF in one laboratory, or on a specific set of equipment, could be reproduced reliably at another time and place would most certainly broaden the use of the OTF for the specification and proving of optical systems. Optical quality could then be generally specified not only in terms of geometric fidelity between image shape and object shape but also in terms of a fidelity of contrast between given points in object and corresponding points in image.

As one reviews the sweat and frustration spent on OTF instrumentation, hindsight suggests a guiding adage that might have made life more productive for workers in the field: As the refinement of optical systems is pushed to higher and higher precision in the balancing of smaller and smaller residual aberrations, with ever higher numerical apertures and ever wider fields of view, old standards of optical measurement must be correspondingly raised. Measurement procedures require more careful planning, the measuring equipment itself must become more elaborate, and the resulting measurements have to be carried out to greater accuracy than ever before.

Through the work of a group at the SIRA Institute, Ltd., the British Calibration Service was among the first to pin down the sources of error and spell out the limitations of OTF measurement equipment. They prepared a set of standard test lenses that were measured by a number of different laboratories. Comparison of the resulting data by an error analysis indicated how the limitations might be reduced, and specifications for designing and building an adequate test facility were assembled. SIRA has also suggested how existing facilities for measuring the OTF could be improved.

As later sections of this chapter indicate, schemes for measuring the OTF proliferated during the period from the late 1940s into the early 1960s. Unfortunately most of the resulting equipment, though often ingenious, emphasized speed and convenience rather than precision. Two excellent review papers by Murata and by Rosenhauer and Rosenbruch [1, 2] describe this interesting period. The weight of the evidence, discussed in some detail near the end of this chapter, seems to favor the interferometric method of measuring wave-front distortion and subsequent calculation of the OTF by autocorrelation.

A present challenge is to measure consistently the OTF at off-axis points in the image field. On-axis results are now reasonably reproducible between laboratories with tolerable error. Even with the same kind of equipment, though, consistency begins to be quite difficult at 10° off axis.

Certain construction difficulties contribute to off-axis measurement errors. A well-made lens that exhibits excellent performance in a practical application often falls short when the MTF or spread function is checked for consistency at a given radius while the lens is rotated about its optical axis. This might be due to a residual, though minute, decentering or tilting of individual elements. Lens seats and screw threads can be out of tolerance or adjustment to increase further the noncoincidence of the mechanical axis (the center line of the cylindrical barrel) with the theoretical optical axis.

Maintaining cylindrical symmetry in laboratory measurements is fundamental. In an MTF instrument itself, it is usually taken for granted that the object-field slide, the image-field slide, and the seating flange of the lens holder are all parallel. However, when a lens is mounted in the system, the optical axis of the lens generally does not coincide exactly with the axis of the MTF instrument. As a result, the image plane of the lens does not coincide with the plane of the image field slide. By making a sequence of off-axis MTF or spread function measurements for different lens aspects, a lens position can be found that produces balanced readings, that is, symmetry about the instrument axis; and the image plane of the lens intersects the plane normal to the axis of the MTF instrument along a line parallel to the edge of the image-field slide [3].

OTF standards are beginning to appear in various countries. In Great Britain OTF measuring equipment must meet certain performance criteria to gain approval by the British Calibration Service [4]. In the United States the American Standards Institute (ANSI) has established a standard on the OTF [5]. A number of other countries either already have OTF standards or are preparing them [6–9].

To complement the work that is going forward on OTF measurements, optical authorities recognize that a broadened optical quality concept requires also explicit standards on veiling glare light, light transmittance, distortion, and angular aberration, all of which are quite independent of OTF measurements.

In earlier chapters, the OTF is often discussed in terms of its two parts, the MTF (modulation transfer function) and the PTF (phase transfer function). As stated near the end of Chapter 6, the MTF, besides being the far easier of the two to measure, has proved of much greater significance than the PTF. Small wonder, then, that purported work on OTF measurement procedures turns out to be about only the MTF part of that function.

As the various equipments for measuring the OTF are reviewed, two paradoxes become apparent. First, extended objects are sometimes used for test purposes, but OTF theory is based on a point source in object space at a given distance r from the axis; furthermore, image-forming properties of the optical system are, in general, dependent on r. Second, when other than laser sources are employed, the object is illuminated by light having a considerable range in the wavelength λ. The image-forming properties of the optical system are in general dependent also on λ.

When an extended object is part of an OTF measuring system, the variation of measured results must either be negligible over the range of the distance r involved, or a well-defined average must be accepted for each result. The concept of isoplanatism, defined in Chapter 2, provides the basis for accepting extended objects in measurement situations.

When the test illumination has a relatively broad spectrum, the optical sys-

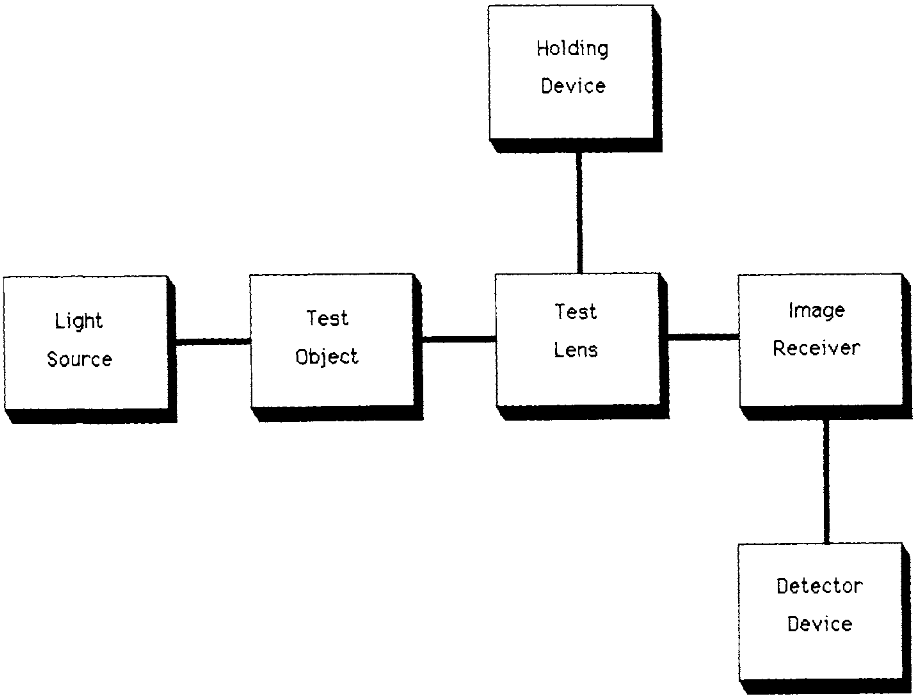


Figure 8.1. Essential components of an OTF measuring equipment.

tem must be free of chromatic aberration over that spectrum to give results equivalent to testing with a narrow range of λ. Otherwise, a well-defined average, again, must be accepted for each result.

COMPONENTS OF A MEASURING SYSTEM

The basic components of an experimental setup for measuring the OTF are illustrated in the block diagram of Fig. 8.1. The essential components of any such system are (1) the illumination system or *light source*, (2) a *test object*, which can take the form of a slit, a half plane, or a grating, (3) the *test lens*, which is the lens being tested, (4) a *holding device* to hold the test lens with the required positioning and displacement accuracy, (5) an *image receiver* with the required placement and motion accuracy in the image plane of the test lens, and (6) a suitable and convenient *electronic detection device* to provide an electrical response for a data output. The variations of the details of these components are discussed in the subsequent sections of this chapter. The major differences among setups lie in the form of the test object and the consequent form of the image receiver. Other differences occur in the nature and particular characteristics required of the detection device.

REQUIREMENTS OF THE COMPONENTS

Common to the large number of current designs of OTF measuring equipment are certain basic components, which will be discussed briefly here by outlining the principles involved and by describing some of the sources of error.

Precision in optical measurements starts with the quality of the optical bench—especially in techniques as sensitive to error as an OTF measurement. Errors peculiar to the bench itself must be known and minimized; these compromise the measurement more and more, as already mentioned, as the object point is moved further off axis.

The OTF measurement requires a particularly stable device for holding the test lens. The device must also provide precise adjustments for displacements so that measurements can be made in a defined image plane for different field angles and for different azimuths. As stated in Chapter 7, an observer's threshold in the perception of change in image quality is about a 10% change in the area under the MTF curve. With this as reference, a feeling for the bench displacement precision required can be gained from an example by Rosenhauer and Rosenbruch [2]: A displacement of the image plane by 3 μm causes a change of about 10% in the MTF for a lens with a numerical aperture of 0.25 (f/number of 2). Mechanical precision requirements are dependent upon focal length and aperture in such a way as to make errors increase with increasing numerical aperture. When the required optical error tolerance is determined for a given test, a careful evaluation of the corresponding mechanical error allowable for the image plane scanning device must be made to check the adequacy of the available bench setup. As stated earlier, off-axis field points at angles as low as 10° begin to challenge typical laboratory equipment.

A particularly convenient method of determining bench errors has been discussed by Marchant and Ironside [10] and by Marchant [11]. Figure 8.2 shows schematically the bench parts involved. The test lens is clamped to a mounting face that is assumed perfectly flat and parallel to the *reference surface*. If the test lens were perfectly corrected and perfectly made, its actual image plane would be exactly parallel to the reference surface and would assume the position of the *ideal image plane* indicated by the dashed line in the figure. OTF measurements are typically made at various points in the image plane. In making these measurements, the *image analyzing device*, typically a slit, should move in the ideal image plane; but, in practice, many different kinds of mechanical errors in the bench, such as curved slide ways and general flexure when the bench is turned to different field angles θ, combine to cause the analyzing device to move in some such path as indicated by the curved line. The consequent error δz is thus a function of position in the image plane and of the accumulated mechanical bench errors, which are in turn functions of the field angle.

Marchant and Ironside [10] determined the influence δz-type errors had on the MTF of a certain high-quality, wide-angle lens ($f/4.5$, 30-cm focal length). In Fig. 8.3 the MTF is plotted for various field angles as a function of the

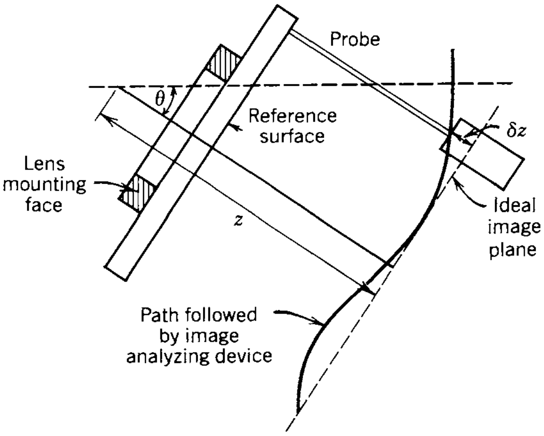


Figure 8.2. Diagram illustrating a method of defining and measuring optical bench error δz [10].

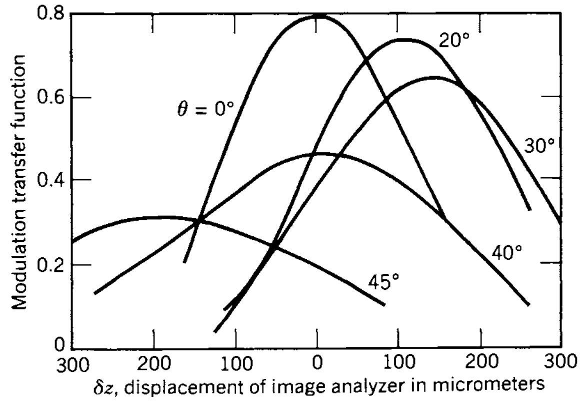


Figure 8.3. "Through-focus" OTF curves for a high-quality, wide-angle lens (30 cycles/mm, aperture: $f/4.5$) [10].

displacement of the image analyzer from the ideal image plane, which was chosen as the position where the on-axis ($\theta = 0°$) MTF attained its peak value. These *through-focus* curves were all recorded at a spatial frequency of 30 cycles/mm. Because of the choice of image plane position, some of the off-axis curves, for instance the one for $\theta = 30°$, have a steep slope at zero displacement; so the MTF value for these curves is comparatively sensitive to small changes in analyzer displacement near the ideal image plane.

The relative shapes and positions of the through-focus curves, such as those of Fig. 8.3, vary considerably from lens to lens. A flat-field lens of otherwise poor correction, for instance, would produce shallow curves that reach their peaks very nearly at the same image plane position. As a result, the error tolerance for δz would be comparatively great.

Marchant and others [12] have analyzed the MTF measurements made by nine different laboratories on a standard wide-angle lens. Three figures, Figs. 8.4–8.6, show graphical results of these measurements. Figure 8.4 is a plot of one set of results. The MTF for both the radial and tangential directions, from on-axis up to 40° off-axis, are shown. The ideal image plane, zero on the displacement scale, was placed at the peak of the tangential on-axis MTF curve. To facilitate comparison of the various curves, the on-axis data are for a spatial frequency of 30 cycles/mm; and the off-axis data are for 10 cycles/mm. To compare various measurements at different laboratories, nine sets of data, recorded under supposedly identical conditions, were plotted on a common pair of axes as in Figs. 8.5 and 8.6. In Fig. 8.5, the radial MTFs 10° off-axis are plotted against spatial frequency; in Fig. 8.6 the tangential MTFs 20° off-axis are plotted also against spatial frequency.

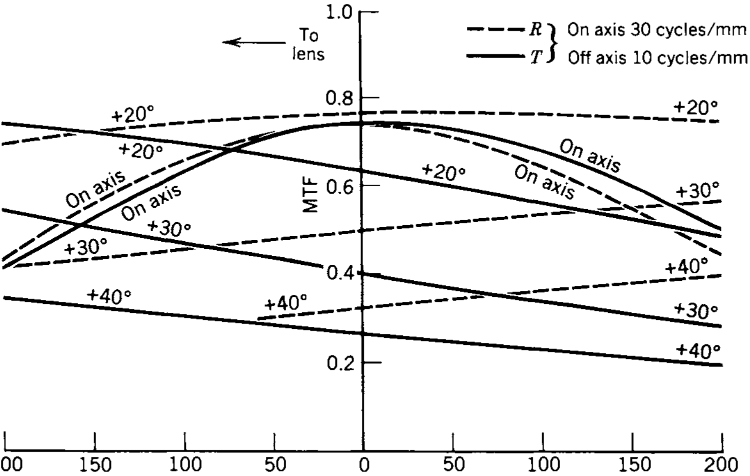


Figure 8.4. Variation of MTF with focus position and field angle [12].

At first sight, the discrepancies in the results from the nine laboratories may appear to discredit the MTF for the specification and performance assessment of lenses. However, the study by Marchant and others came up with the following conclusions: Analysis indicates that the differences among the laboratories' results can be attributed largely to errors in setting up the lens on the

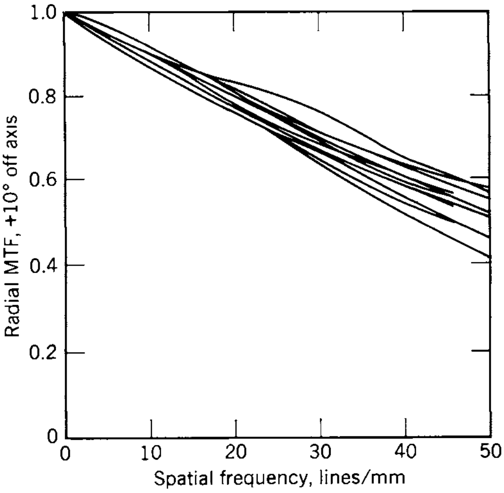


Figure 8.5. Comparison of MTF measurements made at different laboratories. Radial MTFs, 10° off-axis, versus spatial frequency [12].

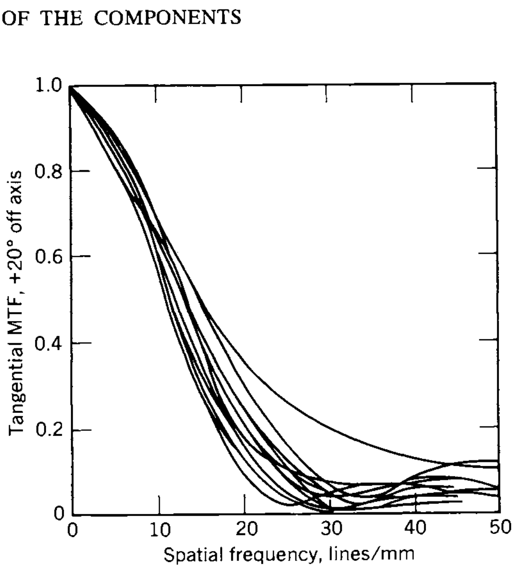


Figure 8.6. Comparison of MTF measurements made at different laboratories. Tangential MTF, 20° off-axis, versus spatial frequency [12].

optical bench, to mechanical misalignment of the bench itself, and to errors in the spatial frequency calibration of the MTF measuring equipment. With diligent attention to these sources of error, it is estimated that MTF measurements can be repeated to within ±0.05.

A substantial body of evidence suggests that ±0.05 uncertainty in the MTF is small relative to the least-detectable difference in picture quality. So, with attention to this uncertainty and the quality of the laboratory making measurements, the MTF can be reliably employed in procurement specifications. It appears especially appropriate to use the MTF as a quality control tool in photographic lens production.

What maximum value of δz can be tolerated in making acceptable MTF measurements? A general answer is probably impossible to formulate, but a useful approximation can be reached by limiting the scope of optical parameters in the test lens. As in aberration considerations, the ultimate precision required in any measurement corresponds to reducing errors to match the errors caused only by diffraction (diffraction-limited lens). With this provision and with the ±0.05 latitude in MTF already discussed, the tolerance on δz becomes a function of spatial frequency and relative aperture. With reference to Fig. 8.4, we make the conservative assumption that the through-focus curve of the MTF being measured is so located along the displacement axis that the maximum slope occurs at the origin (position of the image plane). Then, with rather arbitrary assumptions as to the typical ranges of spatial frequencies and of apertures, the simplified relation for the maximum error that can be tolerated is

$$\delta z = [54\ (f/\text{No.})/\omega]\ \mu\text{m}, \tag{8-1}$$

where $\Delta(\text{MTF}) = 0.05$, $\lambda = 0.546\ \mu\text{m}$, and ω is the spatial frequency in cycles/mm. Although the mechanical tolerances calculated by this formula may appear extremely small, a review of the data-taking represented by Figs. 8.2 and 8.3 suggests that this tight a restriction on δz is a legitimate target for the bench manufacturer [10].

A reliable value for δz cannot generally be reached by dealing with the total of such errors as those caused by departures from straightness of slide-ways and the lack of parallelism between transverse slides and the lens mounting flange. The various errors must be assessed individually. Finally, to get an accurate evaluation of δz, it should be measured directly as a function of field angle θ.

Most OTF measuring instruments involve a narrow slit either as the test object or, on the other side of the lens, to scan the aerial image formed by the test lens of an extended object. Errors occur when the nominal size and geometry of the slit are accepted without a careful check. Typical spectrometer slits are calibrated for various widths, but each control setting must be checked against actual slit width. This is especially important when the precision of the measurement calls for a correction in the MTF value based on slit width. Also, the slit should be minutely examined to assure constancy of width along its length, straightness of its edges, and coplanarity of the slit jaws. The difficulty of this task can be appreciated when it is found, as demonstrated in a later section, that the slit width for MTF measurements is of the order of a micrometer.

For MTF measurements at large field angles or at large apertures, the response of the image receiver to obliquely incident beams becomes important. Some mechanically adjustable slits, however finely honed, have a depth many times the width; so the opening presents a tunnel rather than a theoretically desirable slit to the incident beam. The cross section of the beam admitted through such an opening can be a complicated function of the field angle. This ''tunnel effect'' can be minimized by ruling the slits in thin metal films. Another approach to the problem is to form a reduced image of a regular spectrometer slit with a low-power microscope objective. Kuttner [13] has shown the significant effects of microscope objectives on MTF measurements made at large apertures and at large field angles.

OTF measuring devices make use of grating transparencies in a number of different ways. These line patterns, usually involving a sine wave variation, can be test objects or be in the image field either as scanners or as patterns scanned by images of special test objects. In these applications, the purpose of the measurement is usually to determine how the contrast of the line pattern varies with spatial frequency or with the wavelength of transmitted light. Knowing the precise frequency in such measuring is critical to getting a true MTF curve shape [10].

As we have established in earlier chapters, the validity of the optical transfer function requires incoherent illumination, with the exception of the interferometric or autocorrelation method of measurement. This incoherence is usually attained by arranging a sufficiently large aperture in the illumination system and providing ground glass or some other diffuser.

Since aberrations are dependent on distance to the object, the bench setup must comply with the specified distance, either by actual spacing or by simulation with an optical auxiliary such as a collimator. When optical auxiliaries are introduced, care must be taken neither to disturb the conditions for incoherence nor to compromise the measured OTF by including the effects of aberrations in the auxiliaries. Backing out the contributions of auxiliary optics from an overall measurement is usually an extremely difficult procedure; so, whenever possible to avoid this difficulty, the quality of the auxiliary should far exceed that of the test lens, or the auxiliary should be dispensed with altogether. The unwanted effects of auxiliary optics may show up either as part of the geometrical aberrations or in the wave-front aberration, depending on the measurement method used.

DIRECT METHODS

Perhaps the most direct method of measuring the OTF is the one represented by Fig. 8.7*a*. A grating is placed in the specified object position for the test lens with the grating lines parallel to, for example, the η-axis. With this setup, the variation of the transmittance τ as a function of the position coordinate ξ, perpendicular to the grating lines, is

$$\tau(\xi) = \tau_a + \tau_b \cos \pi(\xi/\xi_1), \tag{8-2}$$

which corresponds to Eq. (2-3) with a change in notation. When the grating transparency is illuminated by an incoherent beam having a uniform incidance H at the grating, the distribution of exitance M over the grating, on the side toward the entrance pupil, is

$$M(\xi) = H\,\tau(\xi) = M_a + M_b \cos \pi(\xi/\xi_1). \tag{8-3}$$

(See the section on "Distributions of Physical Quantities" in Chapter 2.) In Eqs. (8-2) and (8-3), ξ_1 is the half period of the spatial frequency in the grating; so the spatial frequency is given by

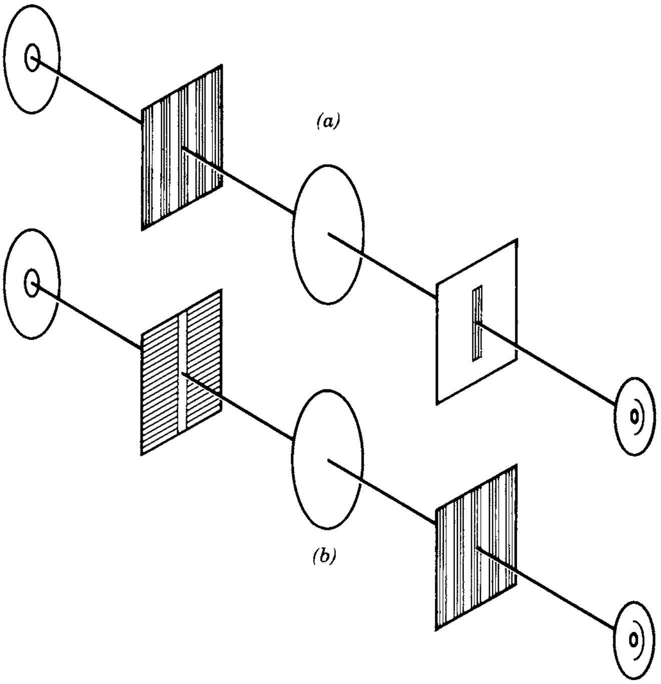


Figure 8.7. Two configurations for the direct measurement of the OTF: (*a*) Sinusoidal grating as the object; slit in the image plane. (*b*) Slit as the object; sinusoidal grating in the image plane.

$$\omega_1 = 1/(2\xi_1). \tag{8-4}$$

Methods for varying the spatial frequency are discussed in a subsequent paragraph.

Contrast in the object, according to Eq. (2-16) and with appropriate adjustment of nomenclature, is

$$C = M_b/M_a = \tau_b/\tau_a. \tag{8-5}$$

The test lens forms a sinusoidal aerial image (image in space without a screen) at the defined image plane. To determine the distribution of this aerial image experimentally, the flux density is measured photometrically by probing the light field at the image with a radiant energy detector. As the detector is moved at a constant speed in a path perpendicular to the sinusoidal bars in the image plane, it produces an electrical time-dependent signal of the form

$$i(t) = i_a + i_b \cos[(\pi t/t_1) + \phi], \tag{8-6}$$

where ϕ is the time phase advance (time phase lag if negative). The time and spatial quantities are related as follows:

$$t = \xi/v; \qquad t_1 = \xi_1/v, \tag{8-7}$$

where v is the velocity of scanning in a direction parallel to the ξ-axis. The measured contrast in the image is

$$C_i = i_b/i_a \tag{8-8}$$

and the modulation transfer function at the spatial frequency ω_1 is

$$T(\omega_1) = (i_b/i_a)/M_b/M_a) = (i_b M_a)/\, i_a M_b). \tag{8-9}$$

Each data point on the MTF curve requires a separate grating for each particular spatial frequency for which the MTF is measured.

The straightforward derivation of Eq. (8-9) is, in fact, oversimplified. In a subsequent section, it is shown that the finite length of the grating and the width of the slit both strongly affect the value of $T(\omega_1)$. These factors also enter into the measured value of ω_1. To take into account the grating length and the slit width, a special computation is required after each measurement; in fact, when precise results are needed, the additional compensating computation really denies the existence of a truly direct measurement procedure.

Certain experimental difficulties in the described procedure require attention. First, the radiant energy detector in the probe and its associated dc amplifier should be designed with as high signal-to-noise ratio as practicable. The detecting surface of the detector has to be in the shape of a slit, as narrow as possible to provide fine resolution along the ξ direction. However, the signal-to-noise ratio improves with increased cross section of the slit; so the slit should be long in the η direction to compensate for the required narrow width. Second, the alignment of the detector probe long dimension must be precisely parallel with the lines of the grating image; otherwise the advantage of narrow width is lost. Third, although the ξ–η-axes define an image plane, the moving probe may, according to the conditions of the test, have to follow instead the actual surface of the optimum image (maximum contrast). The location of the probe scan path relative to the ξ–η plane must be recorded as part of the measurement.

A variation of the procedure just described fixes the detecting probe slit and moves the grating instead, producing again a relative motion between image and probe. Another variation interchanges the positions of the grating and the slit; so, instead of the configuration of Fig. 8.7*a*, the parts are arranged as shown in the schematic of Fig. 8.7*b*. As indicated, the slit and the detector can no longer be incorporated in one physical entity since they are on opposite sides of the lens. Again, a time-varying electrical signal is produced by the detector as relative motion occurs between the slit image and the grating. Equivalence between these variants is discussed in more detail in subsequent paragraphs.

EFFECT OF FINITE GRATING LENGTH

In our oversimplified view of OTF measurement, we have assumed that the grating extends a great distance in each direction of the ξ coordinate. This cannot be attained practically, the actual extent being only from, say, some limited $-\xi_a$ to $+\xi_a$, with the arbitrarily placed origin of ξ located so that the object distribution function (representing the truncated sinusoidal distribution of the actual grating) is an even function. The frequency spectrum of the truncated sinusoidal distribution is

$$m(\omega) = \int_{-\infty}^{+\infty} M_{\mathcal{R}} M(\xi) \exp(-i2\pi\omega\xi)\, d\xi, \tag{8-10}$$

where $M_{\mathcal{R}}$ is a rectangular function (sometimes designated rects M) defined as

$$\begin{aligned} M_{\mathcal{R}} &= 1 \quad \text{when} \quad -\xi_a \leqq \xi \leqq \xi_a, \\ M_{\mathcal{R}} &= 0, \quad \text{when} \quad |\xi| > \xi_a, \end{aligned} \tag{8-11}$$

and $M(\xi)$ is defined by Eq. (8-3) where ξ extends to positive and negative values without limit. An alternative symbolic form for Eq. (8-10) is

$$m(\omega) \rightarrow \{M_{\mathcal{R}}M_a + M_{\mathcal{R}}M_b \cos(\pi\xi/\xi_1)\}, \tag{8-12}$$

where the arrow denotes "equals the Fourier transform of." To find the Fourier transform of the expression in braces, each of the two terms is operated on separately, which can be done by inspection with the help of the convolution theorem (see Appendix B). The first term in braces is a rectangular function having the amplitude M_a; its transform $m_1(\omega)$ is the well-known sinc function:

$$m_1(\omega) = [2\xi_a M_a \sin(2\pi\xi_a\omega)]/(2\pi\xi_a\omega). \tag{8-13}$$

This term is plotted in Fig. 8.8.

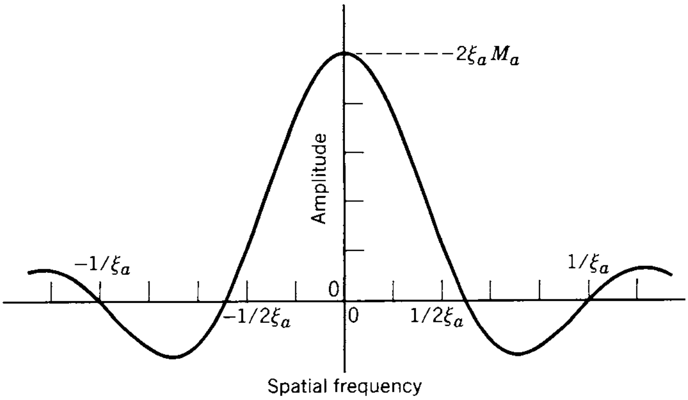


Figure 8.8. Sinc function term in the frequency spectrum of a truncated sinusoidal distribution.

The second term, $m_2(\omega)$, is the rects function with an amplitude of M_b multiplied by a sine wave of unit amplitude and unlimited extent. The Fourier transform of the sine wave is a unit value delta function at the spatial frequency ω_1 of the grating. The convolution of the rects spectrum with the delta function is the sinc function centered at $\pm\omega_1$; hence

$$m_2(\omega) = 2\xi_a M_b \left\{ \sin[2\pi\xi_a(\omega_1 - \omega)] / [2\pi\xi_a(\omega_1 - \omega)] \right\}. \tag{8-14}$$

The complete spectrum is therefore

$$\begin{aligned} m(\omega) = m_1(\omega) + m_2(\omega) &= 2\xi_a \Big\{ M_a \big\{ [\sin(2\pi\xi_a\omega)] / (2\pi\xi_a\omega) \big\} \\ &\quad + M_b \big\{ \sin[2\pi\xi_a(\omega_1 - \omega)] / [2\pi\xi_a(\omega_1 - \omega)] \big\} \Big\}. \end{aligned} \tag{8-15}$$

This spectrum has three segments: $m_1(\omega)$ centered at the origin and two segments in $m_2(\omega)$, one centered at $-\omega_1$ and the other at $+\omega_1$. The widths of the graphical lobes in the three segments are determined by ξ_a. The first zero of $m_1(\omega)$ is at $1/(2\xi_a)$.

As indicated by the foregoing brief analysis of the total spectrum, the effect of working with a grating of finite length is that, instead of having a single frequency ω_1, the measuring apparatus simultaneously generates a distribution of frequencies in the vicinity of $\omega = 0$ and another distribution around the desired ω_1, as shown in the plot of positive amplitudes in Fig. 8.9. It is obvious that as the length of the grating $2\xi_a$ is increased, the "skirts" of each peak are drawn in closer to the frequency of the peak and, in the limit, squeeze in to produce a single frequency at ω_1 and a single zero frequency, a "dc," at the

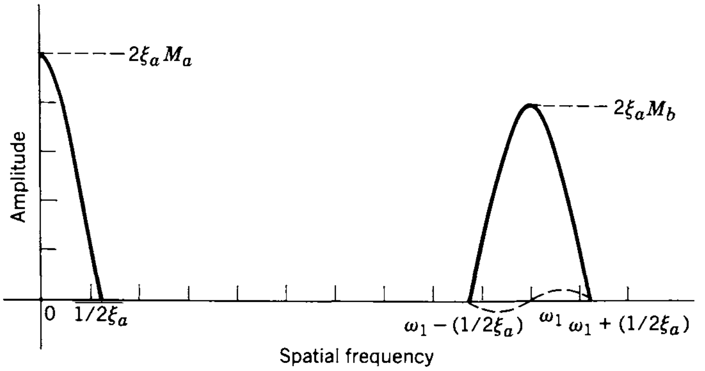


Figure 8.9. Frequency spectrum of a truncated sinusoidal distribution (minor lobes omitted).

origin. Because the grating has to be finite, a distribution of appreciable width will always occur around the frequency ω_1; and an exacting task in the laboratory is to measure both the height of the peak and the frequency at which it occurs. These measurements obviously become easier as the lobe is narrowed. Where to make the compromise between manageable grating length and a narrow distribution of frequencies involves a number of factors including measurement judgment.

The example in Fig. 8.9 shows the relative widths of the frequency distributions about $\omega = 0$ and about $\omega = \omega_1$ when the grating width contains eight complete cycles of the spatial frequency. The maximum of the peak centered at ω_1 is never greater than the maximum of the peak in the other distribution because

$$M_b \leqq M_a, \tag{8-16}$$

as discussed in connection with Eq. (2-3).

Parts of the spectrum are omitted in Fig. 8.9. A mathematically complete spectrum would include negative frequencies in a distribution about $\omega = -\omega_1$ to form a "mirror image" of the positive frequencies shown in the figure. Negative frequencies are not plotted because they have no physical significance. Also, that part of the spectrum distributed about $\omega = 0$ would not vanish at $1/(2\xi_a)$ as the figure suggests; there is a succession of lobes of decreasing height going on out to higher and higher frequencies, and these higher frequencies must be considered until the lobe maxima begin to have negligible values. Each lobe of Fig. 8.8 contains a continuum of spatial frequencies. The phase of spatial frequencies in alternate lobes changes by π radians; or, alternatively, the amplitude of every frequency in even-numbered lobes can be assigned a negative sign. Thus we must take into account both the wide $m_1(\omega)$ segment of the spectrum centered at $\omega = 0$ and another wide segment, the $m_2(+\omega_1)$ segment, centered at $\omega = \omega_1$; the two segments overlap.

It is of interest to consider the spatial frequencies in the two lobes of the $m_1(\omega)$ segment on either side of ω_1, one lobe just above in frequency and the other just below. (Since the value of ω_1 in this example happens to be an integral multiple of $1/(2\xi_a)$, the $m_1(\omega)$ segment is zero at ω_1.) These lobes are illustrated in Fig. 8.9 by the broken line curve. (The maxima are not drawn to scale.) The frequency ω_1 is at the eighth zero of the $m_1(\omega)$ segment. Just below ω_1 and just above ω_1 the amplitude maxima are -0.042 and $+0.037$, respectively, times the amplitude at $\omega = 0$. By doubling the number of cycles in the grating width to 16 cycles, the ratio of the larger adjacent lobe maximum to the $\omega = 0$ amplitude would be reduced to -0.021. When the spatial frequencies in the $m_1(\omega)$ segment are added algebraically at each frequency to the spatial frequencies of the $m_2(+\omega_1)$ segment, the resulting amplitudes just below ω_1 will be reduced and those just above will be augmented. These effects shift the frequency of the measured peak to a slightly higher value than ω_1.

Because a wider grating decreases the amount of significant overlap of the two distributions, the potential error in measuring the frequency of a test grating image is reduced.

Accurate measurement of spatial frequency in the image of a sinusoidal grating is often critical. Our discussions of frequency, beginning in Chapter 3, deal largely with the reduced coordinate system in which the optical magnification is assumed to be unity. In the real-space coordinates of optical measurements, the magnification may not be known to the required accuracy of the overall measurement, in which instance the spatial frequency in the image must be carefully measured.

CHANGING SPATIAL FREQUENCY

The practical laboratory problem of how to vary spatial frequency has been solved in a number of different ways. In a typical scheme, a succession of gratings is applied to provide as many different frequencies as the MTF curve requires. One frequently used method is indicated in Fig. 8.10. The gratings are on the cylindrical surface of a rotatable drum to facilitate the successive presentation of the different gratings. A quite different way to employ a grating drum is to put a continuous grating (of constant frequency or of a succession of different frequencies) on the cylindrical surface by making the lines parallel to the drum axis. Then rotation of the drum provides the scanning motion across a stationary slit in the image surface.

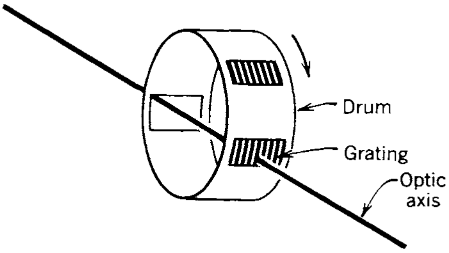


Figure 8.10. Schematic showing several gratings mounted on a drum for rapid change of spatial frequency.

Rosenhauer and Rosenbruch [2] employed the second drum technique to get the oscillogram pattern shown in Fig. 8.11. As the drum was rotated at a constant angular velocity, a photomultiplier tube behind the slit picked up the light from the grating and produced the pattern on an oscilloscope, which was photographed for the figure. With a number of stipulations, the envelope of the oscillogram is the MTF curve. Among the requirements of this method are:

1. The photomultiplier–amplifier–oscilloscope system must be linear (output proportional to input).
2. The contrast in the successive gratings must be constant.
3. The frequency and the number of cycles in each successive grating must be so chosen as to make linear the spatial frequency scale of the envelope.

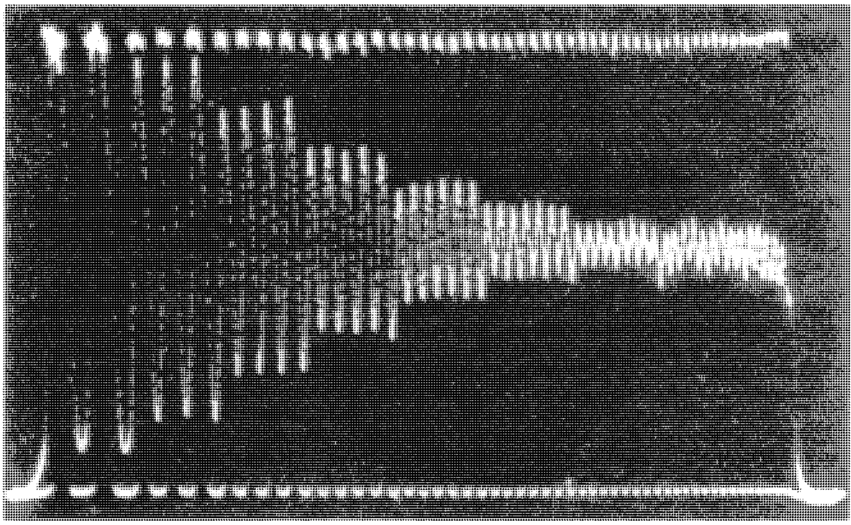


Figure 8.11. Pattern produced by a rotating drum-mounted grating having a succession of different frequencies with lines parallel to the drum axis [2].

Creating a grating with a sinusoidal transmittance has photographic problems. Sinusoidal variation of flux density in a light field can be accomplished by a number of different methods; but when such a distribution is to be recorded on film, nonlinear photographic response limits the accuracy of the sinusoidal variation in the processed film. Several workers [14–17] have kept distortion to less than 1% by carefully controlling the choice of film, exposure, and the photographic processing procedure. To achieve linearity, a common objective in the entire process is to make the photographic *gamma* constant and near unity—no mean feat.

When only one qualified grating is available for an MTF measurement, some means are needed for varying the *effective* frequency of the grating. One technique is to employ an auxiliary optical system with magnification either greater or less than unity to form an image of the grating, which then functions as the object for the test lens in the measurement setup. As indicated earlier, the MTF of the auxiliary optics must be taken into account in arriving at the MTF for the test lens.

A nearly sinusoidal distribution with continuously variable spatial frequency can be produced by crossing two square-wave gratings of relatively high spatial frequency. The spatial frequency of the resulting Moiré pattern depends upon the angle of crossing. One mechanism for this effect is to put the gratings on two disks rotating in opposite directions about the optic axis; the Moiré frequency then varies continuously as the disks rotate.

THE AREA GRATING

To avoid the photographic linearity problem, described earlier, in making a film grating with a sinusoidal transmittance as a function of the coordinate ξ, the grating area covered by the slit can be made a combination of areas having either maximum transmittance (ideally unity) or minimum transmittance (ideally zero). In other words, the grating is made up of areas that are either perfectly (nearly) transparent or completely opaque—nothing between these two extremes. We can analyze this approach by returning to the grasshopper example of Chapter 2. The grating pattern of Fig. 8.12*b* is analogous to the picket fence, and Eq. (2-1) can be rewritten, with a change in notation, as

$$\begin{aligned}
\tau(\xi) = 1 \quad & \text{when} \quad (-\xi_1/2) \leqq \xi \leqq (+\,\xi_1/2) \\
& \text{and} \\
& \qquad (2n + 3/2)\xi_1 \leqq |\xi| \leqq (2n + 5/2)\xi_1, \\
\tau(\xi) = 0 \quad & \text{when} \quad (2n + 1/2)\xi_1 \leqq |\xi| \leqq (2n + 3/2)\xi_1, \\
n = 0, 1, 2, 3, 4, \ldots. & \qquad\qquad (8\text{-}17)
\end{aligned}$$

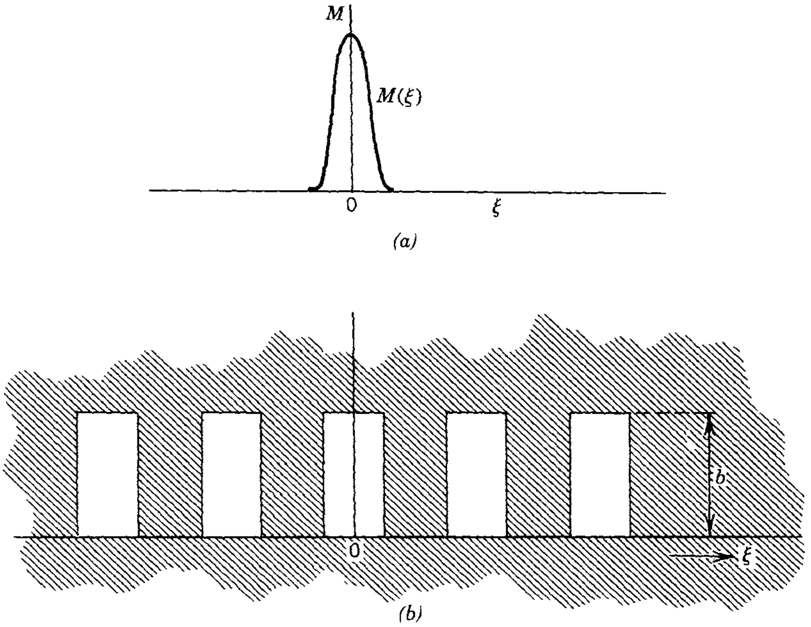


Figure 8.12. Scanning a square wave grating with a slit: (*a*) line spread function of the slit; (*b*) square wave grating.

This grating can be considered an array of parallel transparent bars as in Fig. 8.12*b*. The grating is assumed to be an image receiver whose surface is in the image plane, and a slit is the test object. The test lens has the line spread function $M(\xi)$ shown in Fig. 8.12*a*, which, by definition, is a plot of the incidance in the slit image on the grating. We assume that the slit image falls on the grating so that the long dimension of the image is parallel to the long sides of the bars and that the image is at least as long as from bottom to top of a bar, b. The scale along the ξ-axis for the line spread function $M(\xi)$ is the same as that for the grating function $\tau(\xi)$. The light flux in the slit image that strikes the transparent part of the grating passes through, and the remainder is rejected; so the amount that passes through depends on the position of the image relative to the grating. In the measurement procedure, the image scans the grating by motion of either the slit or the grating. The relative movement of the slit is parallel to the ξ coordinate axis. As the slit image is slid along the grating, the spread function at each point along the way is expressed mathematically by the shifted spread function; when its center is located at a point ξ_0, the spread function is given by

$$M'(\xi_0) = M(\xi_0 - \xi). \tag{8-18}$$

At a point ξ along the axis, a slice out of the grating can be expressed as a rectangular elemental area $b\ d\xi$. Because of the orientation of the slit image, the incidance is uniform over the elementary rectangle, and the total flux reaching the slice is $bM(\xi_0 - \xi)\ d\xi$. When this quantity is multiplied by the grating transmittance function $\tau(\xi)$, the resulting expression is the light flux $d\Phi$ that passes through the elemental area of the grating:

$$d\Phi = b\tau(\xi)M(\xi_0 - \xi)\ d\xi. \tag{8-19}$$

The total flux transmitted through the grating as a function of ξ_0 is the integral of Eq. (8-19):

$$\Phi(\xi_0) = b\int_{-\infty}^{+\infty} \tau(\xi)M(\xi_0 - \xi)\ d\xi. \tag{8-20}$$

Equation (8-20) is recognized as the correlation integral of the grating transmittance function with the line spread function.

To help visualize the process represented by Eq. (8-20), a transparency will be hypothetically substituted for the lens spread function in the image plane. The spread function $M(\xi)$, as shown in Fig. 8.13*a*, is plotted on a transparency with the height at $\xi = 0$ equal to the height b of the grating bars. To convert the ordinate scale of $M(\xi)$ to the same units used for b, a constant C_1 is applied:

$$C_1M_\mathrm{m} = b \quad \text{or} \quad C_1 = b/M_\mathrm{m}. \tag{8-21}$$

where M_m is the peak value of the $M(\xi)$ occurring in Eqs. (8-18)–(8-20). The area between the curve and the ξ-axis is cut out or otherwise made perfectly transparent (unity transmittance) and the rest of the transparency is opaque (zero transmittance) as represented in Fig. 8.13*a*. The transparency, shown in Fig. 8.13*a*, is placed on the grating with the ξ-axes coincident. Then, as the transparency is slid in the ξ direction with the origin of the transparency at ξ_0 as shown in Fig. 13*b*, the total flux Φ_2 passing through the two films is

$$\Phi_2(\xi_0) = b/M_\mathrm{m}\int_{-\infty}^{+\infty} \tau(\xi)M(\xi_0 - \xi)\ d\xi. \tag{8-22}$$

Comparison of Eq. (8-22) with Eq. (8-20) shows them to be identical except for the scale constant $1/M_\mathrm{m}$. Therefore, the hypothetical transparency–grating combination is a legitimate model for visualizing the projection of a spread

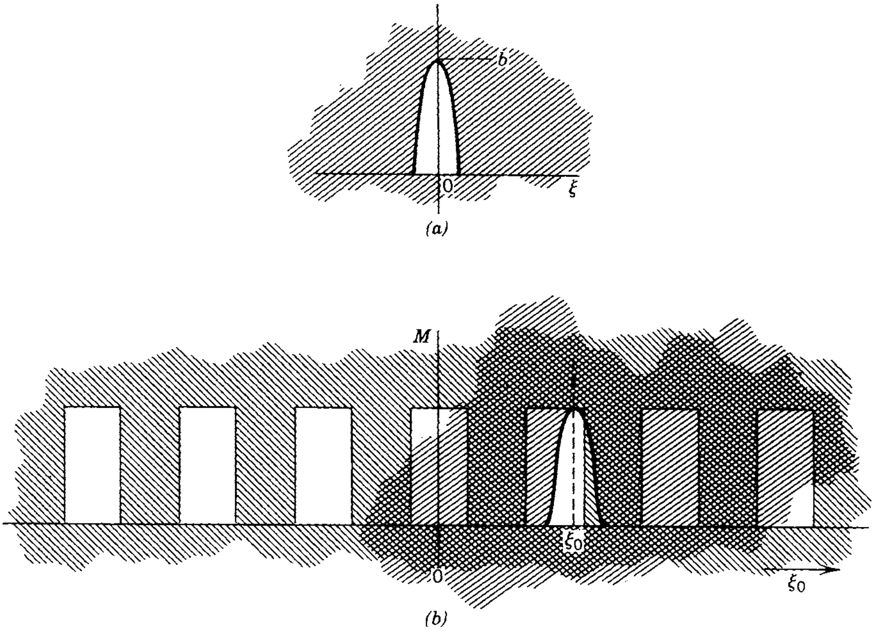


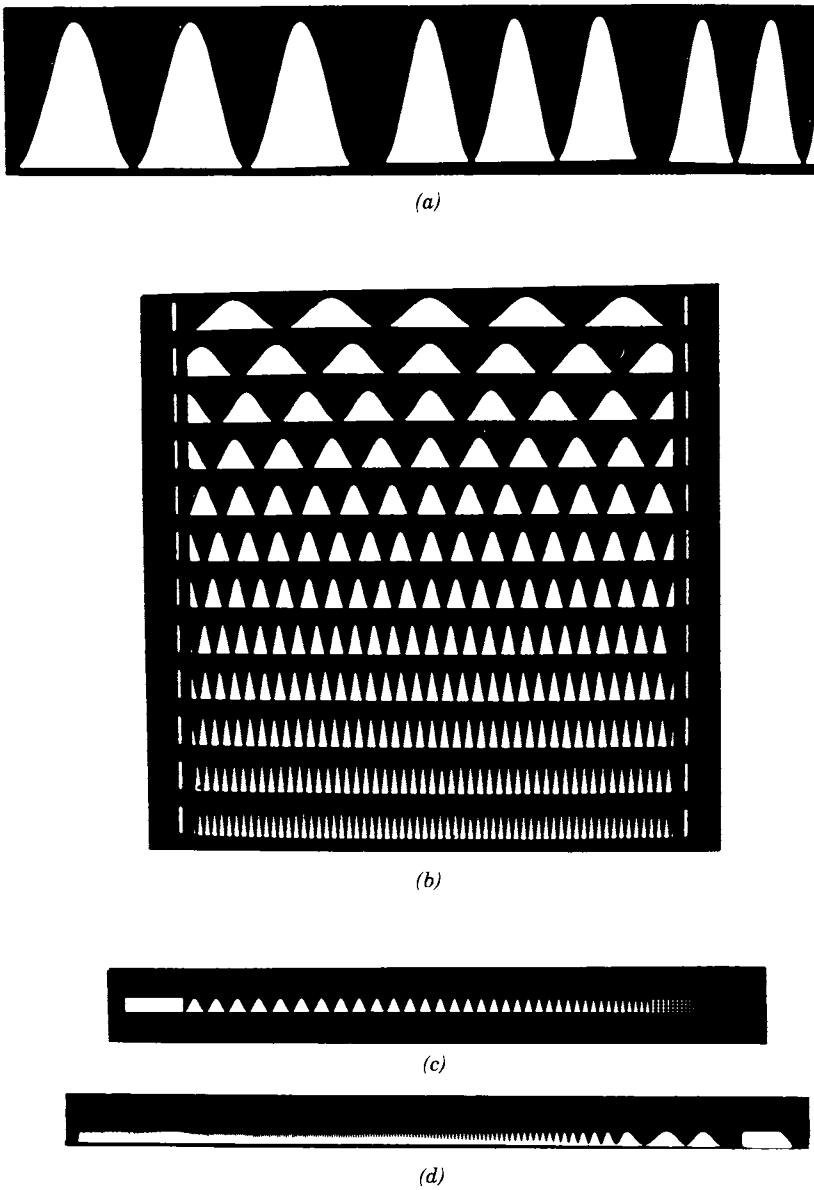
Figure 8.13. Superposition of two transparencies: (*a*) spread function transparency; (*b*) superposition of spread function transparency on grating transparency.

function on a grating. Although we used a specific function (the square wave) for $M(\xi)$ so that simple illustrative figures could be drawn, the development for any other $M(\xi)$ function would have been the same as for our square wave example. Particularly, the function could have been sinusoidal to be in line with our previous grating discussions. Examples of actual sinusoidal area gratings are shown in Fig. 8.14.

Though the application of the area grating is a current art in MTF measurements, it is reminiscent of area sound track techniques in the early "talking" moving pictures.

The measurement of $\Phi(\xi_0)$, as represented by Eq. (8-20), constitutes the raw data for evaluating the spectrum $m(\omega)$, the lens MTF, which is the Fourier transform of the lens spread function $M(\xi)$. We resort to the convolution theorem (see Appendix B, Eq. (B-32)) to extract the MTF from Eq. (8-20). If $\varphi(\omega)$ is the transform of $\Phi(\xi_0)$, then

$$\varphi(\omega) = b[\tau_1(\omega)][m(\omega)]. \qquad (8\text{-}23)$$



(a)

(b)

(c)

(d)

Figure 8.14. Examples of sinusoidal area gratings (from Murata [1]).

Solving for the spectrum $m(\omega)$, which is the MTF, yields

$$m(\omega) = (1/b)[\varphi(\omega)]/[\tau_1(\omega)]. \tag{8-24}$$

In this formula for $m(\omega)$, the functions $\varphi(\omega)$ and $\tau_1(\omega)$ must still be evaluated by taking the Fourier transforms of $\Phi(\xi_0)$ and $\tau(\xi)$, respectively. Since $\Phi(\xi_0)$ is recorded by a scanning process, ξ_0 must also be evaluated from $\xi_0 = vt$, where v is the velocity of the scan and t is time, the actual independent variable in the scanning process. In practice, rather than recording Φ continuously, this quantity is usually observed as discrete values at regular time intervals, t_0, t_1, t_2, etc.

In operating on the grating function $\tau(\xi)$, its finite nature has to be recognized as indicated earlier in the analysis of the finite sinusoidal grating distribution function $m(\omega)$. A similar analysis of the bar pattern of Eq. (8-17) produces the result:

$$f(\omega) = [\sin(2\pi N\xi_1\omega)]/[2\pi\omega\cos(\pi\xi_1\omega)], \tag{8-25}$$

where N is the total number of bars in the grating.

EFFECT OF SLIT WIDTH

So far, in the mathematical formulation of the slit–lens–grating combination, the lens and grating functions have provided for physical limitations in the lens and grating, but the slit has been treated as an ideal line having negligible width. Actually, the slit must be wide enough to allow the passage of enough light flux for relatively noise-free measurement. Just as we found that a lens produces a spread function image for an ideal line object and that the finite nature of a grating complicates its function, we now find that the appreciable width of a slit also has to be involved in reducing the data for an MTF measurement.

For the following development, the slit will be regarded as the image receiver as shown schematically in Fig. 8.7*a*. Setting up an expression to represent the light flux passing through the combination of the image and scanning slit parallels the writing of Eqs. (8-20) and (8-22), which are usually referred to as *convolution integrals*; in this instance, the integral is

$$\Phi_3(\xi_0) = \int_{-\infty}^{+\infty} \tau_3(\xi)M_3(\xi_0 - \xi)\,d\xi, \tag{8-26}$$

where $\Phi_3(\xi_0)$ is the light flux "passing through" the image–slit combination,

$\tau_3(\xi)$ is the rectangular slit function, and $M_3(\xi)$ is the grating image function. For simplicity, the constant preceding the integral has been assumed unity. By again applying the convolution theorem, as we did to arrive at Eq. (8-23), Eq. (8-26) leads to

$$\varphi_3(\omega) = [\tau_3(\omega)][m_3(\omega)], \tag{8-27}$$

where $\varphi_3(\omega)$, $\tau_3(\omega)$, and $m_3(\omega)$ are the Fourier transforms of $\Phi_3(\xi_0)$, $\tau_3(\xi)$, and $M_3(\xi)$, respectively. Solving for $m_3(\omega)$, which is the MTF, gives

$$m_3(\omega) = [\varphi_3(\omega)]/[\tau_3(\omega)]. \tag{8-28}$$

Although this formula for the MTF appears to give the desired result, certain peculiarities of the laboratory optical setup have to be explored before we can be sure that the calculated $m_3(\omega)$ is indeed the true MTF for the lens under test. Somewhat like the mechanical tolerances discussed earlier in this chapter, certain design tolerances either have to be maintained or special calculation procedures, often complicated, have to be followed to compensate for departures from ideal design. For example, a grating may have a sufficient number of cycles, though finite, to be treated as a continuous grating, but should it be necessary to limit the number of cycles below some tolerable limit, calculation of the MTF from observed data must take into account the finite nature of the grating. Also, as another example, should the width of a slit exceed a limit determined by the required accuracy of the MTF measurement, it can no longer be treated mathematically as a line.

A previous section, "Effect of Finite Grating Length," has already treated the grating problem at some length. The actual spectrum produced by the finite grating is a combination of sinc functions rather than the single frequency resulting from an ideal sine variation; so reduction of the measured data requires a corresponding correction when the difference is significant. Also, the total value of the contrast peak near ω_1 (Fig. 8.9) actually occurs at a frequency slightly different from ω_1 because of the contribution of the sinc function at the origin. If not taken into account, this could contribute to errors both in frequency and in contrast values.

When the Fourier transform of the rectangular slit function $\tau_3(\xi)$ is taken to get the spectrum $\tau_3(\omega)$ of the slit, a sinc function results:

$$\tau_3(\omega) = [\sin(2\pi\xi_s\omega)]/(2\pi\xi_s\omega), \tag{8-29}$$

where $2\xi_s$ is the slit width. The first zero of this function occurs at the lowest value of ω where the sine in the numerator is zero and the denominator has a nonzero value:

$$\begin{aligned} 2\pi\xi_s\omega_s &= \pi, \\ 2\xi_s &= 1/\omega_s. \end{aligned} \tag{8-30}$$

Ideally, as the slit scans spatial frequencies of equal contrast, we would prefer that the signal at the detector be of constant amplitude. However, Eqs. (8-29) and (8-30) indicate that as the frequency increases, the slit attenuates the signal more and more until at ω_s the slit is as wide as a spatial period and no net signal is received as the pattern is being scanned. Any observation through the slit has to be corrected for this attenuation. As in any other measuring system, attenuation of the signal brings the measured quantity closer to the noise level, which affects the accuracy of the observation.

To get some idea of how wide a slit can be tolerated, the attenuation curve $m_3(\omega)$ can be compared with the "perfect" (diffraction-limited) MTF curve. In particular, the relative values of the first zero frequency ω_s and the MTF cutoff frequency ω_c are of interest. In Chapter 5, Eq. (5-70), the cutoff frequency is shown to be

$$\omega_c = 2 \text{ N.A.}/\lambda, \tag{8-31}$$

where N.A. is the numerical aperture and λ is the nominal wavelength of the illuminating light. If we assume a numerical aperture of 0.25, which corresponds approximately to an $f/2$ optical system, and a wavelength of 500 nm, ω_c is found to be 1000 lines/mm. Comparison of the MTF ("perfect") and the sinc curves in Fig. 8.15 confirms the anticipated requirement that ω_s must be greater than ω_c; so $\omega_s > 1000$ lines/mm. From Eq. (8-30), where $2\xi_s$ is the width of the slit,

$$\begin{aligned} &1/(2\xi_s) > 1000, \\ &2\xi_s < 1/1000 \text{ mm} \quad \text{or} \quad 1\ \mu\text{m}. \end{aligned} \tag{8-32}$$

So, for the assumed example, the slit should be as much smaller than a micrometer as the detector minimum signal requirement allows.

SQUARE WAVE GRATINGS

Instead of a sinusoidal grating, a square wave grating is sometimes used as the test object in setups represented by the schematic of Fig. 8.7. Besides a fun-

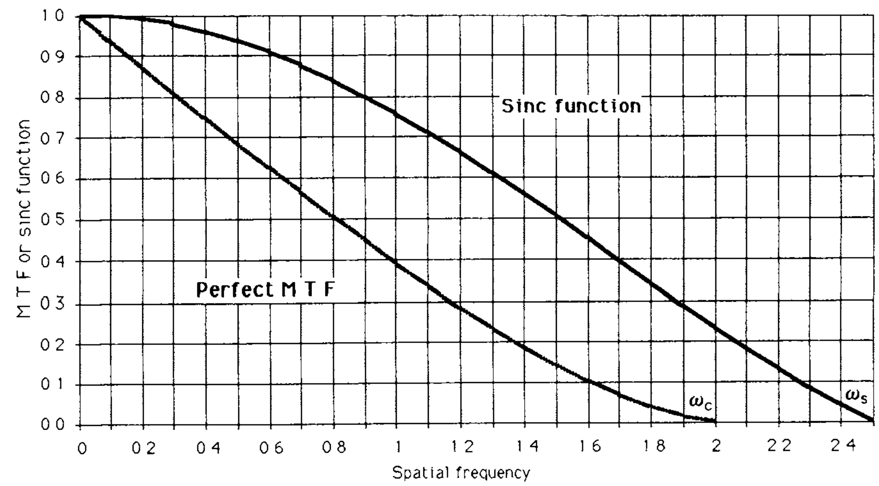


Figure 8.15. Comparison of the "perfect" MTF and the sinc function.

damental frequency, such a grating produces an infinite series of harmonics as indicated by Eq. (2-4), which must be taken into account either by calculation or by electronic filtering.

Filtering becomes quite complicated when a sequence of square wave gratings having different frequencies is put on a rotating drum. One way of avoiding this problem is to substitute a rotating sector disk (Fig. 8.16) for the conventional grating. When the disk image is scanned by a small aperture at a fixed radius from the disk center, both the optical spatial fundamental and the electrical time fundamental are constant as the disk rotates at constant speed. However, as the aperture is moved toward the center of the rotating disk, the spatial frequency increases, but the electrical frequency remains constant, making electronic filtering a relatively easy task.

Another scheme places the square wave gratings of different frequencies side by side rather than in sequence on a rotating drum. Then the spatial frequencies are selected by moving the circular aperture parallel to the axis of rotation. Constant electrical frequency, as in the disk technique, can be realized by making the drum rotational speed inversely proportional to the spatial frequency being observed.

As in the measuring techniques employing sinusoidal gratings, the signal-to-noise ratio for square wave gratings can be greatly improved whenever one can use the relatively large transmission area of even a narrow slit instead of a small circular aperture. Also, when the straight-line boundaries of a square wave grating are scanned by a round aperture, the process produces its own peculiar harmonic structure different from that of a slit.

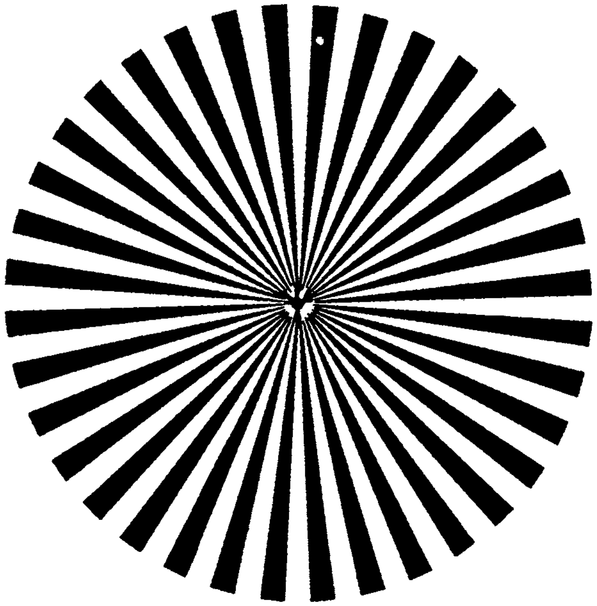


Figure 8.16. Square wave grating in the form of a sector disk.

INDIRECT METHODS

Although considerable calculation is involved in converting raw measurement data into MTF results in most of the methods already discussed, they are generally referred to as *direct* methods. The OTF or MTF can also be evaluated by observing the images of objects that are less suggestive of the sinusoidal functions that form the basis of the OTF. Fourier analysis is characteristically employed to reduce the data to useful form. This second group is loosely referred to as *indirect* methods.

Simple examples of objects commonly employed in indirect methods are slits and edges. The frequency spectrum of a slit in terms of its width has already been discussed in connection with scanning periodic images. The knife edge, also called a *half plane*, consists of what may be regarded as a special case of a square wave grating—a total of one cycle in all space, the ultimate in low spatial frequencies. In the usual setup, the edge or line separating the two half planes is arranged to intersect the optic axis. The bright-to-dark contrast is made as high as practical, typically about 1000 to 1. The image of an edge is a distribution called an *edge trace*, which is represented by the symbol $M_e(\xi)$. The derivative of the edge trace is the *line spread function* $M_\ell(\xi)$:

$$M_\ell(\xi) = d[M_e(\xi)]/d\xi. \tag{8-33}$$

Inversely, as shown by Eq. (2-29) in different notation, the edge trace is the integral of the spread function:

$$M_e(\xi_e) = \int_{-\infty}^{\xi_e} M_\ell(\xi)\, d\xi. \tag{8-34}$$

The Fourier transform of the line spread function turns out to be the optical transfer function.

In the laboratory, the image of an incoherently illuminated slit or edge is scanned by a second slit. The incidance at the image plane is converted to a proportional electrical signal by a detector. A rule of thumb based on experience is that the width of the scanning slit should be no greater than a third of the spatial period of the test lens cutoff frequency.

Indirect methods have the advantage of avoiding the complicated test objects of direct methods but suffer a disadvantage in requiring complicated electronic systems to perform Fourier analysis. They share the complication of having to take into account the spectrum of every slit that is part of the laboratory apparatus.

The image of a slit is the convolution of the ideal line spread function $M_s(\xi')$ with a rectangular function $M_{\mathcal{R}}(\xi)$ involving the slit width. This rectangular function has already been defined by Eq. (8-11) to take into account the "width" of a finite number of cycles in a grating. In reduced coordinates, the light flux $\Phi_s(u)$ in the slit image is

$$\Phi_s(u) = \int_{-\infty}^{+\infty} M_{\mathcal{R}1}(u')M_s(u - u')\, du'. \tag{8-35}$$

Reduced coordinates, as used in Eq. (8-35), are discussed at some length in Chapter 3 where the key defining equations are Eqs. (3-39), (3-40), (3-43), and (3-45). If the conditions for Eq. (3-35) are met, no distinction need be made between the scale of variables in object and image spaces (magnification, for instance, is normalized to unity), so primes are no longer reserved to distinguish between object and image variables. (The primes in Eq. (8-35), for example, identify the conventional variable of integration in the convolution expression, and do *not*, of themselves, label the associated symbols as coordinates in image space. Whether a given variable is in object or image space has to be gleaned from the context.)

When the image $\Phi_s(u)$ of the first slit is scanned by a second slit, the resulting flux $\Phi_m(u)$ is found by the convolution of $\Phi_s(u)$ with a rectangular function $M_{\mathcal{R}2}$ representing the second (scanning) slit:

$$\Phi_{\mathrm{m}}(u) = \int_{-\infty}^{+\infty} M_{\mathcal{R}2}(u')\Phi_{\mathrm{s}}(u - u')\, du'. \tag{8-36}$$

Successive convolutions, as indicated in Eqs. (8-35) and (8-36), are discussed in Appendix B leading up to Eq. (B-46). In the notation of the present chapter, the successive convolutions are represented by

$$\Phi_{\mathrm{m}}(u) = M_{\mathcal{R}2}(u) * M_{\mathcal{R}1}(u) * M_{\mathrm{s}}(u). \tag{8-37}$$

Then, by applying the convolution theorem to each successive convolution in Eq. (8-37), the following product of a three-term multiplication results:

$$m_{\mathrm{m}}(s) = m_{\mathcal{R}2}(s) m_{\mathcal{R}1}(s) m_{\mathrm{s}}(s), \tag{8-38}$$

or

$$m_{\mathrm{s}}(s) = m_{\mathrm{m}}(s) / [m_{\mathcal{R}2}(s) m_{\mathcal{R}1}(s)], \tag{8-39}$$

where $m_{\mathrm{m}}(s)$ is the observed spectrum and $m_{\mathrm{s}}(s)$ is the spectrum that would have been produced by the lens for an ideal line source. Since an ideal line source (ideal slit) can be represented by a Dirac delta function, whose spectrum (Fourier transform) consists of all frequencies of equal amplitude and zero phase shift (see Appendix B, "The Delta Function"), $m_{\mathrm{s}}(s)$ is the MTF of the lens when normalized to unity at zero frequency.

The spectra in the above equations are expressed in terms of the reduced spatial frequency s. Actual measurements in the laboratory are generally made in terms of the real space coordinate ξ; the spectrum m_{c}, the Fourier transform of the observed $\Phi_{\mathrm{c}}(\xi)$ through the scanning slit, then is in terms of the real spatial frequency ω. The reader is referred to Chapter 3 for conversion relations between real and reduced quantities to evaluate, for instance, $m_{\mathrm{m}}(s)$ from $m_{\mathrm{c}}(\omega)$ or vice versa. In real-space coordinates, Eq. (8-39) would be written

$$\mathrm{MTF} = m_{\mathrm{s}}(\omega) / [m_{\mathcal{R}2}(\omega) m_{\mathcal{R}1}(\omega)]. \tag{8-40}$$

INTERFEROMETRIC METHODS

An implied condition of MTF measurement in all of the methods discussed thus far is that the illuminating light beam be incoherent. When measuring high-quality optics such as microscope objectives, or when these optics are used as an auxiliary part of the test setup, it is difficult to ensure strict incoherence. One way to avoid this problem is to use an interferometric method, in which coherence is required.

The laser provides a stable, powerful, and monochromatic light source for interferometric measurements. Its high intensity enables a correspondingly high signal-to-noise ratio. A disadvantage is that a polychromatic OTF is not directly obtained but requires a synthesis of measurements made with lasers of different wavelengths. One objection often heard against interferometric measurements is that they are time-consuming and relatively expensive; however, if corresponding OTF results are to be obtained from any other kind of setup, the differences in time and expense are not found significant.

As emphasized in previous chapters, the fundamental measure of optical performance is the pupil function, from which the OTF can be calculated by an autocorrelation procedure. In turn, the pupil function can be calculated from ray-trace data. The least complicated, experimentally, of the interferometric methods is measurement of the wave-front shape in the exit pupil of the lens being tested; the OTF is subsequently calculated by an autocorrelation procedure. This method is discussed further in Chapter 9 in company with other topics involving extensive calculation. Other methods discussed in the following sections of this chapter are generally characterized by measurement of the total light flux in the overlapping area of sheared (that is, laterally displaced) wave fronts. Among the interferometers discussed in this chapter, the most common types are each some modification of the Michaelson interferometer. We refer to all of these as forms of the Michaelson interferometer even though certain instruments may be closer in design to what is known as the "Twyman–Green" interferometer.

THE INTERFEROMETER

A schematic of a Michaelson interferometer is shown in Fig. 8.17. A laser beam is passed through a beam expander to provide plane wave fronts incident from the left at Q. The incident wave train is represented by

$$\xi = \xi_0 \cos(2\pi\nu_0 t - \varphi), \tag{8-41}$$

where ξ_0 is the scalar quantity representing amplitude in the wave, ν_0 is the optical (time) frequency, and φ is the phase angle at Q. The wave is divided into two parts at the beamsplitter, a portion being reflected upward and the remainder transmitted to the right. The reflected portion is again reflected at the mirror M_1 so that it travels back down toward the beamsplitter. At the beamsplitter, the wave divides a second time, part being reflected toward the source

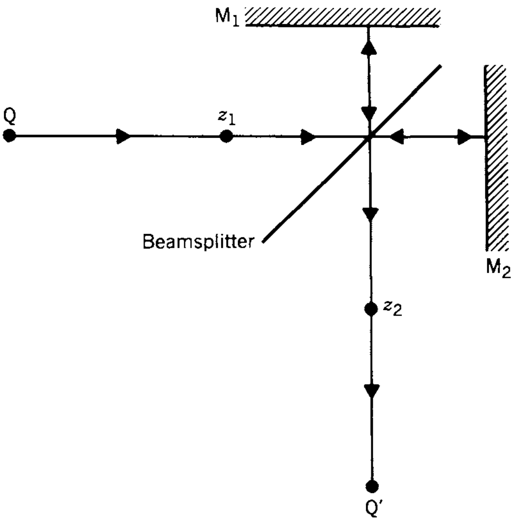


Figure 8.17. Schematic of a Michaelson interferometer.

and the rest transmitted downward toward Q'. Similarly, the portion transmitted at the first encounter with the beamsplitter reflects from mirror M_2, strikes the beamsplitter again, and a part reflects downward toward Q'. With perfect alignment of the beamsplitter and the mirrors, the two parts of the wave train are traveling downward toward Q' in exactly parallel directions.

If φ_2 is the phase angle of one of the waves at z_2, the other wave has a phase angle of $\varphi_2 - (2\pi\delta/\lambda)$ to take into account any difference δ in the two paths. If we assume that the amplitude of a wave is divided into two equal parts by the beamsplitter, the two waves at z_2 can be represented as

$$\begin{aligned} \xi_1 &= (\xi_0/4)\cos(2\pi\nu_0 t - \varphi_2), \\ \xi_2 &= (\xi_0/4)\cos[2\pi\nu_0 t - \varphi_2 + (2\pi\delta/\lambda)]. \end{aligned} \tag{8-42}$$

The resultant field quantity is

$$\xi = \xi_1 + \xi_2. \tag{8-43}$$

Adding two sinusoidal functions of the same frequency results in a sinusoidal function of that frequency as the sum with a phase angle somewhere between the phase angles of the two components. Because ξ_1 and ξ_2 have been assumed of equal amplitude, the resultant has the phase angle $[\varphi_2 - (\pi\delta/\lambda)]$, halfway between those of the components. After some trigonometric manipulation, the amplitude of the resultant is found to be a square root, $[2 + 2\cos(2\pi\delta/\lambda)]^{1/2}$, times the amplitude of a component. By a trigonometric half-angle identity, the square root becomes $2\cos(\pi\delta/\lambda)$; so we can write

$$\xi = (\xi_0/2)[\cos(\pi\delta/\lambda)][\cos(2\pi\nu_0 t - \varphi_2 + \pi\delta/\lambda)]. \tag{8-44}$$

The radiant flux density of the resultant wave, with a convenient choice of units, can be calculated by finding the average of ξ_2 over a large number of cycles:

$$\mathcal{W} = (\xi_0^2/4)[1 + \cos(2\pi\delta/\lambda)]. \tag{8-45}$$

From this expression, it is obvious that the flux density becomes zero for negative unity values of the cosine. This occurs when

$$2\pi\delta/\lambda = n\pi \quad \text{or} \quad \delta = n\lambda/2, \qquad \text{for } n = 1, 3, 5, 7, \ldots. \tag{8-46}$$

Maximum values of the flux density occur for positive unity values of the cosine:

$$2\pi\delta/\lambda = m\pi \quad \text{or} \quad \delta = m\lambda/2, \qquad \text{for } m = 0, 2, 4, 6, \ldots. \tag{8-47}$$

The effect of a mirror flaw on the performance of the interferometer can be studied by assuming a specific defect in the mirror surface and finding what it does to the combined wavefront. The specific results can be extended to flaws in general if the assumed defect can be thought of as a building block for more complicated flaws. A suitable assumption for this purpose is a dimple, a circularly symmetrical depression of z depth and having the following shape in the mirror surface:

$$z = p\lambda(1 - r), \qquad r \leqq 1, \tag{8-48}$$

where p is the depth at the center of the depression in wavelengths of light, λ is the wavelength of the light beam, and r is the distance, parallel to the mirror surface, from the center of the depression. Then the portion of the wave front at r has to travel the extra distance, because of the flaw, of twice the depth at that point:

$$2z = 2p\lambda(1 - r). \tag{8-49}$$

Thus, δ of Eqs. (8-46) and (8-47) becomes a function of z:

$$\delta(z) = \delta_0 + 2p\lambda(1 - r), \tag{8-50}$$

where δ_0 is the constant path difference for a flawless mirror surface.

When

$$\delta(z) = n\lambda/2, \tag{8-51}$$

a dark ring, usually referred to as a *fringe*, results because, according to Eq. (8-46), the flux density is zero. When

$$\delta(z) = m\lambda/2, \tag{8-52}$$

a bright ring results because, according to Eq. (8-47), the flux density is a maximum. The fringes, of course, are circular because of the assumed symmetry of the flaw. The central fringe is a circular disk, dark or bright depending on whether $n\lambda/2$ or $m\lambda/2$, respectively, applies.

Fringes in the described example are said to be formed at infinity because the interferometer beams are parallel. The fringes can be more easily observed by placing a converging lens in the beam near z_2 to form an image of the fringe pattern at Q′.

To explore the effects of moving mirrors M_1 and M_2, we assume all surfaces to be perfectly plane and the interferometer to be adjusted so that the wave fronts are perfectly parallel at z_2. The flux density $\mathcal{W}$, as given by Eq. (8-45), is uniform over a cross section of the beam. Moving either mirror along the optical axis, that is, perpendicular to its surface, varies δ and, therefore, the flux density. If the movement is a linear displacement, say of M_2, with time, the flux density will vary sinusoidally between a maximum of $\xi_0^2/2$ and zero. A condensing lens placed near z_2 in Fig. 8.17 produces the configuration of Fig. 8.18 and concentrates the flux so that a detector can efficiently convert the variations in light flux to an electrical sine wave.

When one of the fixed mirrors is tilted slightly to form a wedge between the two superposed wave fronts, the separation of the wave fronts is a linear function of the distance x perpendicular to z. Ideally this would produce "straight-line" fringes. For small angles, the distance δx between two successive bright fringes, as indicated in the schematic of Fig. 8.19, is

$$\delta x = \lambda/\alpha. \tag{8-53}$$

where α is the wedge angle between wave fronts in radians and λ is the wavelength of the light beam. Because of the double passage of the beam at the

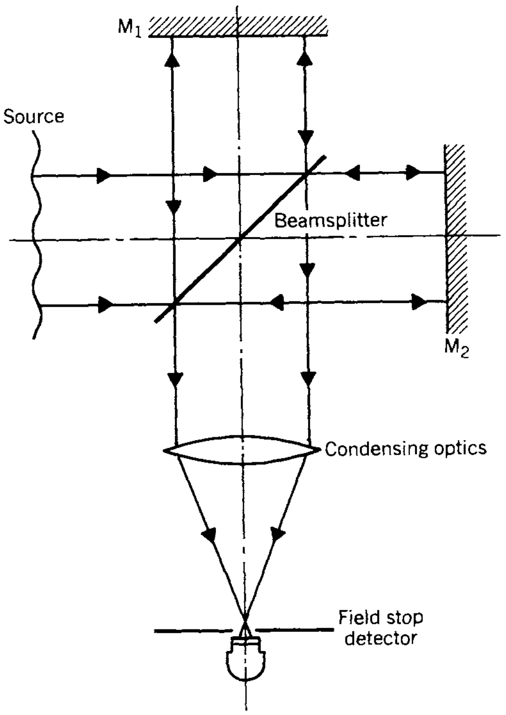


Figure 8.18. Application of condensing optics in a Michaelson interferometer.

mirror surface, the corresponding wedge angle between the mirror surface and its position for parallel wave fronts is $\alpha/2$.

In general, the existence of fringes indicates some kind of flaw in the interferometer setup that causes a variable path difference in the two beams. With m and n defined as in Eqs. (8-46) and (8-47), a fringe or wavelength variation of wavefront spacing occurs between two points 1 and 2 when either

$$m_2 = m_1 + 2 \quad \text{or} \quad n_2 = n_1 + 2. \tag{8-54}$$

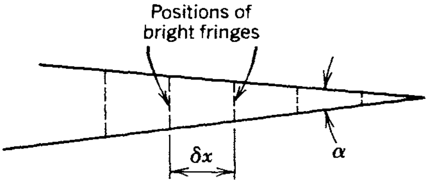


Figure 8.19. Production of fringes by a wedge between wave fronts.

A fringe pattern is thus a sort of contour map. Measurements on a photograph of a fringe pattern can provide accurate information about the wave-front shape of one beam relative to the other.

H. H. Hopkins [18] was the first to suggest applying the interferometer to measuring the OTF. He replaced the plane mirrors M_1 and M_2 of Fig. 8.18 with corner reflectors as illustrated in Fig. 8.20. Because this configuration shifts a ray laterally, an instrument so designed is called a *shearing interferometer*. In Fig. 8.17, each wave front is "inverted" upon reflection from each plane mirror including the beamsplitter; that is, when looking in the direction of propagation of each wave, one finds that the right edge of the incident wave becomes the left edge of the reflected wave. However, in Fig. 8.20 each wave front retains its right-left relations as it reverses direction at a corner reflector. In doing this, as indicated, a ray does not return upon itself but is displaced laterally ("sheared"). The amount of shear can be varied by a lateral displacement of the corner reflector. In Fig. 8.20, one ray is sheared a distance p by mirror M_1, and the other ray is sheared a distance m by mirror M_2. By experimentally tracing a few rays through the mirror system, one quickly discovers that the two wave fronts, except for special positioning of the corner reflectors, are only partially superposed; of course, fringes will be formed only in the superposed, or overlapping, region. If lens L is placed to transmit the incoming wave front to produce the distorted form W as shown in Fig. 8.20, integration of $\mathcal{W}$, Eq. (8-45), in the superposed region for selected values of p and m, provides data for plotting the MTF. The superposed region in the shearing interferometer is

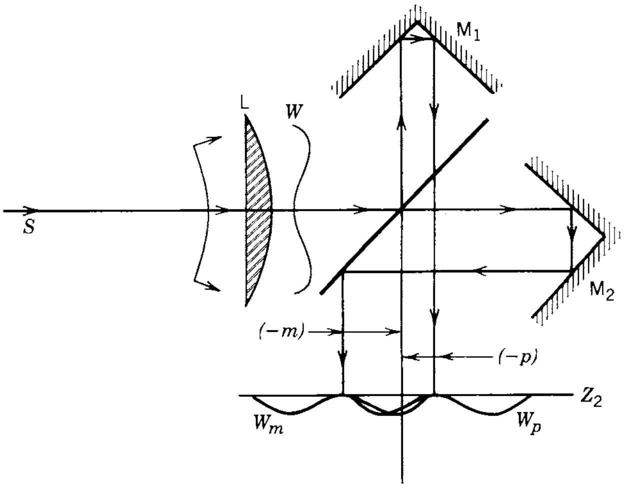


Figure 8.20. Schematic of a shearing interferometer with corner reflectors [18].

the physical analog of the mathematical overlapping area in the autocorrelation of the pupil function, Eq. (5-15).

The complex equations, with time factors suppressed, for the beams approaching the plane at z_2 in Fig. 8.20 are

$$\hat{f}_1(x, y) = U_1(x, y) \exp\left\{ikW[(x - p), y]\right\}, \tag{8-55}$$

and

$$\hat{f}_2(x, y) = U_2(x, y) \exp\left\{ikW[(x + m), y] + ik\delta\right\}, \tag{8-56}$$

where $U_1(x, y)$ and $U_2(x, y)$ are the wave amplitudes, $W(x, y)$ is the displacement difference between the actual wave front and a hypothetical reference wave front (in this instance, a plane), and δ is the path difference between the superposed wave fronts. Certain assumptions can greatly simplify the mathematics of the interferometer. Although beamsplitters do not generally divide a beam perfectly into two parts of equal amplitude, the assumption that $U_1 = U_2 = 1$ introduces negligible error in practical calculations of MTF. To bring the interferometer mathematics in line with the reduced coordinates introduced in Chapter 3 and the autocorrelation of the pupil function in Chapter 5, it is convenient to assume that each wave front has a circular boundary of radius r beyond which the amplitude is zero. Further, p is defined as the displacement of the center of one wave relative to the center of the other, which is at the origin, that is, $m = 0$ in Fig. 8.20. Then the reduced spatial frequency is defined as $s = p/r$. When these modifications are made in Eqs. (8-55) and (8-56), the total of the two wave fronts in the overlapping region is

$$\hat{f}_1 + \hat{f}_2 = \exp\left\{ikW[(x - s), y]\right\} + \exp[ikW(x, y) + ik\delta]. \tag{8-57}$$

Then the flux density is

$$\begin{aligned}(\text{Amplitude})^2 &= (\hat{f}_1 + \hat{f}_2)(\hat{f}_1^* + \hat{f}_2^*)\\ &= 2 + 2 \cos k\left\{W[(x - s), y] - W(x, y) - \delta\right\}.\end{aligned} \tag{8-58}$$

The total light flux $\Phi(s)$ in the superposed region $\mathcal{A}$ is

$$\Phi(s) = \iint_{\mathcal{A}} \left\langle 2 + 2 \cos k\left\{W[(x - s), y)] - W(x, y) - \delta\right\}\right\rangle dx\, dy. \tag{8-59}$$

For a later application of this light flux formula, it is desirable to change its form by applying trigonometric identities as follows:

$$\Phi(s) = 2 \iint_{\mathcal{A}} dx\, dy + 2 \cos k\delta \iint_{\mathcal{A}} \cos k\left\{ W[(x - s), y] - W(x, y) \right\}$$
$$\cdot\, dx\, dy + 2 \sin k\delta \iint_{\mathcal{A}} \sin k\left\{ W[(x - s), y] - W(x, y) \right\} dx\, dy. \tag{8-60}$$

AN INTERFEROMETRIC MEASURING EQUIPMENT

Although Baker [19] describes an "interference photometer," Fig. 8.21, as one that gives the frequency response of a lens directly for almost any condition of illumination and independent of the type of object, our discussion here is confined to a coherent beam in this type of instrument because lasers have become readily available.

The apparatus of Baker's report is essentially an analog computer incorporating the lens under test. The setup is based on an interferometer of the Michaelson type with corner reflectors as suggested by Hopkins and illustrated in Fig. 8.20. These optical components are on the right side of the diagram in Fig. 8.21. The rest of the measurement setup consists of a flicker photometer in which one beam passes through the lens L_1 under test, and the other beam passes through a high quality collimating lens L_2. As suggested by the diagram, the test lens L_1 creates an aberrant wave front, and the collimating lens L_2 provides a plane (reference) wave front.

Baker's apparatus has a sodium vapor lamp as a light source. Its image is focused on the slit S_0, the light from which reaches the outside faces of the roof reflector R after the beam passes through a collimating lens and the polarizing unit Pol_1-WP-Pol_2. The Wollaston prism WP splits the light into two slightly divergent beams polarized at right angles to each other. By the time they reach the roof reflector, the two beams are separated enough so that one beam reflects upward and the other downward in Fig. 8.21. The next reflector in each path causes the beams to become parallel, well separated, each focused by a lens (C_1 or C_2) on a slit. The function of the field lenses F_1 and F_2 is to focus the exit pupil of the previous lens onto the lenses C_1 and C_2, respectively. The lower slit in the diagram is in the focal plane of L_1, the lens under test; and the upper slit is in the focal plane of the reference lens L_2.

After passing through L_1 and L_2, the beams are recombined at z before pas-

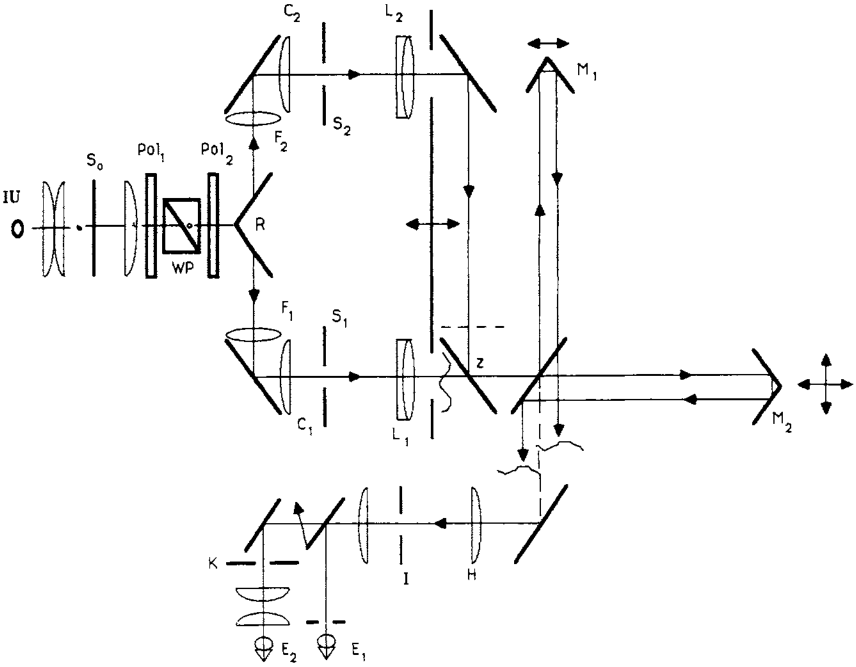


Figure 8.21. An interferometric measuring equipment: IU, illumination unit; S_0, S_1, S_2, slits; Pol_1, Pol_2, polarizers; WP, Wollaston prism; R, roof reflector; F_1, F_2, field lenses; C_1, C_2, slit focusing lenses; L_1, lens under test; L_2, reference (ideally perfect) lens; M_1, M_2, interferometer corner reflectors; H, lens forming interferogram at slit I; E_1, eye position for visual inspection; E_2, detector position for measurements; K, shaped aperture [19].

sage through the interferometer. The lens H, which typically has great depth of focus, images the pupils of L_1 and L_2 at the aperture I where an interference pattern, or ***interferogram***, is formed. With this pattern as the source and the following field lens to focus its image, two alternative optical systems are available to observe the interference pattern. When the mirror following the field lens is in the position shown in the diagram, an eye at E_1 can observe the image; however, when this mirror is swung upward out of the way for a measurement, the image is formed at K and can be viewed from E_2 through an eyepiece. After visual adjustment of the instrument is completed, a highly transmissive diffusing element is placed between I and K to produce a uniform illumination over K; and an electronic detector is placed at E_2 where it responds to the flux in the light beam.

The ratio of light flux in the L_1 and L_2 arms can be set by adjusting the elements in the polarizing assembly. First, the polarizer Pol_1 is set at a known angle; then, by rotating Pol_2 about the optic axis, the operator can smoothly vary the distribution of light flux from a hundred percent in one arm to a hundred percent in the other. The element S_0 is a spectrometer slit, which opens symmetrically; and S_1 and S_2 are iris diaphragms, each of which can be widely opened and fully illuminated to provide an effective extended source. This is useful during the initial setting up of the equipment.

The corner reflectors M_1 and M_2 are mounted on lathe-bed ways and can be accurately positioned laterally by adjustment of micrometer screws. Mirror M_2 can also be moved parallel to the incoming beam so that the optical paths in the two interferometer arms can be made virtually equal.

The aperture I is shaped so that it masks off the noninterfering light of the sheared wave fronts. Thus, the transmitted cross section is common to the two displaced circular wave fronts.

In this text, depending upon the context, the optical transfer function has been expressed mathematically a number of different ways. Examples are Eq. (2-19), where real-space spatial frequency is the independent variable; Eq. (5-48), where the normalized spatial frequency is substituted; and Eqs. (5-94) and (5-95), where the transfer function $\hat{b}_0$ (which is the OTF before normalization at zero frequency) is shown to be the autocorrelation function of the pupil function.

The OTF in terms of the normalized spatial frequency is

$$\hat{O}(s) = T(s) \exp[i\phi(s)] = T(s) \cos \phi(s) + iT(s) \sin \phi(s). \qquad (8\text{-}61)$$

The OTF as an autocorrelation function is

$$\hat{O}(s) = (1/\mathfrak{a}_0) \iint_{\mathfrak{a}} \exp i k \left\{ W[(x - s), y] - W(x, y) \right\} dx\, dy. \qquad (8\text{-}62)$$

where $1/\mathfrak{a}_0$ is the normalizing coefficient to make $\hat{O}(0) = 1$. So, from Eq. (8-61),

$$T(s) \cos \phi(s) = (1/\mathfrak{a}_0) \iint_{\mathfrak{a}} \cos k \left\{ W[(x - s), y] - W(x, y) \right\} dx\, dy, \qquad (8\text{-}63)$$

$$T(s) \sin \phi(s) = (1/\mathfrak{a}_0) \iint_{\mathfrak{a}} \sin k \left\{ W[(x - s), y] - W(x, y) \right\} dx\, dy. \qquad (8\text{-}64)$$

By substituting from Eqs. (8-63) and (8-64) into Eq. (8-60) we obtain

$$\Phi(s, \delta) = 2 \iint_{\mathcal{A}} dx\, dy + 2\mathcal{A}_0 T(s) \cos k\delta \cos \phi(s)$$

$$+ 2\mathcal{A}_0 T(s) \sin k\delta \sin \phi(s), \qquad (8\text{-}65)$$

$$\Phi(s, \delta) = 2 \iint_{\mathcal{A}} dx\, dy + 2\mathcal{A}_0 T(s) \cos[k\delta - \Phi(s)]. \qquad (8\text{-}66)$$

To simplify the right side of Eq. (8-66), the equation is divided through by $2\mathcal{A}_0$; then, because the first term is a constant for a given value of s, that is, a fixed amount of shear, it will be designated $B(s)$:

$$\Phi(s, \delta)/(2\mathcal{A}_0) = \Phi_1(s, \delta) = B(s) + T(s) \cos[k\delta - \phi(s)]. \qquad (8\text{-}67)$$

In the diagram of Fig. 8.21, the detector at E_2 senses the light flux $\Phi_1(s, \delta)$. As the path difference δ is varied at a constant speed, the detector signal varies sinusoidally with time; and the amplitude of the variation is proportional to $T(s)$, the MTF. The phase angle $\phi(s)$ is the phase transfer function (PTF) part of the OTF. If Pol_2 in Fig. 8.21 is adjusted so that only the plane wave from L_2 reaches the point z, Eq. (8-45) applies for the flux. Comparing Eq. (8-67) with Eq. (8-45) indicates that under the plane wave condition, $B(s)$ and $T(s)$ are equal and that the phase shift (PTF) is zero. With this as a reference, it is apparent that the difference in the observed phase angle between the aberrant and the plane waves is the PTF.

Baker's experimental results with the described equipment compare favorably with theoretical results calculated by Hopkins [20] for a defocused telescope objective. Hopkins' theoretical results are discussed in Appendix A.

OTHER INTERFEROMETRIC EQUIPMENTS

A number of other OTF measuring equipments have been described in the literature [21–33], but these in general involve practical improvements in experimental techniques rather than new basic principles of measurement.

A particularly interesting version is Kelsall's automatically recording instrument [21], which is capable of making rapid and accurate measurements on a wide variety of optical assemblies. To achieve stability of adjustment, the moving parts required for wave-front shearing and path changing are mounted independently of the interferometer bed. The stationary part of the interferometer

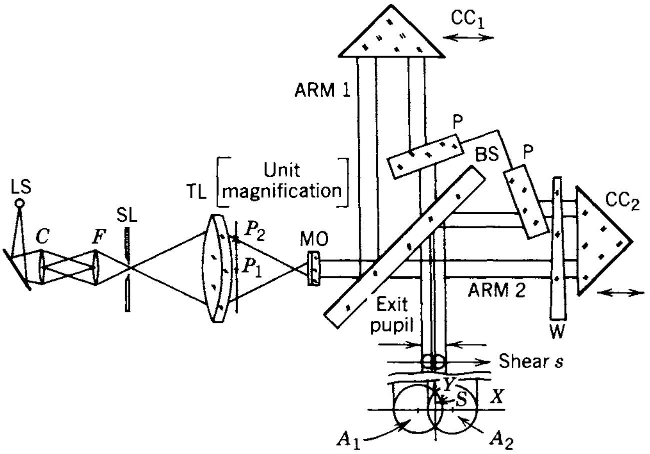


Figure 8.22. Kelsall's interferometer with shear plates [21].

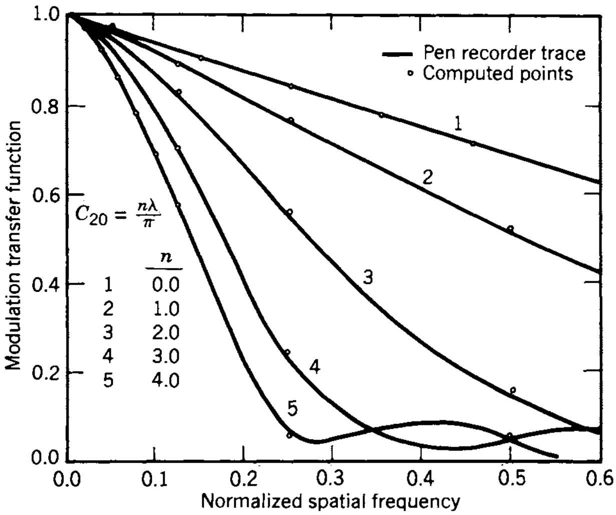


Figure 8.23. Comparison of pen recorder traces (curves), made by Kelsall's measuring equipment [21], with corresponding theoretical calculations (plotted points) by H. H. Hopkins for defocusing [20].

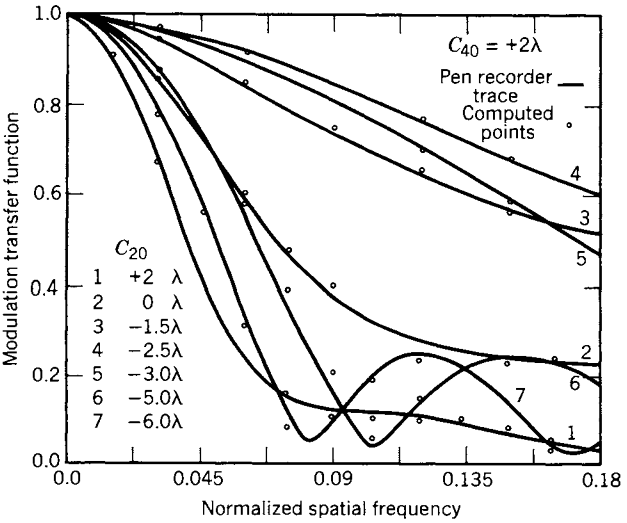


Figure 8.24. Comparison of pen recorder traces (curves), made by Kelsall's measuring equipment [21], with corresponding theoretical calculations (plotted points) by Black and Linfoot for primary spherical aberration [34].

is mounted on a rigid base plate supported by a vibration-free mounting. By this separation, it is possible to shear the wave fronts and alter the path difference without any detectable effect on other adjustments. Conversely, neither relative shearing nor tilting of the wave front occurs when adjustments are made.

As the schematic in Fig. 8.22 shows, Kelsall's interferometer shears the

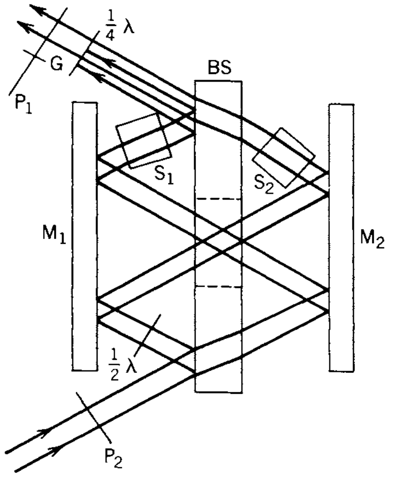


Figure 8.25. Optical transfer function measuring interferometer of Montgomery: M_1, M_2, mirrors; BS, beam splitter; S_1, S_2, shear plates; P_1, P_2, polarizers. Collimated beam from test lens enters from bottom left [23].

wave fronts with two plane-parallel optical flats, optically one in each arm of the interferometer but mechanically mounted together and swung on a common pivot; pivoting the flats performs the same function as the lateral movement of the corner reflectors.

Examples of Kelsall's results are shown in Figs. 8.23 and 8.24. In the first, a study of defocusing, the curves are replicas of the pen recorder trace made by the measuring apparatus; the plotted points are theoretical calculations by H. H. Hopkins. Figure 8.24 shows a similar comparison for primary spherical aberration. The theoretical points are those reported by Black and Linfoot [34] and discussed in Appendix A.

Figure 8.25 is a schematic of an interferometer, reported by Mongomery [23, 24], using polarizing techniques.

Wyant [24] describes an interferometer in which gratings accomplish the lateral shearing. This interferometer functions with a white light source; he elaborates on three conditions to make such a source work in a lateral shearing interferometer.

REFERENCES

1. K. Murata, Instruments for the Measuring of Optical Transfer Functions. In *Progress in Optics*, Vol. 5, E. Wolf (Ed.). North-Holland, Amsterdam, 1966.
2. K. Rosenhauer and K.-J. Rosenbruch, The Measurement of the Optical Transfer Functions of Lenses. In *Reports on Progress in Physics*, Vol. 30, Part 1, A. C. Stickland (Ed.). Institute of Physics and the Physical Society, London, 1967.
3. P. Kuttner, ''Interlaboratory Comparisons of MTF Measurements and Calculations. *Opt. Acta* **22,** 265 (1975).
4. British Calibration Service, Document 0751, Measurements of the Optical Transfer Function of Optical Systems. British Calibration Service, Stuart House, Soho Square, London.
5. ANSI, Guide to Optical Transfer Function Measurement and Reporting, ANSI PH3.57-1978. American National Standards Institute, New York.
6. P. Kuttner, Review of German Standards on Image Assessment. *Opt. Eng.* **17,** 90 (1978).
7. T. Nakamura, Y. Sekine, T. Nito, and T. Ose, OTF Standardization in Japan. *SPIE Proc.* **98,** 120 (1976). (Please see the note following Ref. 2 of Chapter 1.)
8. J. Simon, Measurement of the OTF in France. *SPIE Proc.* **98,** 125 (1976). (Please see the note following Ref. 2 of Chapter 1.)
9. Ir. J. A. J. van Leunen, OTF (Optical Transfer Functions) Standardization [in Holland]. *SPIE Proc.* **98,** 132 (1976). (Please see the note following Ref. 2 of Chapter 1.)

10. A. C. Marchant and E. A. Ironside, Influence of Optical Bench Errors on Accuracy of OTF Measurements. *Opt. Tech.* **2,** 85 (1970).

11. A. C. Marchant, Accuracy in Image Evaluation: Setting Up an OTF Standards Laboratory. *Opt. Acta* **18,** 133 (1971).

12. A. C. Marchant, E. A. Ironside, J. F. Attryde, and T. L. Willams, The Reproducibility of MTF Measurements. *Opt. Acta* **22,** 249 (1975).

13. P. Kuttner, The Influence of Microscope Objectives on the Measurement of the Modulation Transfer Function of Optical Systems. *Opt. Acta* **16,** 761 (1969).

14. R. L. Lamberts, Sine-wave Response Techniques in Photographic Printing. *J. Opt. Soc. Am.* **51,** 982 (1961).

15. N. S. Kapany and J. N. Pike, Scanning Integrator for Producing Sinusoidal Test Objects. *J. Opt. Soc. Am.* **46,** 867 (1956).

16. D. H. Kelly, M. E. Lynch, and D. S. Ross, Improved Sinusoidal Test Charts. *J. Opt. Soc. Am.* **48,** 858 (1958).

17. J. W. Coltman, The Specification of Imaging Properties by Response to a Sine Wave Input. *J. Opt. Soc. Am.* **44,** 468 (1954).

18. H. H. Hopkins, Interferometric Methods for the Study of Diffraction Images. *Opt. Acta* **2,** 23 (1955).

19. L. R. Baker, An Interferometer for Measuring the Spatial Frequency Response to a Lens System. *Proc. Phys. Soc. (London) Ser. B* **68,** 871 (1955).

20. H. H. Hopkins, The Frequency Response of a Defocused Optical System. *Proc. R. Soc. London Ser. A* **231,** 91 (1955).

21. D. Kelsall, Optical Frequency Response Characteristics in the Presence of Spherical Aberration Measured by an Automatically Recording Interferometric Instrument. *Proc. Phys. Soc. (London)* **73,** 465 (1959).

22. A. J. Montgomery, New Interferometer for the Measurement of Modulation Transfer Functions. *J. Opt. Soc. Am.* **54,** 191 (1964).

23. A. J. Montgomery, Two Methods of Measuring Optical Transfer Functions with an Interferometer. *J. Opt. Soc. Am.* **56,** 624. (1966).

24. J. C. Wyant, OTF Measurements with a White Light Source: An Interferometric Technique. *Appl. Opt.* **14,** 1613 (1975).

25. P. Hariharan and D. Sen, A Simple Interferometric Arrangement for the Measurement of Optical Frequency Response Characteristics. *Proc. Phys. Soc. (London)* **75,** 434 (1960).

26. T. Tsurata, Measurement of Transfer Functions of Photographic Objectives by Means of a Polarizing Shearing Interferometer. *J. Opt. Soc. Am.* **53,** 1156 (1963).

27. W. H. Steel, A Polarization Interferometer for the Measurement of Transfer Functions. *Opt. Acta* **11,** 9 (1964).

28. W. A. Minnick and J. D. Rancourt, Transfer Function Calculation Techniques for Real Optical Systems. *SPIE Proc.* **13,** 87 (1969). (Please see the note following Ref. 2 of Chapter 1.)

29. D. Kelsall, Rapid Interferometric Technique for MTF Measurements in the Visible or Infrared Region. *Appl. Opt.* **12,** 1398 (1973).

30. Several papers on the interferometric method were reported in Vol. 46 of the *Proceedings of the SPIE* (*SPIE Proc.* **46,** (1974) on Image Assessment and Specification, Dutton (Ed.), a seminar held in Rochester, NY, May 20–22, 1974. (Please see the note following Ref. 2 of Chapter 1.) R. E. Swing, The Case for the Pupil Function; D. R. Herriott and J. H. Bruning, Modulation Transfer Function by Measurement of the Pupil Function; K. Murata, H. Fujiwara, and R. Sato, Two-Dimensional Measurement of Optical Transfer Functions by Holographic Techniques; D. Kelsall, Rapid Real-Time MTF Measurements in a Field Environment; R. Moore and F. H. Slaymaker, Comparison of OTF Data Obtained from Edge-Scan and Interferometric Measurements.

31. F. T. Arecchi, M. Bassan, S. F. Jacobs, and G. Molesini, MTF Measurement via Diffraction Shearing with Optically Superimposed Gratings. *Appl. Opt.* **18,** 1247 (1979).

32. C. P. Grover and H. M. van Driel, Autocorrelation Method for Measuring the Transfer Function of Optical Systems. *Appl. Opt.* **19,** 900 (1980).

33. F. T. Arecchi, M. Bassan, and G. Molesini, A Simple Inexpensive MTF Meter. *Opt. Acta* **27,** 1263 (1980).

34. G. Black and E. H. Linfoot, Spherical Aberration and the Information Content of Optical Images. *Proc. R. Soc. London Ser. A* **239,** 522 (1957).

35. J. C. Wyant, A Simple Interferometric MTF Instrument. *Opt. Commun.* **19,** 120 (1976).

36. M. A. Gan, S. Ustinov, V. U. Kotov, P. A. Sergeev, and I. N. Tsvikevich, Computer Analysis of Interferograms and the Determination of the Point Spread Function and Optical Transfer Function during the Testing of Optical Systems. *Sov. J. Opt. Technol.* **45,** 448 (1978).

37. K. H. Wormack, Frequency Domain Description of Interferogram Analysis. *Opt. Eng.* **23,** 396 (1984).

9

Calculation of the OTF: Analytical Methods

INTRODUCTION

The specific calculation required for the analytical determination of the OTF depends on a number of factors. First, the choice depends on whether the calculation is (1) a purely theoretical one for a hypothetical optical system having a given kind and amount of aberration, (2) an anticipated evaluation from system design data for optics in the design stage, or (3) a practical determination from experimental results for an actual system that is being tested. In the calculation from experimental results, the method depends on the information available, for example, whether the pupil has been measured interferometrically or the aerial image of a point source (the aerial spread function) has been scanned by one of several methods: by a knife edge, an elongated slit, or a ''point'' sensor.

When the pupil function is known, the OTF can be calculated by either of two equivalent procedures. In one, the OTF is obtained directly by the autocorrelation of the pupil function. In the other, the amplitude spread function is first found by taking a Fourier transform of the pupil function. Then the intensity spread function is calculated by squaring the modulus of the amplitude spread function. Finally, the OTF is obtained by a Fourier transform of the intensity spread function. The latter combination is called the *double-transform method.*

The shape of a wave front at the instant it passes through the exit pupil is a fundamental characterization of the optical system and contains all the data about the imaging properties of the system; all of this information is retrievable—at least in principle. In Chapter 4, the wave-front shape is described mathematically by the wave aberration function, which gives the phase at points on the wave front relative to the phase at corresponding points on a reference sphere. The aberration function is thus a description of the errors, that is, the aberrations, produced by a lens or optical system in terms of the phase-difference distribution over the wave front at the exit pupil. Since typical variations of amplitude over the wave front generally produce negligible effects on imaging properties compared with the effects of even small changes of phase, this chapter will be limited to discussion of only the wave aberration function, which is the phase characteristic of the pupil function, and will assume the amplitude across the wave front to be constant.

Because the concept of *wave aberration* is based on a coherent wave train, involving it in the calculation of the OTF does not require specifying the degree of coherence in the illuminating light beam and hence does not depend on the type of object being imaged. This independence is a significant advantage in the interferometric measuring procedure for finding the wave-front shape.

The autocorrelation method and the double-transform method are more fundamental in principle than computations based on measurements of a spread function. Spread function methods assume incoherent light beams and can be in error because the actual experimental light beams in the laboratory are sometimes partially coherent, which may not be evident without a careful examination of the apparatus. Choice of a measurement-calculation procedure for obtaining the OTF involves a number of interrelated factors, both theoretical and practical; but under "average" restraints, it appears that the autocorrelation method has been the most useful.

To calculate the OTF from the aberration function, the aberration function must be expressed in analytic form. However, in the usual laboratory measurement, the amounts of wave-front distortion at specific, generally predetermined, coordinates are measured experimentally. That is, at each of a finite number of coordinate positions, say at (x_i', y_i') when x and y are the independent variables, a discrete value of the aberration function W_i, which is the dependent variable, is determined. These values can be assembled in tabular form or plotted as curve relations; but until mathematical expressions for the curves can be written, calculation of the OTF cannot proceed. Finding the mathematical expression that describes an experimental curve is called *curve fitting*.

Commonly used expressions to which the data are to be "fit" are various series: Power, exponential, polynomial, and sinusoidal (Fourier) series are familiar examples. The process involves numerical methods which are described in Chapter 10. The OTF is quite sensitive to variations in the aberration function; this sensitivity manifests itself particularly at the midrange and higher spatial frequencies. So any approximations, either in the observed data or in the subsequent curve fitting, must be examined carefully to be sure that they are valid over the entire wave front defined by the aperture. The actual calculation from the analytic aberration function to the OTF must itself maintain an accuracy consistent with the required accuracy in the end results; available procedures to meet this requirement are not simple. There are several to choose from, each of which has its peculiar advantages.

Hopkins [1] was able to set up an expression for the OTF, in terms of a converging series of Bessel functions, for an optical system free of aberrations but suffering from a defect of focus. M. De [2] originated a corresponding expression for astigmatism with defocusing due to Petzval curvature; his expression is a product of a series of Bessel functions.

In the following two sections of this chapter, we redo the calculations of Hopkins and De. For astigmatism with Petzal curvature, we obtain a new result. In each development, the OTF, $\hat{O}(s)$, is given by an exact expression; but the calculations are made by appropriately truncating the series, which are all rapidly convergent. In Chapter 10 on numerical methods, we illustrate a calculation involving curve-fitting the data.

THE OTF CALCULATED FOR DEFOCUSING

In this section we calculate the OTF for an optical system free of aberration but having a defect of focus. The procedure was introduced by Hopkins in 1955 [1]. The OTF, $\hat{O}(s)$, is the autocorrelation of the pupil function $\hat{G}(x, y)$ as given by Eq. (5-2). For reasons discussed earlier in this chapter, the modulus $G(x, y)$ is assumed constant—in fact, unity. Then, by combining the relations expressed in Eqs. (5-15) and (5-45) we obtain

$$\hat{O}(s) = (1/A_0) \iint_{\mathfrak{a}} \hat{G}\{[x + (s/2)], y\}\, \hat{G}^*\{[x - (s/2)], y\}\, dx\, dy, \tag{9-1}$$

where A_0 is new notation for $b_0(0)$ of Eq. (5-45), that is, the value of the integral of Eq. (9-1) when $s = 0$. Under the conditions imposed, the pupil function reduces to

$$\hat{G}(x, y) = \exp[ik\mathrm{W}(x, y)]. \tag{9-2}$$

A general expression for the aberration function $W(x, y)$ is given in polar coordinates by Eq. (4-33). The third term on the right side represents defocusing. Since all other aberrations are excluded in the present development,

$$W(\rho, \varphi) = {}_0C_{20}\rho^2. \tag{9-3}$$

The applicable rectangular-to-polar conversion equation is the familar $\rho^2 = x^2 + y^2$. The amount of defocusing is expressed by the coefficient ${}_0C_{20}$. To state this coefficient in terms of α', the number of quarter wavelengths:

$$ {}_0C_{20} = \alpha'\lambda/4. \tag{9-4}$$

When the expression $k = 2\pi/\lambda$, which is usually written as a constant coefficient of the aberration function, is combined with ${}_0C_{20}$, then

$$ k\,{}_0C_{20} = \alpha'\pi/2. \tag{9-5}$$

The graphical representation of the shifted pupil function (Fig. 5.2) can be dimensioned as shown in Fig. 9.1. Since the defocusing expression of Eq. (9-3) indicates radial symmetry, the OTF can be calculated along any radial direction; in our development, as shown in the figure, we choose a shift along the x-axis. With the combinations and assumptions discussed, the integrand of Eq. (9-1) becomes

$$\begin{aligned}&\exp\Big\langle (i\alpha'\pi/2)\Big\{\big[(x+s/2)^2+y^2\big]\\ &\qquad -\big[(x-s/2)^2+y^2\big]\Big\}\Big\rangle = \exp(i\alpha'\pi s x).\end{aligned} \tag{9-6}$$

At $s = 0$, the integrand $\exp(0)$ is unity. Then the indicated integration over the unit-radius circle $\mathcal{A}$ is π. Because $\hat{O}(0) = 1$, $A_0 = \pi$. Then Eq. (9-1) becomes

$$\hat{O}(s) = (1/\pi)\int_y \int_x \exp(i\alpha x)\,dx\,dy, \tag{9-7}$$

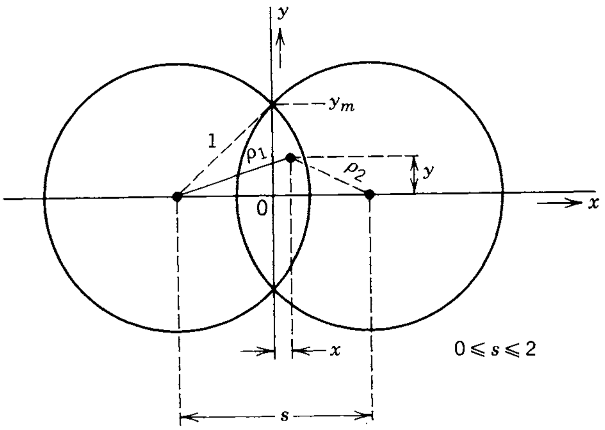


Figure 9.1. Geometry for the autocorrelation calculation.

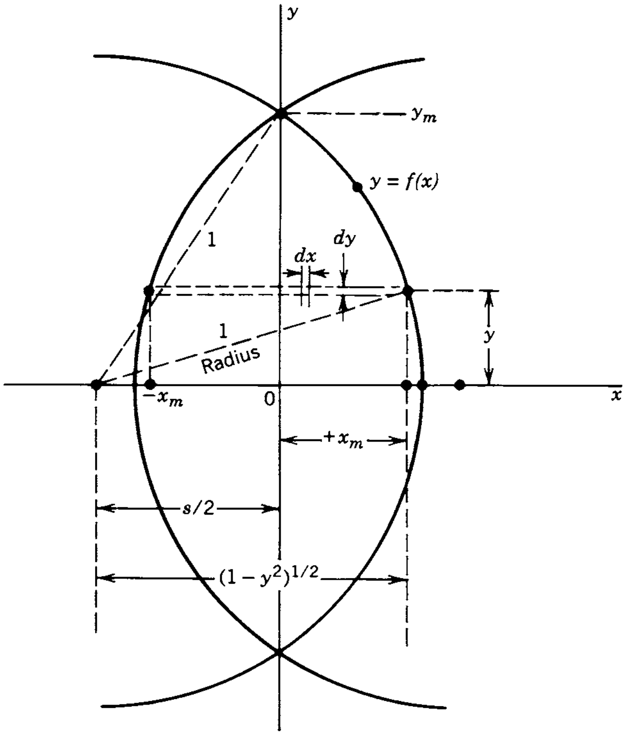


Figure 9.2. Geometry to determine integration limits.

where $\alpha = \alpha' \pi s$. As shown in Fig. 9.2, integration on x is between the upper limit $+x_m$ and the lower limit $-x_m$, that is, $\pm[(1 - y^2)^{1/2} - (s/2)]$. After integration,

$$\hat{O}(s) = (2/\pi\alpha) \int_y \sin(\alpha x_m)\, dy, \tag{9-8}$$

where we have applied the identity

$$\sin \alpha x_m = (1/i) \sinh(i\alpha x_m) = (1/i2)\left[\exp(i\alpha x_m) - \exp(-i\alpha x_m)\right]. \tag{9-9}$$

When the value for x_m is substituted in Eq. (9-8),

$$O(s) = (2/\pi\alpha) \int_y \sin \alpha\left[(1 - y^2)^{1/2} - (s/2)\right] dy. \tag{9-10}$$

Because of the conditions imposed on the pupil function, the OTF, $O(s)$, is real; hence no caret ($\hat{}$) is needed on the OTF symbol. The geometry of Figs. 9.1 and 9.2 is symmetrical about the x-axis, so integration can be made over only positive values of y and the resulting integral multiplied by 2. Limits on y are then zero and $[1 - (s/2)^2]^{1/2}$. A change in variable is needed to facilitate integration. The substitution chosen allows the integrand to be expressed as a series of Bessel functions; then integration can be accomplished term by term. Since y has already been limited to values between zero and unity, the following variable substitutions are permissible:

$$y = \sin\varphi, \qquad \varphi = \operatorname{arc\,sin} y,$$

$$dy = \cos\varphi \, d\varphi,$$

$$(1 - y^2) = (1 - \sin^2\varphi) = \cos^2\varphi,$$

$$\left[1 - (s/2)^2\right]^{1/2} = \beta = \operatorname{arc\,cos}(s/2). \tag{9-11}$$

By applying the identity for the sine of the difference of two angles, the integrand of Eq. (9-10) can be expanded as

$$\begin{aligned}\sin\left[\alpha(1 - y^2)^{1/2} - (\alpha s/2)\right] &= \sin\left[\alpha\cos\varphi - (\alpha s/2)\right] \\ &= \sin(\alpha\cos\varphi)\cos(\alpha s/2) \\ &\quad - \cos(\alpha\cos\varphi)\sin(\alpha s/2).\end{aligned} \tag{9-12}$$

The factors $\sin(\alpha\cos\varphi)$ and $\cos(\alpha\cos\varphi)$ are each defined by a series of Bessel functions (see, for example, Watson [3, p. 22] and Abramowitz [4, p. 361]):

$$\sin(\alpha\cos\varphi) = 2\sum_{n=1}^{\infty}(-1)^n\left[J_{2n+1}(\alpha)\right]\cos[(2n + 1)\varphi], \tag{9-13}$$

$$\cos(\alpha\cos\varphi) = J_0(\alpha) + 2\sum_{n=1}^{\infty}(-1)^n\left[J_{2n}(\alpha)\right]\cos(2n\varphi). \tag{9-14}$$

Here the $J_n(\alpha)$ are Bessel functions of the first kind, order n and argument α. When the series, Eqs. (9-13) and (9-14), are substituted into the right side of Eq. (9-12), the resulting expression is the integrand of Eq. (9-10), which now has the following form:

$$\begin{aligned}O(s) = {}& (4/\pi\alpha)\cos\alpha s/2\left\{2\sum_{n=1}^{\infty}(-1)^n\left[J_{2n+l}(\alpha)\right]\right.\\&\left.\cdot\int_0^{\beta}\cos\left[(2n+1)\varphi\right]\cos\varphi\,d\varphi\right\}\\&-(4/\pi\alpha)\sin\alpha s/2\left[J_0(\alpha)\right]\int_0^{\beta}\cos\varphi\,d\varphi\\&-(4/\pi\alpha)\sin\alpha s/2\left\{2\sum_{n=1}^{\infty}(-1)^n\left[J_{2n}(\alpha)\right]\int_0^{\beta}\cos 2\mathrm{n}\varphi\cos\varphi\,d\varphi\right\}.\end{aligned} \tag{9-15}$$

The integrals of Eq. (9-15) are standard forms given in most tables of integrals. See, for example, Selby [5, p. 437, No. 317] and Gradshteyn [6, p. 140, No. 2.532:3]. After integrating and regrouping, Eq. (9-15) becomes

$$\begin{aligned}O(s) = {}& (4/\pi\alpha)\cos\alpha s/2\left\{\beta J_1(\alpha) + (1/2)\sin 2\beta\left[J_1(\alpha) - J_3(\alpha)\right]\right.\\&\left.-(1/4)\sin 4\beta\left[J_3(\alpha) - J_5(\alpha)\right] + \cdots\right\}\\&-(4/\pi\alpha)\sin\alpha s/2\left\{\mathit{sin}\,\beta\left[J_0(\alpha) - J_2(\alpha)\right]\right.\\&-(1/3)\sin 3\beta\left[J_2(\alpha) - J_4(\alpha)\right]\\&\left.+(1/5)\sin 5\beta\left[J_4(\alpha) - J_6(\alpha)\right] - \cdots\right\}.\end{aligned} \tag{9-16}$$

The series in Eq. (9-16) are convergent and are in a convenient form to be evaluated numerically. However, depending on what resources are available in the way of Bessel function tables or equivalent computer software, the user may want to select convenient values of the argument α to reduce the tedium of getting the numerical Bessel function values for insertion in Eq. (9-16). A study of a given kind of aberration, such as defocusing in the present development, often consists of plotting a family of OTF curves, each differing from the next by a fixed increment in the coefficient (${}_0C_{20}$ for defocusing). If we label the increment $b\lambda$, where λ is the wavelength of light and b is the fraction of the wavelength that makes the argument α a convenient value, then each curve is for a given ${}_0C_{20}$, that is, an integral multiple m of the increment $b\lambda$; so, dropping the subscripts, we have

$$C = mb\lambda. \tag{9-17}$$

It remains to determine the value of b for convenient values of α. From Eq. (9-17) and the definitions of α' and α in Eq. (9-4) and following Eq. (9-7) respectively, we find

$$\begin{aligned} \alpha &= \alpha'\pi s = (4C/\lambda)\pi s, \\ C &= \alpha\lambda/(4\pi s), \\ b &= C/(m\lambda) = \alpha/(4\pi m s). \end{aligned} \tag{9-18}$$

Suppose that the desired increment in α is 0.5 and the increment in s for the OTF plot is 0.1. The values of m have already been established as integrals, that is, having increments of unity. When these increments are substituted in Eq. (9-18),

$$b = 5/4\pi \cong 0.3978873577 \cong 0.4 \text{ wavelength.} \tag{9-19}$$

Table 9.I Modulation Transfer Function Values for Defocusing ${}_0C_{20} = 5m\lambda/(4\pi)$

Normalized Frequency s	Perfect Lens $m = 0$	Defocusing for $m = 1$	Defocusing for $m = 2$
0.0	1.0000	1.0000	1.0000
0.1	0.9364	0.9104	0.8365
0.2	0.8729	0.7884	0.5675
0.3	0.8097	0.6586	0.3155
0.4	0.7471	0.5368	0.1338
0.5	0.6850	0.4317	0.0252
0.6	0.6238	0.3460	−0.0293
0.7	0.5636	0.2788	−0.0488
0.8	0.5046	0.2276	−0.0499
0.9	0.4470	0.1898	−0.0476
1.0	0.3910	0.1630	−0.0449
1.1	0.3368	0.1443	−0.0376
1.2	0.2848	0.1299	−0.0291
1.3	0.2351	0.1161	−0.0184
1.4	0.1881	0.1021	−0.0071
1.5	0.1443	0.0883	0.0033
1.6	0.1041	0.0729	0.0175
1.7	0.0682	0.0533	0.0241
1.8	0.0374	0.0327	
1.9	0.0133	0.0147	
2.0	0.0000		

A tabular sequence of values would then be

For $s = 0.1$,

$$\begin{array}{llllll} m = & 0 & 1 & 2 & 3 & 4 \\ \alpha = & 0 & 0.5 & 1.0 & 1.5 & 2.0 \end{array} \tag{9-20}$$

For $s = 0.2$,

$$\begin{array}{llllll} m = & 0 & 1 & 2 & 3 & 4 \\ \alpha = & 0 & 1.0 & 2.0 & 3.0 & 4.0 \end{array} \tag{9-21}$$

and so on. The first three columns of the routine described here ($m = 0, 1, 2$) were used to calculate the points tabulated in Table 9.I and plotted in Fig. 9.3.

A helpful tool for evaluating Bessel functions of successive orders for a given α, once two in the sequence have been established, is the recurrence formula

$$J_n(\alpha) = [2(n-1)/\alpha]J_{n-1}(\alpha) - J_{n-2}(\alpha). \tag{9-22}$$

Levi [7, 8], using a different mathematical approach from the one discussed here, presents a large table of values for $O(s)$ as a function of s and different

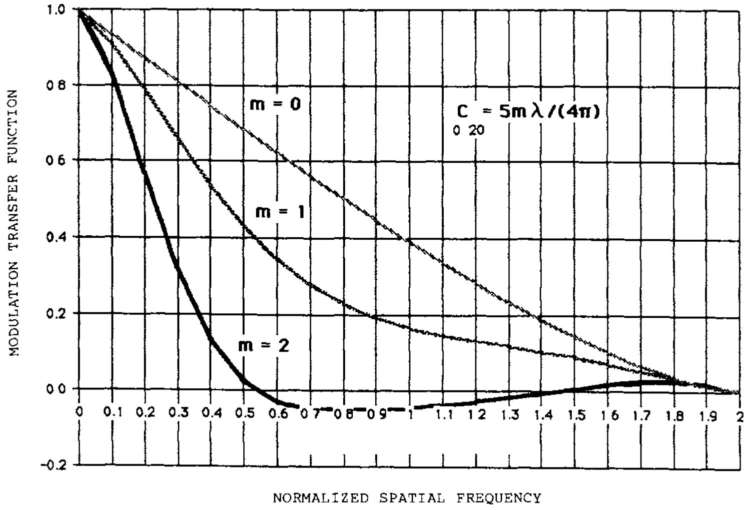


Figure 9.3. Modulation transfer function (MTF) curves for a perfect optical system and for two amounts of defocusing.

amounts of defocusing. Curves for $O(s)$ as plotted by Hopkins [1] are given in Appendix A.

THE OTF CALCULATED FOR ASTIGMATISM

M. De [2] calculated the OTF for primary astigmatism with Petzval curvature by a method similar to that of Hopkins, which is discussed in the previous section. Although our approach in this section is similar to that of De, we show a different result.

As in the previous section, the present OTF calculation begins with the power series expression for the aberration function $W(x, y)$ given in polar coordinates by Eq. (4-33). Following Eq. (4-39), the power series terms are sorted out and assigned names based on the traditional Seidel designations. Among these are two terms, commonly paired in optics, that are together called *astigmatism and Petzval curvature*:

$$W_a = {}_2C_{20}\, r^2\rho^2 + {}_2C_{22} r^2\rho^2 \cos^2 \varphi. \tag{9-23}$$

Since the OTF is always calculated for a fixed r, this independent variable can be suppressed in Eq. (9-23) by incorporating it into the coefficient of each term. Then the first right term, $C_{20}\rho^2$, takes the form of a defocusing term, which is the subject of the previous section. The reference image point is a point on a plane perpendicular to the axis at the paraxial image point; the significance of this reference is discussed following the calculations in the present section.

For integration steps that are to follow, it will be convenient to express the aberration function of Eq. (9-23) in rectangular coordinates (x_0, y_0):

$$\begin{aligned} W_a &= C_{20}\rho^2 + C_{22}\rho^2 \cos^2 \varphi = C_{20}(x_0^2 + y_0^2) + C_{22}y_0^2 \\ &= C_{20}x_0^2 + (C_{20} + C_{22})y_0^2. \end{aligned} \tag{9-24}$$

The coordinate transformation equations in slightly different notation are given in Eq. (4-31).

Examination of the contour plot for primary astigmatism, Fig. 4.20, indicates that we should expect the OTF in the tangential direction to be different from the OTF in the sagittal direction; in fact, the OTF depends upon the angle ψ of any arbitrary direction between these two. So, for an all-inclusive analysis, it is convenient to transform the rectangular coordinates in accordance with the OTF angle:

$$x_0 = x \cos \psi - y \sin \psi,$$

$$y_0 = x \sin \psi + y \cos \psi, \qquad 0 \leqq \psi \leqq \pi/2. \qquad (9\text{-}25)$$

When these expressions are substituted in Eq. (9-24) and the appropriate trigonometric identities applied:

$$W(x, y) = C_{20}(x^2 + y^2) + C_{22}(x^2 \sin^2 \psi + y^2 \cos^2 \psi + xy \sin 2\psi). \qquad (9\text{-}26)$$

The integrand in an equation corresponding to Eq. (9-6) in the previous section is

$$\begin{aligned}&\exp\left\langle i\not{k}\left\{W[(x + s/2), y] - W[(x - s/2), y]\right\}\right\rangle \\ &\quad = \exp\left\{i\not{k}s[2x(C_{20} + C_{22} \sin^2 \psi) + yC_{22} \sin 2\psi]\right\}. \qquad (9\text{-}27)\end{aligned}$$

Since ψ is constant for a given OTF, in the corresponding series of integrations (for $s = 0.1, 0.2, 0.3, \ldots$), the mathematical notation is simplified by defining two new constants:

$$p = 2\not{k}s(C_{20} + C_{22} \sin^2 \psi), \qquad (9\text{-}28)$$

$$q = \not{k}sC_{22} \sin 2\psi. \qquad (9\text{-}29)$$

Then the expression for the OTF corresponding to Eq. (9-7) of the previous section is

$$\hat{O}(s, \psi) = (1/\pi) \int_y \exp iqy \left(\int_x \exp ipx\, dx \right) dy. \qquad (9\text{-}30)$$

In a search for symmetry to simplify the integration process, we find that there is no simple symmetry about the y-axis nor about the x-axis because, as Eq. (9-26) shows,

$$W(x, y) \neq W(-x, y),$$

$$W(x, y) \neq W(x, -y). \qquad (9\text{-}31)$$

However, there is a more subtle symmetry as shown by

$$W(x, y) = W(-x, -y),$$

$$W(-x, y) = W(x, -y). \qquad (9\text{-}32)$$

The significance of these symmetries can be appreciated by referring to Fig. 9.1 plotted in the (x, y) coordinates. The overlapping area in the first quadrant is equivalent to the overlapping area in the third quadrant. (The two would be alike in all respects if the first were rotated about the y-axis and then about the x-axis.) The same kind of relation exists between the second and fourth quadrants. Because of the symmetry, integrating over the upper half from $-x_m$ to $+x_m$ is the same as integrating over the lower half from $+x_m$ to $-x_m$. Therefore, the OTF, Eq. (9-30), is found by integrating over the upper half of the overlapping area and then multiplying by 2. As in an earlier integration, the limits are apparent from Fig. 9.2:

$$\hat{O}(s, \psi) = (2/\pi) \int_0^{y_m} \exp(iqy) \left[\int_{-x_m}^{+x_m} \exp(ipx)\, dx \right] dy,$$

$$y_m = \left[1 - (s/2)^2\right]^{1/2}, \qquad x_m = \left[(1 - y^2)^{1/2} - s/2\right]. \tag{9-33}$$

As in Eq. (9-7) of the previous section, the integration on x is straightforward:

$$\hat{O}(s, \psi) = (4/\pi p) \int_0^{y_m} \left\{\exp(iqy) \sin p\left[(1 - y^2)^{1/2} - s/2\right]\right\} dy. \tag{9-34}$$

As in the previous development, the integration on y can be accomplished by expanding the integrand into a series of Bessel functions. Expansion is facilitated by the following changes in variables:

$$y = \sin \omega, \qquad dy = \cos \omega\, d\omega,$$

$$(1 - y^2)^{1/2} = \cos \omega,$$

$$\beta = \text{arc} \cos (s/2) = \text{arc} \sin \left[1 - (s/2)^2\right]^{1/2}. \tag{9-35}$$

With the above substitutions, Eq. (9-34) becomes

$$\hat{O}(s, \psi) = (4/\pi p) \left[\int_0^{\beta} \exp(iq \sin \omega) \sin (p \cos \omega) \cos(ps/2) \cos \omega\, d\omega \right.$$

$$\left. - \int_0^{\beta} \exp(iq \sin \omega) \cos(p \cos \omega) \sin(ps/2) \cos \omega\, d\omega \right]. \tag{9-36}$$

In Eq. (9-36), constant terms can be placed before the integral signs and the exponential terms expanded to give

$$\begin{aligned}\hat{O}(s,\psi) &= (4/\pi p)\cos(ps/2)\int_0^{\beta}\left[\cos(q\sin\omega)\sin(p\cos\omega)\cos\omega\right.\\&\quad\left.+\,i\sin(q\sin\omega)\sin(p\cos\omega)\cos\omega\right]d\omega\\&\quad-(4/\pi p)\sin(ps/2)\int_0^{\beta}\left[\cos q\sin\omega)\cos(p\cos\omega)\cos\omega\right.\\&\quad\left.+\,i\sin(q\sin\omega)\cos(p\cos\omega)\cos\omega\right]d\omega.\end{aligned}\tag{9-37}$$

The following Bessel function series expansions of terms in Eq. (9-37) expedite integration [3, p. 22; 4, p. 361]:

$$\cos(z\sin\varphi)=\sum_{n=0}^{\infty}\epsilon_n J_{2n}(z)\cos(2n\varphi),\tag{9-38}$$

$$\sin(z\sin\varphi)=2\sum_{n=0}^{\infty}J_{2n+l}(z)\sin(2n+1)\varphi,\tag{9-39}$$

$$\cos(z\cos\varphi)=\sum_{n=0}^{\infty}(-1)^n\epsilon_n J_{2n}(z)\cos(2n\varphi),\tag{9-40}$$

$$\sin(z\cos\varphi)=2\sum_{n=0}^{\infty}(-1)^n J_{2n+l}(z)\cos(2n+1)\varphi.\tag{9-41}$$

In Eqs. (9-38) and (9-40), $\epsilon_n = 1$ when $n = 0$, and $\epsilon_n = 2$ when $n \neq 0$. After making the indicated substitutions and rearranging, Eq. (9-37) becomes

$$\begin{aligned}\hat{O}(s,\psi) &= 4/\pi p)\Bigg\{\cos(ps/2)\Bigg[2\sum_{n=0}^{\infty}\sum_{m=0}^{\infty}(-1)^m\epsilon_n J_{2n}(q)\\&\quad\cdot J_{2m+l}(\mathrm{p})\int_0^{\beta}\mathrm{f}_1(\omega)\,\mathrm{d}\omega\\&\quad+\mathrm{i}4\sum_{j=0}^{\infty}\sum_{k=0}^{\infty}(-1)^k J_{2j+l}(q)\,\mathrm{J}_{2k+l}(\mathrm{p})\int_0^{\beta}f_2(\omega)\,d\omega\Bigg]\\&\quad-\sin(ps/2)\Bigg[\sum_{n=0}^{\infty}\sum_{m=0}^{\infty}(-1)^m\epsilon_n\epsilon_m J_{2n}(q)\,J_{2m}(p)\int_0^{\beta}f_3(\omega)\,d\omega\\&\quad+i2\sum_{j=0}^{\infty}\sum_{k=0}^{\infty}(-1)^k\epsilon_k J_{2j+l}(q)\,J_{2k}(p)\int_0^{\beta}f_4(\omega)\,d\omega\Bigg]\Bigg\}.\end{aligned}\tag{9-42}$$

In Eq. (9-42)

$$f_1(\omega) = \cos(2n\omega)\cos[(2m+1)\omega]\cos\omega, \tag{9-43}$$

$$f_2(\omega) = \sin[(2j+1)\omega]\cos[(2k+1)\omega]\cos\omega, \tag{9-44}$$

$$f_3(\omega) = \cos(2n\omega)\cos(2m\omega)\cos\omega, \tag{9-45}$$

$$f_4(\omega) = \sin[(2j+1)\omega]\cos(2k\omega)\cos\omega. \tag{9-46}$$

The functions defined in Eqs. (9-43)–(9-46), which are the integrands in Eq. (9-42), integrate according to formulas in Ref. 7, p. 140, nos. 3 and 5. When the integration is completed, the OTF can be expressed as

$$\begin{aligned}\hat{O}(s,\psi) = (1/\pi p)\Bigg\{&\cos(ps/2)\left[2\sum_{n=0}^{\infty}\sum_{m=0}^{\infty}(-1)^m\epsilon_n J_{2n}(q)\,J_{2m+l}(p)\,f_5(\beta)\right]\\ &-\sin(ps/2)\left[\sum_{n=0}^{\infty}\sum_{m=0}^{\infty}(-1)^m\epsilon_n\epsilon_m J_{2n}(q)\,J_{2m}(p)\,f_6(\beta)\right]\\ &-i\cos(ps/2)\left[4\sum_{j=0}^{\infty}\sum_{k=0}^{\infty}(-1)^k J_{2j+l}(q)\,J_{2k+l}(p)\,f_7(\beta)\right]\\ &+i\sin(ps/2)\left[2\sum_{j=0}^{\infty}\sum_{k=0}^{\infty}(-1)^k\epsilon_k J_{2j+l}(q)\,J_{2k}(p)\,f_8(\beta)\right]\Bigg\}.\end{aligned} \tag{9-47}$$

In Eq. (9-47), the four functions of β are

$$\begin{aligned}f_5(\beta) = {}&\sin[2(m+n+1)\beta]/[2(m+n+1)]\\ &+\sin[2(m-n+1)\beta]/[2(m-n+1)]\\ &+\sin[(2n-2m)\beta]/(2n+2m)\\ &+\sin[(2n+2m)\beta]/(2n+2m),\end{aligned} \tag{9-48}$$

$$\begin{aligned}f_6(\beta) = {}&\sin[(2n+2m+1)\beta]/(2n+2m+1)\\ &+\sin[(2m-2n+1)\beta]/(2m-2n+1)\\ &+\sin[(2n-2m+1)\beta]/(2n-2m+1)\\ &+\sin[(2n+2m-1)\beta]/(2n+2m-1),\end{aligned} \tag{9-49}$$

$$\begin{aligned} f_7(\beta) = {} & \Big\{\cos\big[(2k + 2j + 3)\beta\big] - 1\Big\}/(2k + 2j + 3) \\ & - \Big\{\cos\big[(2k - 2j + 1)\beta\big] - 1\Big\}/(2k - 2j + 1) \\ & + \Big\{\cos\big[(2k + 2j + 1)\beta\big] - 1\Big\}/(2k + 2j + 1) \\ & + \Big\{\cos\big[(2j - 2k + 1)\beta\big] - 1\Big\}/(2j - 2k + 1), \end{aligned} \tag{9-50}$$

$$\begin{aligned} f_8(\beta) = {} & \Big\{\cos\big[2(j + k + 1)\beta\big] - 1\Big\}/\big[2(j + k + 1)\big] \\ & - \Big\{\cos\big[2(k - j)\beta\big] - 1\Big\}/\big[2(k - j)\big] \\ & + \Big\{\cos\big[2(j + k)\beta\big] - 1\Big\}/\big[2(j + k)\big] \\ & + \Big\{\cos\big[2(j - k)\beta\big] - 1\Big\}/\big[2(j - k)\big]. \end{aligned} \tag{9-51}$$

It is interesting to note in all four of the above equations that the indeterminate form $0/0$ occurs whenever the coefficient of β is zero. When these ambiguities are resolved by L'Hôpital's rule, the ratios for zero coefficients of β in Eqs. (9-48) and (9-49) are shown to have the value β; for zero coefficients of β in Eqs. (9-50) and (9-51), the ratios are shown to be zero.

Equation (9-47) is an exact expression for the OTF in terms of the normalized spatial frequency, the amount of primary Petzval curvature, the primary astigmatism, and the orientation with respect to the sagittal direction. Though quite unwieldy, this expression, of course, can be programmed on a computer having subroutines for the trigonometric and Bessel functions. M. De [2] does not include an Imaginary component in the expression for the OTF because, he says, "the aberration being symmetrical about the x- and y-axes, the response is wholly real." However, we have not been able to verify that the imaginary term of Eq. (9-47) should in general be zero. The spread function in Figs. 2.17*b* and *d* for astigmatism and the wave shape in Figs. 4.20 and 4.22 hardly suggest freedom from phase shift for all possible related OTFs.

When the OTF angle ψ is zero, that is, when the OTF is in the sagittal direction, Eqs. (9-28) and (9-29) show that the constant q is zero and the constant p reduces to the α defined in Eq. (9-18); then the integrand in Eq. (9-27) becomes identical with the corresponding defocusing integrand of Eq. (9-6). So the OTF in the sagittal direction for primary Petzval curvature and primary astigmatism is the same as the OTF for defocusing. This conclusion is confirmed by Eq. (9-24) because the angle φ has the value $\pi/2$ in the sagittal direction ($\psi = 0$), and the aberration function equation becomes the same as Eq. (9-3), which describes defocusing.

In the tangential direction ($\psi = \pi/2$), the constant q is again zero. This alone tends to simplify Eq. (9-47) because the Bessel functions involved, Bessel functions of the *first kind*, all have a zero value for a zero argument except the zero-order Bessel function, which is unity for a zero argument. Both of the imaginary terms in Eq. (9-47) have $J_{2j+l}(q)$ as a factor; so for a zero q, which occurs in both the sagittal and tangential directions, the OTF, $\hat{O}(s, 0)$ or $\hat{O}(s, \pi/2)$, is indeed real and can be written without the complex caret (ˆ). Both of the real terms in Eq. (9-47) have $J_{2n}(q)$ as a factor; so, for zero q, this Bessel function is unity for zero n and zero for all other values of n. When the sagittal and tangential direction ($\psi = 0$ and $\psi = \pi/2$) values are applied in Eq. (9-47), it reduces to the following for both directions:

$$O(s, \psi_a) = (1/\pi p) \sum_{m=0}^{\infty} (-1)^m [2 \cos(ps/2)\, J_{2m+l}(p) f_5'(\beta) - \epsilon_m \sin(ps/2)\, J_{2m}(p) f_6'(\beta)]. \qquad (9\text{-}52)$$

Expressions for $f_5'(\beta)$ and $f_6'(\beta)$ in Eq. (9-52), derived from Eqs. (9-48) and (9-49), are

Table 9.II Optical Transfer Function Values for the Direction $\psi = \pi/6$

Normalized Frequency	Perfect	For $p = 3s$, $q = 3s$		For $p = 5s$, $q = 5s$	
s	Lens	MTF	PTF	MTF	PTF
0.0	1.0000	1.0000	0.0	1.0000	0.0
0.1	0.9364			0.9240	−0.314
0.2	0.8730			0.8395	−0.579
0.3	0.8097	0.8000	−0.524	0.6962	−0.758
0.4	0.7471	0.7158	−0.628	0.4964	−0.873
0.5	0.6850	0.6093	−0.681	0.3041	−1.035
0.6	0.6238	0.4863	−0.688	0.2124	−1.459
0.7	0.5636	0.3635	−0.671	0.2412	−1.794
0.8	0.5046	0.2595	−0.694	0.2700	−1.876
0.9	0.4470	0.1744	−0.867	0.2211	−1.912
1.0	0.3910	0.1721	−1.202	0.1248	−2.058
1.1	0.3368	0.1897	−1.458	0.0649	−2.362
1.2	0.2848	0.2005	−1.554	0.0585	−2.147
1.3	0.2351	0.1803	−1.555	0.0735	−1.766
1.4	0.1881	0.1340	−1.491	0.0652	−1.625
1.5	0.1443	0.0187	−1.354	0.0404	−1.625
1.6	0.1041	0.0454	−1.114	0.0284	−1.718
1.7	0.0682			0.0273	−1.579
1.8	0.0374			0.0230	−1.610
1.9	0.0133			0.0109	−1.591
2.0	0.0000			0.0000	

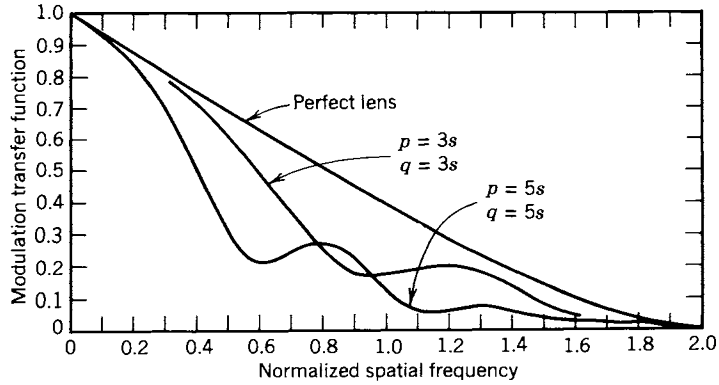


Figure 9.4. Modulation transfer function (MTF) curves for a perfect optical system and for two amounts of astigmatism and Petzval curvature at an OTF direction of $\pi/6$.

$$f_5'(\beta) = 2 \sin[2(m+1)\beta]/[2(m+1)] + 2 \sin(2m\beta)/(2m), \tag{9-53}$$

$$f_6'(\beta) = 2 \sin[(2m+1)\beta]/(2m+1) + 2 \sin[(1-2m)\beta]/(1-2m). \tag{9-54}$$

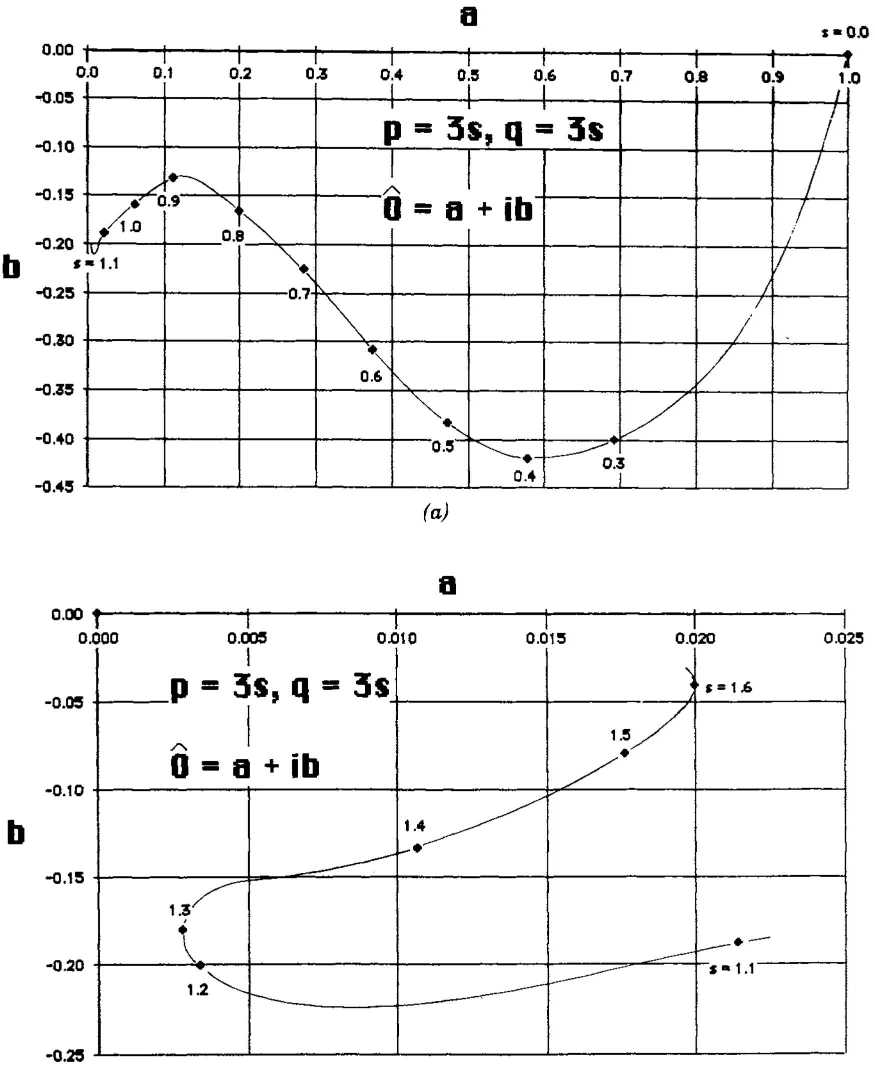
The OTF in the tangential direction is distinguished from the OTF in the sagittal direction by the value of p, the argument of the remaining Bessel functions. In the tangential direction, from Eq. (9-28),

$$p_T = 2ks(C_{20} + C_{22}). \tag{9-55}$$

In the sagittal direction,

$$p_S = 2ksC_{20}. \tag{9-56}$$

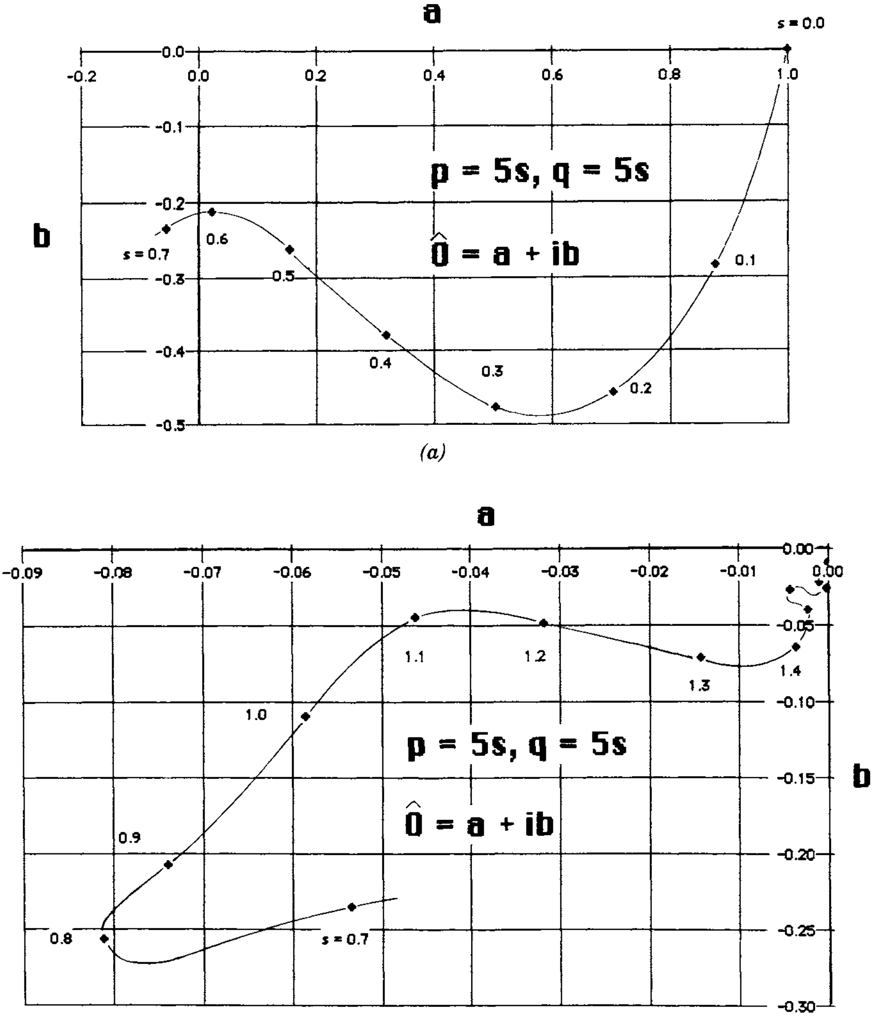
Since it has already been pointed out that when $\psi = 0$ (sagittal direction), corresponding integrands for the combined Petzval curvature and primary astigmatism and for defocusing are identical, we expect that the consequent expressions for the OTF, Eqs. (9-52) and (9-16), are also the same. This is indeed true, but considerable rearrangement of terms is required for the identity to be evident. (As in the integrand comparison, p and α are the same when ψ is zero.) Since Eq. (9-52) serves for both $\psi = 0$ (sagittal direction) and $\psi =$



(a)

(b)

Figure 9.5. Optical transfer function (OTF) plotted on a complex plane (Argand diagram) for astigmatism and Petzval curvature at an OTF direction of $\pi/6$, Bessel function arguments $p = 3s$ and $q = 3s$.



(a)

(b)

Figure 9.6. Optical transfer function (OTF) plotted on a complex plane (Argand diagram) for astigmatism and Petzval curvature at an OTF direction of $\pi/6$, Bessel function arguments $p = 5s$ and $q = 5s$.

$\pi/2$ (tangential direction), the distinction being the value of p as detailed in Eqs. (9-55) and (9-56), we might ask how the Petzval-astigmatism OTF in the tangential direction relates to the defocusing OTF. Because the two expressions for p are identical except that in the tangential direction the coefficient C_{20} is replaced by $(C_{20} + C_{22})$, we can conclude that the OTF in the tangential direction, like the OTF in the sagittal direction, has a defocusing OTF shape but with a different value for the defocusing coefficient.

Though the OTFs in the sagittal and tangential directions, because the expressions for them are relatively simple, may usually be chosen for calculation, an OTF calculation discussion would hardly be complete without some reference to calculations in some intermediate directions. For these, all the terms in Eq. (9-47) must be considered. How one proceeds with the actual calculations depends on what the objectives are and particularly what computer resources are at hand. Of course, if all the necessary software is available, the optical worker need only (1) assume the amount of aberration by selecting val-

Table 9.III Optical Transfer Function Values for the Direction $\psi = \pi/3$

Normalized Frequency s	Perfect Lens	For $p = 6.5s$, $q = 3s$		For $p = 10.8s$, $q = 5s$	
		MTF	PTF	MTF	PTF
0.0	1.0000	1.0000	0.0	1.0000	0.0
0.1	0.9364	0.9014	−0.192	0.8395	−0.314
0.2	0.8730	0.7659	−0.367	0.5800	−0.519
0.3	0.8097	0.6051	−0.456	0.2303	−0.587
0.4	0.7471	0.4290	−0.428	−0.0115	−1.736
0.5	0.6850	0.2708	−0.307	−0.0846	−2.959
0.6	0.6238	0.1500	−0.276	−0.0839	−3.168
0.7	0.5636	0.0932	−0.878	−0.0899	−3.813
0.8	0.5046	0.1151	−1.473	−0.0723	−3.911
0.9	0.4470	0.1090	−1.666	−0.0559	−4.185
1.0	0.3910	0.0724	−1.908	−0.0693	−4.267
1.1	0.3368	0.0638	−2.136	−0.0430	−4.349
1.2	0.2848	0.0770	−1.994	−0.0448	−4.369
1.3	0.2351	0.0735	−1.839	−0.0441	−4.376
1.4	0.1881	0.0560	−1.782	−0.0521	−4.396
1.5	0.1443	0.0572	−1.698	−0.0301	−4.030
1.6	0.1041	0.0597	−1.518	−0.0136	−3.041
1.7	0.0682	0.0405	−1.237	−0.0119	−2.465
1.8	0.0374	0.0222	−0.943		
1.9	0.0133				
2.0	0.0000				

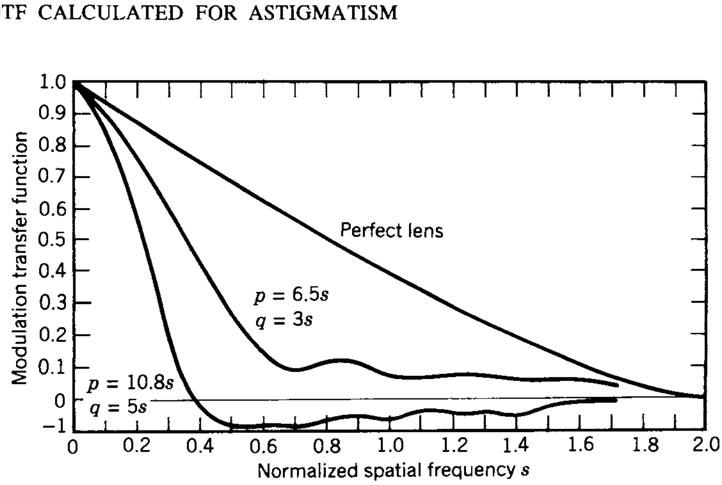


Figure 9.7. Modulation transfer function (MTF) curves for a perfect optical system and for two amounts of astigmatism and Petzval curvature at an OTF direction of $\pi/3$.

ues of C_{20} and C_{22}, (2) determine the desired OTF direction by selecting ψ, and then (3) plug in successive values of s and read the corresponding values of $\hat{O}$ for the OTF curves. However, if the worker is constrained to using tables of Bessel functions and a scientific calculator and wants to avoid elaborate interpolation, certain juggling of assumed values is desirable with some consequent restrictions on freedom of choice. We offer some examples of the latter situation.

$$\psi_1 = \pi/6 \text{ radian} = 30°, \tag{9-57}$$

$$\psi_2 = \pi/3 \text{ radian} = 60°. \tag{9-58}$$

For ψ_1,

$$\sin \psi_1 = 1/2, \qquad \sin^2 \psi_1 = 1/4, \qquad \sin 2\psi_1 = \sin \pi/3 = \sqrt{3}/2. \tag{9-59}$$

For ψ_2,

$$\sin \psi_2 = \sqrt{3}/2, \qquad \sin^2 \psi_2 = 3/4, \qquad \sin 2\psi_2 = \sqrt{3}/2. \tag{9-60}$$

The arguments p and q for the Bessel functions in Eq. (9-47), first defined in Eqs. (9-28) and (9-29), should be convenient for table look-up. This can be attained by proper choice of the aberration constants C_{20} and C_{22} in the defining equations:

$$p = 2ks(C_{20} + C_{22}\sin^2\psi) = 2ksC_{22}[(C_{20}/C_{22}) + \sin^2\psi], \quad (9\text{-}61)$$

$$q = ksC_{22}\sin 2\psi. \quad (9\text{-}62)$$

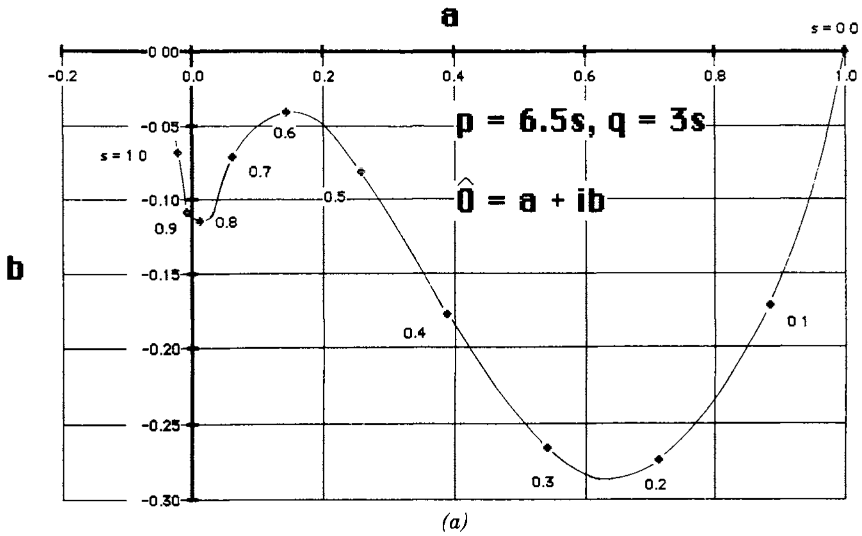
For the direction $\psi = \psi_1 = \pi/6$, a convenient value of the coefficient ratio is

$$C_{20}/C_{22} = (\sqrt{3} - 1)/4 \cong 0.1830127019 \cong 0.18. \quad (9\text{-}63)$$

The first value ("exact value"), given as a fraction, is to be used in calculations; the decimal values are only to indicate the magnitude of the assumed ratio. For this ratio of coefficients, two sets, a and b, of OTF (consisting of both MTF and PTF parts) values are calculated, for which we choose convenient values of C_{20}:

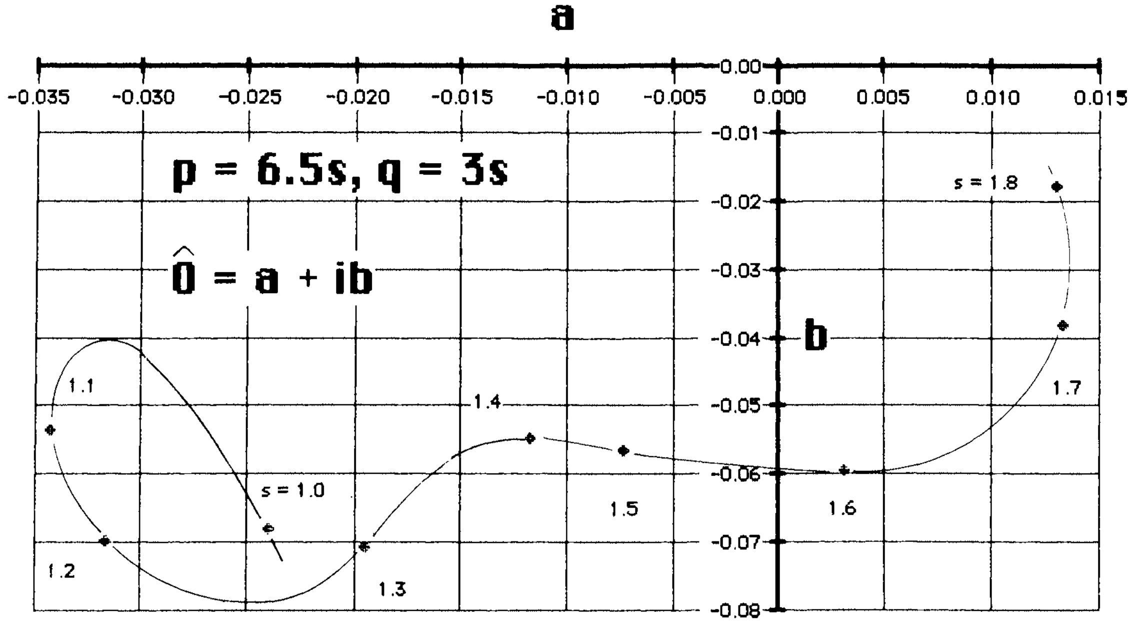
$$(C_{20})_a = [3/(\pi\sqrt{3})]\lambda \cong 0.5513288954\lambda \cong 0.55\lambda,$$

$$(C_{20})_b = [5/(\pi\sqrt{5})]\lambda \cong 0.9188814924\lambda \cong 0.92\lambda. \quad (9\text{-}64)$$



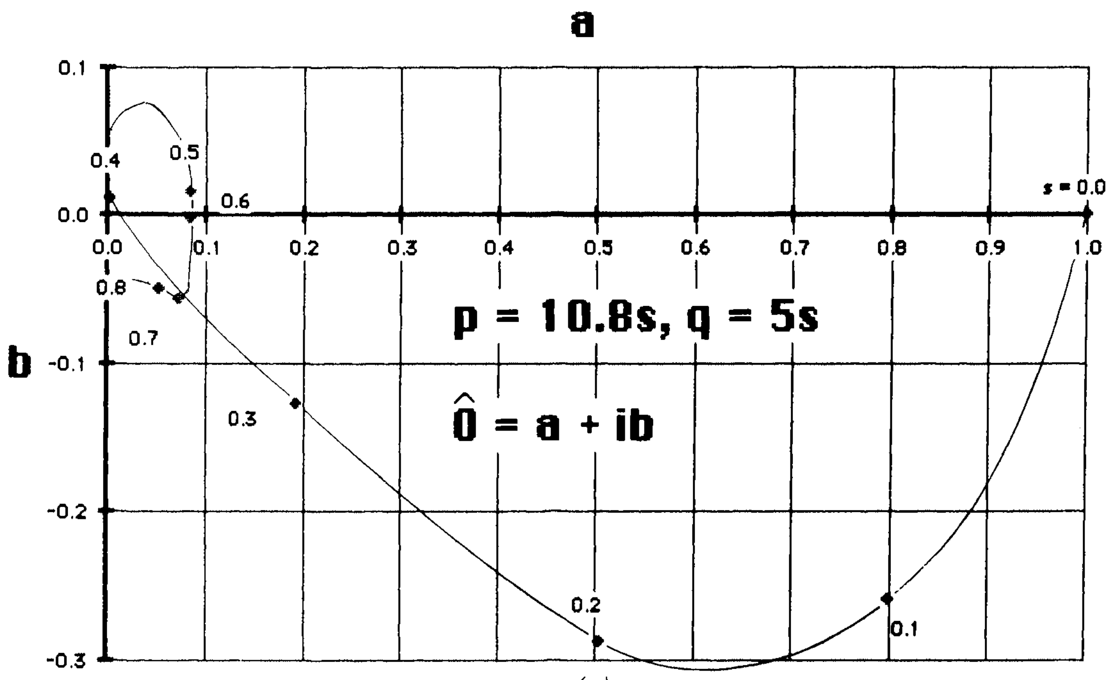
(a)

Figure 9.8. Optical transfer function (OTF) plotted on a complex plane (Argand diagram) for astigmatism and Petzval curvature at an OTF direction of $\pi/3$, Bessel function arguments $p = 6.5s$ and $q = 3s$.



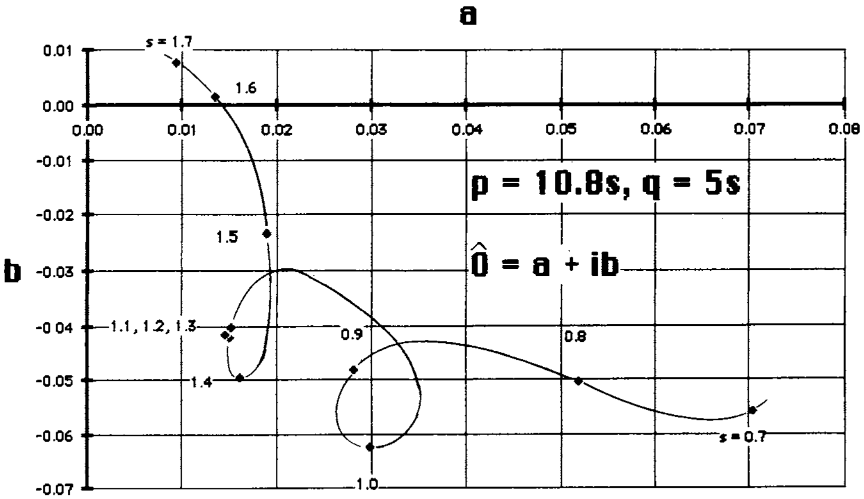
(b)

Figure 9.8. (*Continued*)



(a)

Figure 9.9. Optical transfer function (OTF) plotted on a complex plane (Argand diagram) for astigmatism and Petzval curvature at an OTF direction of $\pi/3$, Bessel function arguments $p = 10.8s$ and $q = 5s$.



(b)

Figure 9.9. (*Continued*)

Then, when p and q are evaluated by Eqs. (9-61) and (9-62), the following results are obtained for $\psi_1 = \pi/6$:

$$p_{1a} = 3s, \qquad q_{1a} = 3s. \tag{9-65}$$

$$p_{1b} = 5s, \qquad q_{1b} = 5s, \tag{9-66}$$

where the normalized spatial frequency will be assigned values as follows:

$$s = 0.1, 0.2, 0.3, \ldots, 1.9, 2.0. \tag{9-67}$$

The calculated MTF and PTF values are tabulated in Table 9.II, and the MTF curves are plotted in Fig. 9.4. The OTF, $\hat{O}(s)$, for the two sets of aberration coefficients is also plotted in Figs. 9.5 and 9.6 on a complex plane. (As we point out in Chapter 5 in connection with Fig. 5.10, these are often referred to as *Argand diagrams*.) Because each curve on the complex plane requires different coordinate scales for clear representation at its two ends, it is broken into parts a and b (with some overlap).

Two further sets of values are presented, this time for the OTF direction $\psi = \psi_2 = \pi/3$. The same aberration values, C_{22} and C_{22}/C_{20}, are applied so that comparisons can made with the calculations for the previous OTF direction ψ_1. The Bessel function arguments p and q are necessarily different from those of the previous direction because they are functions of ψ:

$$p_{2a} = (3 + 6/\sqrt{3})s = 3(1 + 2/\sqrt{3})s \cong 6.464101615s$$

$$\cong 6.5s, \tag{9-68}$$

$$q_{2a} = 3s. \tag{9-69}$$

$$p_{2b} = (5 + 10/\sqrt{3})s = 5(1 + 2/\sqrt{3})s \cong 10.77350269s$$

$$\cong 10.8s, \tag{9-70}$$

$$q_{2b} = 5s. \tag{9-71}$$

Table 9.III gives the calculated values for direction $\psi_2 = \pi/3$. The corresponding curves are shown in Figs. 9.7–9.9. As in the previous Argand diagrams, each curve in Figs. 9.8 and 9.9 is broken into parts a and b so that appropriate coordinate scales can be set up for each end of the curve.

REFERENCES

1. H. H. Hopkins, The Frequency Response of a Defocused Optical System. *Proc. R. Soc. London Ser. A* **231,** 91 (1955).
2. M. De, The Influence of Astigmatism on the Response Function of an Optical System. *Proc. R. Soc. London Ser. A* **233,** 91 (1955).
3. G. N. Watson, *Theory of Bessel Functions.* Cambridge University Press, Cambridge, 1958.
4. M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions.* Dover, New York, 1965.
5. S. M. Selby (Ed.), *Standard Mathematical Tables*, 21st ed., The Chemical Rubber Company, Cleveland, 1973.
6. I. S. Gradshteyn and I. M. Ryzhik, *Tables of Integrals, Series, and Products.* Academic, New York, 1980.
7. L. Levi, *Applied Optics: A Guide to Optical System Design*, Vol. 1. Wiley, New York, 1968, p. 591, Table 69.
8. L. Levi and R. H. Austing, Tables of the Modulation Transfer Function of a Defocused Perfect System. *Appl. Opt.* **7,** 967 (1968).

10

Calculation of the OTF: Numerical Methods

INTRODUCTION

In Chapter 9 analytical methods for calculating the OTF are demonstrated by assuming the kinds and the amounts of aberration and then proceeding to calculate the points on the OTF curves that correspond to these assumptions. The kinds and amounts of aberration are expressed in terms of the power series for the wave aberration function, W; in each example the aberration assumptions are made by setting the values of the coefficients in the power series.

In this chapter, rather than starting with assumed values of coefficients in the wave aberration function, we assume that the initial data are in the form of optical path differences observed in laboratory measurements or calculated from ray trace data in an optical design. When one tries to get OTF values from this kind of information, it is soon discovered that analytical approaches are not practical. Instead, numerical methods must be applied to the problem through electronic computers.

Most numerical methods that are productive in calculating OTF values appear to involve three kinds of processes: (1) numerical interpolation, (2) autocorrelation of the pupil function resulting from numerical interpolation, and (3) direct numerical operation, equivalent to evaluating a definite integral, on optical path differences. The third process combines processes 1 and 2 in one operation.

Figure 10.1 represents, with a single independent variable, the problem in *numerical interpolation*. The function y represented by the curve is to be expressed as a power polynomial in x or, more generally, a series of polynomials. In optics the function could be the wave aberration function W in terms of two or three independent variables. In Fig. 10.1 discrete values of y at x_0, x_1, x_2, and x_3 are shown. The complete sequence of known values, $f(x_i)$, can be expressed mathematically as

$$f(x_i) = \sum_{n=0}^{N} c_n x_i^n, \qquad \text{where} \quad x_i = x_0, x_1, x_2, x_3, \ldots . \tag{10-1}$$

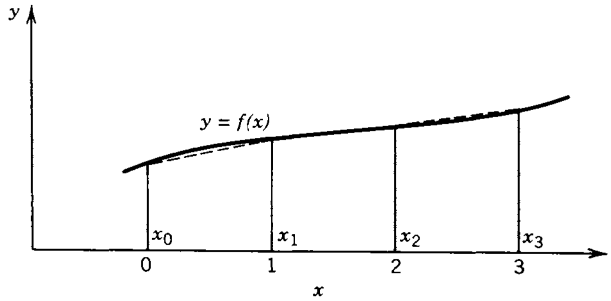


Figure 10.1. Approximation of a curve by a sequence of straight line segments.

The function $f(x)$ is known to have the values $f(x_i)$ at the given abscissas x_i, which are called *interpolation points*. The problem is to determine the constants c_n.

When the numerical interpolation example represented by Fig. 10.1 and Eq. (10-1) is extended to the two-dimensional wave aberration function $W(\rho, \varphi)$, autocorrelation of the pupil function, which involves the wave aberration function, to get the OTF would constitute the second process listed earlier.

The third process, which may be substituted for the combination of the first two processes, can be represented in one dimension by

$$\int_a^b f(x)\,dx = \sum_i H_i f(x_i), \tag{10-2}$$

where the $f(x_i)$ have the same meaning as in Eq. (10-1), and the H_i are weighting factors. When the function $f(x)$ in Eq. (10-2) is represented graphically by the curve in Fig. 10.1, each of the expressions on the two sides of the equation is an expression for the area under the curve from $x = a$ to $x = b$. When the infinite series on the right side is truncated, as it must be in any practical calculation, the summation on the right side is just an approximation of the actual area. This summation process is often referred to as *mechanical quadrature* or just *quadrature* (probably because the approximation is made by one or more quadrilaterals so assembled that the sum of quadrilateral areas approaches the area under the curve). In three dimensions (counting the y-value), the process would be to approximate the volume under a surface and probably be more appropriately termed *cubature*.

To extend the one-dimensional example, Eq. (10-2), to two dimensions, it is replaced by a slightly altered form of the expression in Eq. (5-15), which is the autocorrelation function of the pupil function:

$$\iint_{\mathfrak{a}} \hat{G}(x, y)\, \hat{G}^*[(x - s), (y - t)]\, dx\, dy$$

$$= \sum_{i}^{n} \sum_{j}^{n} H_{ij} \hat{G}(x_i, y_j)\, \hat{G}^*[(x_i - s), (y_j - t)]. \qquad (10\text{-}3)$$

The three processes in numerical methods outlined in the previous paragraphs are discussed in greater detail in the remainder of this chapter. However, the scope of this text does not allow full mathematical rigor in treating numerical methods, nor is it feasible to go into the software subtleties that must be practiced before the three processes can be executed by electronic computers.

Optical path difference data may be obtained experimentally from interferometric measurements on an optical system. In the second section of this chapter, the measurement-calculation procedures to develop these data are discussed in some detail and a number of papers are cited for further review. Emphasis on interferometric methods is largely justified by the modern developments that greatly facilitate the measurement and reduction of data. The laser, for example, provides an intense, monochromatic, stable light source having sufficient temporal coherence to eliminate the tedious requirement of maintaining nearly equal path lengths in the arms of the interferometer. Also, fast-scanning microdensitometers, operating automatically in conjunction with dedicated digital computers, scan and measure the interferogram and reduce the data.

Because the illuminating light beams of most interferometers are nearly coherent, it is appropriate to review the properties of coherent beams that are pertinent to the calculation of the OTF. An excellent paper by Swing [1] for this purpose starts with the theory of partial coherence relating to the imaging properties of optical systems and presents a brief analysis of the propagation of what he defines as the "mutual intensity" and the calculation of the OTF of lenses. (See also a paper by DeVelis and Parrant [2].)

When lenses are cascaded, the transfer function of the combination is principally limited by the pupil function of the poorest lens. In Swing's analysis of a reasonably balanced system, it is apparent that one cannot find the transfer function, in general, simply by calculating the product of the pupil functions of the individual lenses or by calculating the product of the individual transfer functions. When two lenses are cascaded in the finite conjugate mode, as illustrated in Fig. 10.2, with the pupils displaced in the x direction:

$$\hat{O}(s) = (1/A_0) \iint_{\mathfrak{a}} \hat{G}_1\{[x + (s/2)], y\} \cdot \hat{G}_1^*\{[x - (s/2)], y\}$$

$$\cdot\, \hat{G}_2\{[x + (s/2)], y\} \cdot \hat{G}_2^*\{[x - (s/2)], y\}\, dx\, dy, \quad (10\text{-}4)$$

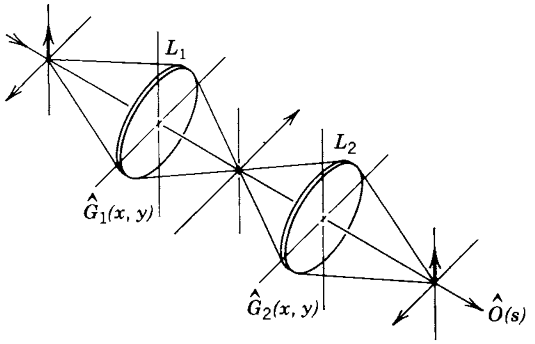


Figure 10.2. Sketch of two lenses cascaded at finite conjugates [1].

where the symbols conform with the notation in Eqs. (9-1) and (9-2). When the system is extended to N lenses similarly cascaded,

$$\hat{O}(s) = (1/A_0) \iint_{\mathcal{A}} \prod_{j=1}^{N} \hat{G}_j \left\{ \left[x + (s/2) \right], y \right\} \cdot \hat{G}_j^* \left\{ \left[x - (s/2) \right], y \right\} dx\, dy. \tag{10-5}$$

Swing's analysis makes evident that the pupil function is the only functional description of a lens suitable for assessing cascaded system response, unless the whole system can be treated as a unit. When a lens is tested, for example, other lenses are usually part of the experimental setup; and if these are not significantly higher in quality than the tested lens, their effects on the overall measurement have to be removed by "backing out" their pupil functions from the integrand of Eq. (10-5). Whatever the quality of the lenses in the measuring apparatus, their characteristics must be known to assure reliable results.

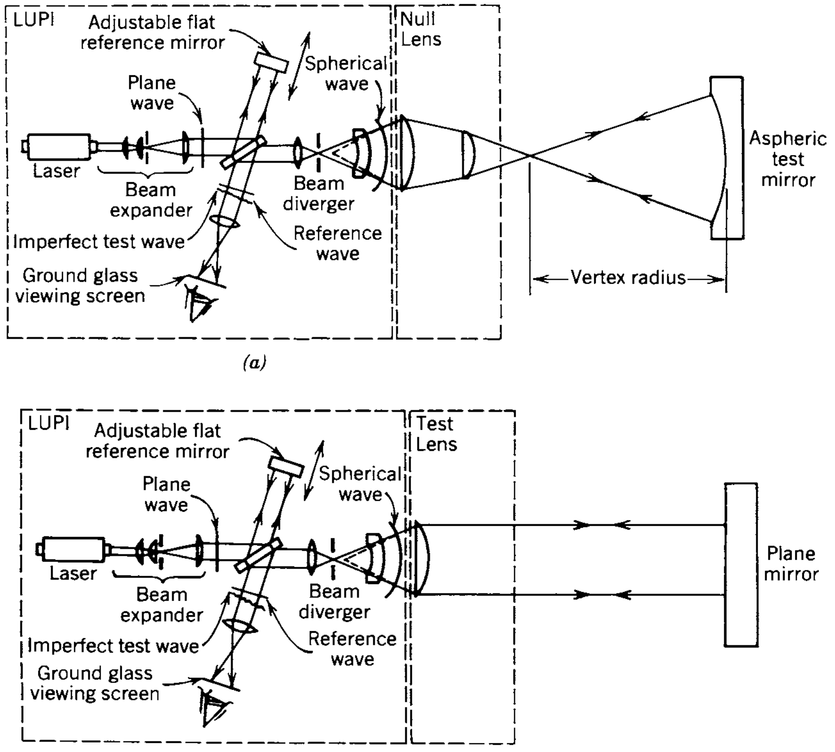
Because the pupil function is recognized, as emphasized by Swing and others, as the fundamental characterization of an optical system, particularly with partially coherent illumination, its measurement is logically the way to get the data for calculating the OTF. Measurement of the pupil function is best done by interferometric methods.

OPTICAL PATH DIFFERENCE DATA BY INTERFEROMETRY

To get optical path difference data of a lens under test, the laboratory worker may employ any one of a number of well-known interferometers, with a variety of modifications, to produce the interference pattern from which the data can be derived. The principles involved are reviewed here by describing the operation of an instrument called a LUPI (Laser Unequal Path Interferometer), which was originally discussed by Minnick and Rancourt in 1968 [3].

LUPI produces interferogram photographs of an interference pattern. Optical path differences are calculated by measuring the fringe spacing in the photographs. The schematic in Fig. 10.3 shows how LUPI is set up to test (*a*) an aspheric mirror and (*b*) a refracting lens. When a refracting lens is to be tested, it is placed in a position to focus at infinity so that its post-refracted rays are ideally parallel. Our discussion of LUPI assumes that a refracting lens is being tested.

A highly coherent beam is produced by the laser, and the beam is expanded to a train of plane wave fronts, actually phase fronts, by the beam expander.



(*a*)

(*b*)

Figure 10.3. An interferometer for measuring optical path differences with (*a*) an aspheric mirror and (*b*) a refracting lens [3].

Each wave front is divided, by division of amplitude, at the beamsplitter; a portion is reflected and another portion is transmitted. The reflected wave fronts become a reference beam and are retroreflected at a flat reference mirror. The angle between the reference mirror surface and the incident wave fronts is adjustable. The reference beam reflected from the reference mirror returns to the beamsplitter and again is divided, a portion being transmitted through it toward a ground glass viewing screen. The second portion of the original expanded wave front that passes through the beamsplitter is caused to diverge as a spherical wave front by the beam diverger. Concentric spherical wave fronts seem to diverge from a virtual point source within the beam diverger. The lens being tested is placed so that its optic axis is colinear with the axis of the interferometer and its object–space focal point coincides with the virtual point source. Spherical wave fronts become plane wave fronts. (The beam is focused at infinity by the lens being tested.) Plane wave fronts are retroreflected at a second plane mirror; the beam then retraces its path to the beamsplitter. A portion is reflected toward the viewing screen where the wave fronts are superposed on the reference wave fronts and interference fringes are produced. A photographic film substituted for the viewing screen records the interferogram. Use of instant film such as high-speed Polaroid positive film facilitates a speedy processing of data.

Linear fringes can be produced by an appropriate adjustment of the reference mirror, and fringe spacing is related to the angle of adjustment between the interfering beams. Circular fringes can also be produced by adjustment of the reference mirror, but linear fringes are preferred for the following reasons:

1. The linear pattern is easily recognized by x–y scanning devices.
2. A reasonably uniform distribution of light flux density is achievable over the aperture.
3. Ambiguities in fringe order, which are possible with circular fringe patterns, are avoided.

With linear fringes, any curvature of a fringe is recognized as an effect caused by a distortion of the wave front. Also, with circular fringes, the zero order is at the center of the pattern, and it must be located, whereas with linear fringes, it is not necessary to locate the zero order.

When fringes are truly linear, that is, when the interfering wave fronts are both perfectly plane, the fringe spacing x is a function of wavelength λ and the angle α between the two interfering wave fronts, as shown diagrammatically in Fig. 10.4. The plane of the diagram is perpendicular to the line of intersection of wave fronts. From the geometry of the figure,

$$\tan \alpha/2 = \lambda/x. \qquad (10\text{-}6)$$

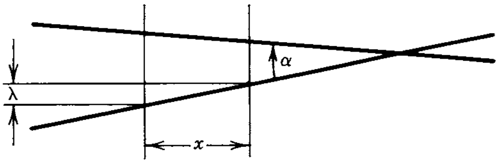


Figure 10.4. Cross section of interfering plane wave fronts.

When "linear fringes" are curved and spacing between fringes is decreasing, an imaginary plane that is tangent to the distorted wave front is making a larger angle α to the reference wave front. When spacing is increasing between curved "linear fringes," the angle α is smaller. Fringe spacing is measured in a direction normal to the curvature of fringes. Thus, from the size and shape of the curved pattern of fringes, which ideally would be linear, and from the spacing between fringes, a contour plot of the distorted wave front can be constructed and optical path difference data deduced.

CALCULATION OF THE ABERRATION POLYNOMIAL

When numerical methods are applied to wave-front-ray-aberration data, a series of points is known; but an expression for the smooth curve that passes through these points is unknown. In optics the wave aberration function, for example, is usually expressed as a series of polynomials such as the right-hand member of Eq. (4-47). The process of evaluating the constants in the polynomial series so that it passes through calculated points is known as *curve fitting* and is discussed by a number of authors including Plight [4] and Barakat [5].

The optical path length D, defined by Eq. (3-4), from the object point along a pupil ray to the exit pupil point E', as defined in Chapter 3, determines the reference path length D_0. When any other ray is extended this same distance D_0 from the object, the end point in image space is a point on the wave front near the exit pupil at the ray's coordinates (x, y) on the reference sphere. The optical path difference (OPD) between the wave front, defined by the D_0 points, and the reference sphere, which also passes through the exit point E', is the *phase advance*, defined as $2\pi n'(D - D_0)/\lambda$. This quantity is discussed in Chapter 3 in connection with Fig. 3.4. The OPD, given in radians, is the value a_i applied in numerical calculations as the wave distortion at (x_i, y_i). It is positive when directed from the reference sphere to the wave front in the direction that the wave is traveling. Values of a_i are called the *wave-front-ray aberrations*.

The data a_i can be made as accurate mathematically as desired during the design optimization procedure, which consists of successive calculation of the

OPD for a number of key rays. However, when the data are obtained experimentally by interferometric measurement of the wave-front shape produced by an actual optical system, the required precision in laboratory methods, as discussed in Chapter 8, is often extremely difficult to attain.

Once the wave-front-ray-aberration data have been obtained, the next task is to select a general mathematical expression that can be made to fit the data in the range of interest. Since the wave aberration function W is the difference between the actual wave front and the reference sphere at the exit pupil, the power series and the Zernike polynomial series representations of W discussed in Chapter 4 are two examples of possible mathematical expressions to fit wave-front-ray-aberration data.

It is of interest to review the elementary beginnings of employing a power series as a general function. The basis of our discussion is the Weierstrass theorem [6]. Though simple in expression, proof of this theorem has involved subtle and highly sophisticated reasoning by a number of eminent mathematicians, including Gauss. The theorem is stated as follows in terms of a single variable [7, p. 19]:

> Let $f(x)$ be an arbitrary continuous function defined in the finite interval $a < x < b$. It is always possible to approximate $f(x)$ over the whole interval (a, b) as closely as we please by a power series in x of sufficiently high degree.

In equation form, the theorem states

$$f(x) = c_0 + c_1x^1 + c_2x^2 + c_3x^3 + \cdots + c_nx^n, \tag{10-7}$$

where n is a sufficiently high degree of the polynomial and the c_n are coefficients. (Ordinarily the exponent 1 is omitted and x^1 is simply written x; on the other hand, if the exponential continuity is to be emphasized, the first right-hand term could be written c_0x^0 with no change in the meaning of the expression.) If the values of $f(x)$ are known at certain discrete values of x, either by experimental determination or by calculations in a theoretical model, we can state this by

$$f(x_i) = a_i, \tag{10-8}$$

and for each known value of a_i, there is an equation

$$f(x_i) = a_i = c_0 + c_1x_i^1 + c_2x_i^2 + c_3x_i^3 + \cdots + c_nx_i^n. \tag{10-9}$$

In expanded form, Eq. (10-9) is

$$\begin{aligned}
f(x_0) &= c_0 + c_1x_0^1 + c_2x_0^2 + c_3x_0^3 + \cdots + c_nx_0^n,\\
f(x_1) &= c_0 + c_1x_1^1 + c_2x_1^2 + c_3x_1^3 + \cdots + c_nx_1^n,\\
f(x_2) &= c_0 + c_1x_2^1 + c_2x_2^2 + c_3x_2^3 + \cdots + c_nx_2^n,\\
&\vdots\\
f(x_n) &= c_0 + c_1x_n^1 + c_2x_n^2 + c_3x_n^3 + \cdots + c_nx_n^n.
\end{aligned} \qquad (10\text{-}10)$$

Since $n + 1$ discrete values, "interpolation points," were chosen for x to find corresponding values of $f(x)$, we recognize in Eq. (10-10) a set of $n + 1$ simultaneous linear equations in c where the "constants" represented by c are paradoxically the variables to be determined by simultaneous solution. Of course, as with all such sets of equations, certain conditions of independence and freedom from inconsistencies must hold, for which the reader is referred to a standard discussion of simultaneous linear equations and their solution.

Since the Weierstrass theorem places no restrictions on the chosen values of x_i, that is, the "positions" of the interpolation points, some thought has to be given to how best to choose these points with accuracy and convenience of calculation in mind. Because in the optical application of quadrature theory we are ultimately concerned with values of the OTF, it seems obvious that we would like to relate the method of choosing the interpolation points to some accuracy criterion of the OTF. Gauss quadrature ranks high among optical practitioners as a basis for effective computer software to find the OTF from OPD data. Although our present scope does not allow exploring the intricacies of such software development, a following section does outline the principles of Gauss quadrature by illustrating the procedure for one independent variable (instead of the two or three independent variables usually involved in optical computations).

EXTENSION TO MORE THAN ONE INDEPENDENT VARIABLE

If we chose to write Eq. (10-7) in summation form, it would appear as

$$f(x) = \sum c_j x^j, \qquad (10\text{-}11)$$

and Eq. (10-10) would be written

$$f(x_i) = \sum_i \sum_j c_j x_i^j. \qquad (10\text{-}12)$$

With similar notation, Eq. (4-33) of Chapter 4, with r treated as a constant, could be written

$$W(\rho, \varphi) = \sum c_{jk} \rho^j \cos^k \varphi. \tag{10-13}$$

If a number of observations a_i of W are made, then

$$a_i = \sum_{i=0}^{n} c_{ijk} \rho_i^j \cos^k \varphi_i, \tag{10-14}$$

and

$$W(\rho_i, \varphi_i) = \sum_{i} \sum_{jk} c_{ijk} \rho_i^j \cos^k \varphi_i, \tag{10-15}$$

where certain limitations are placed on j and k in accordance with Eqs. (4-35)–(4-38). As with the single independent variable, with sufficient observations of a_i, the c_{ijk} may be solved for and $W(\rho, \varphi)$ determined to any accuracy desired.

CHOICE OF ORTHOGONAL POLYNOMIAL

Setting up an expression for $W(\rho, \varphi)$, as discussed in the previous section, is only the first step in finding the OTF. Following that, the autocorrelation integral must be solved, probably by numerical methods. In fact, the form chosen for $W(\rho, \varphi)$ depends in large degree upon the subsequent mathematical operations to calculate the OTF. Those who have labored over this problem have come up with various infinite series of orthogonal polynomials to represent $W(\rho, \varphi)$, which, of course, are truncated to a manageable number of polynomials for actual calculation. Orthogonality has already been discussed in Chapter 4 where Eqs. (4-42) and (4-43) define the property. Also in Chapter 4, the Zernike circle polynomials are treated in detail to give the reader a feeling for how the orthogonal polynomials are applied in general to the wave aberration function $W(\rho, \varphi)$.

Besides being orthogonal, the interpolation polynomials have a number of other properties in common. Not only should each polynomial series represent the wave front, but it should be able to give the shape of the wave front relative to the reference sphere everywhere over the exit pupil. This requires the series to have simple continuity and to be single valued in the interval defined by the boundary of the exit pupil. With rectangular coordinates (x, y), as defined in Chapter 3, practical optical problems require

$$0 \leqq \left| W(x, y) \right| \leqq 2n\pi \qquad \text{for} \quad (x^2 + y^2) \leqq 1. \tag{10-16}$$

The integer n is assumed to be small, say less than 5 in practical optics, and

$$W(x, y) = 0 \qquad \text{for} \quad (x^2 + y^2) > 1. \tag{10-17}$$

In polar coordinates (ρ, φ)

$$0 \leqq \left| W(\rho, \varphi) \right| \leqq 2n\pi \qquad \text{for} \quad \rho \leqq 1 \tag{10-18}$$

and

$$W(\rho, \varphi) = 0 \qquad \text{for} \quad \rho > 1. \tag{10-19}$$

It must be possible to set the polynomial series equal to the known value a_i of the wave-front-ray aberration for each of a finite number of interpolation points (x_i, y_i) or (ρ_i, φ_i).

No one series of orthogonal polynomials can be declared best for the calculation of the OFT from a set of interpolation points. We have already referred to the discussion of Zernike polynomials in Chapter 4 where the general expression for the series is given as Eq. (4-47) and the first 25 Zernike radial polynomials are shown in Table 4.II. Hawkes [8] has carried out an expansion of the wave front in Zernike circle polynomials, which are the most widely used for analyzing diffraction integrals.

Another powerful set for optical calculations is the series of Tschebyscheff polynomials discussed by a number of authors [4]. Tschebyscheff polynomials of the first kind can be arrived at by a recurrence formula when two of three successive polynomials have been found by other means:

$$T_{n+1}(\rho) = 2\rho T_n(\rho) - T_{n-1}(\rho). \tag{10-20}$$

As indicated by the single independent variable, the radial coordinate ρ, the Tschebyscheff polynomials alone can be applied to expressing $W(\rho, \varphi)$ only when there is rotational symmetry. Otherwise each polynomial must be used in combination with another function [4, 9]:

$$Q_p^q(\rho, \varphi) = T_q(\rho) \cos p\varphi, \tag{10-21}$$

where $q \geqq p$, and $|q - p|$ is even. The first 13 Tschebyscheff polynomials are tabulated in Table 10.I.

Legendre orthogonal polynomials result from application of the most precise

Table 10.I The First Thirteen Tschebyscheff Polynomials

$T_0(x) = 1$
$T_1(x) = x$
$T_2(x) = -1 + 2x^2$
$T_3(x) = -3x + 4x^3$
$T_4(x) = 1 - 8x^2 + 8x^4$
$T_5(x) = 5x - 20x^3 + 16x^5$
$T_6(x) = -1 + 18x^2 - 48x^4 + 32x^6$
$T_7(x) = -7x + 56x^3 - 112x^5 + 64x^7$
$T_8(x) = 1 - 32x^2 + 160x^4 - 256x^6 + 128x^8$
$T_9(x) = 9x - 120x^3 + 432x^5 - 576x^7 + 256x^9$
$T_{10}(x) = -1 + 50x^2 - 400x^4 + 1120x^6 - 1280x^8 + 512x^{10}$
$T_{11}(x) = -11x + 220x^3 - 1232x^5 + 2816x^7 - 2816x^9 + 1024x^{11}$
$T_{12}(x) = 1 - 72x^2 + 840x^4 - 3584x^6 + 6912x^8 - 6144x^{10} + 2048x^{12}$

of the many quadrature formulas, the method of Gauss, which is discussed in the next section.

One way of defining the Legendre polynomials is by means of a *generating function* $H(x, r)$ [12, p. 45]:

$$H(x, r) = 1/(1 - 2xr + r^2)^{1/2}. \qquad (10\text{-}22)$$

If this is regarded as a function of r, the right side can be expanded in a power series for sufficiently small values of the variable. The coefficients of the powers of r will be polymomials in x, which we designate as $P_n(x)$ where $n = 0, 1, 2, \ldots$ so that

$$H(x, r) = P_0(x) + P_1(x)\, r + P_2(x)\, r^2 + \cdots . \qquad (10\text{-}23)$$

The polynomial coefficients are the Legendre polynomials, the first five of which are given in Table 10.II. For calculation of further Legendre polynomials, the

Table 10.II The First Five Legendre Polynomials

$P_0(x) = 1$
$P_1(x) = x$
$P_2(x) = (1/2)(3x^2 - 1)$
$P_3(x) = (1/2)(5x^3 - 3x)$
$P_4(x) = (1/8)(35x^4 - 30x^2 + 3)$

recurrence formula may be applied:

$$(n + 1) P_{n+1}(x) - (2n + 1) xP_n(x) + nP_{n-1}(x) = 0. \qquad (10\text{-}24)$$

GAUSS QUADRATURE

More needs to be said about the process of quadrature, sometimes called *approximate quadratures* [13], which is the numerical integration already referred to in the previous pages for proceeding from optical path difference data to the OTF.

When a function $f(x)$ is known to be continuous over an interval of x from a to b but the explicit form of $f(x)$ is unknown, numerical evaluation of its integral depends on substituting a second integral for the original:

$$\int_a^b f(x)\, dx \quad \text{is replaced by} \quad \int_a^b \phi(x)\, dx, \qquad (10\text{-}25)$$

where $\phi(x)$ can be determined in a simple way. Since $f(x)$ is known to have $(n + 1)$ values $y_0, y_1, y_2, \ldots, y_n$ at the $(n + 1)$ interpolation points within the interval (a, b), the second integral may be expressed as

$$\int_a^b \phi(x)\, dx = C_0 y_0 + C_1 y_1 + C_2 y_2 + \cdots + C_n y_n, \qquad (10\text{-}26)$$

where the $(n + 1)$ quantities C_i are independent of the $(n + 1)$ values of the y_i. Therefore, if $f(x)$ is a polynomial of degree $\leqq n$, the error made in replacing the original integral by $\Sigma\, C_i y_i$ may be made to vanish by the proper choice of the C_i. Even if $f(x)$ is a polynomial of degree $> n$, the difference between the true value of the integral and the value in Eq. (10-26) may still be small enough to make this procedure useful.

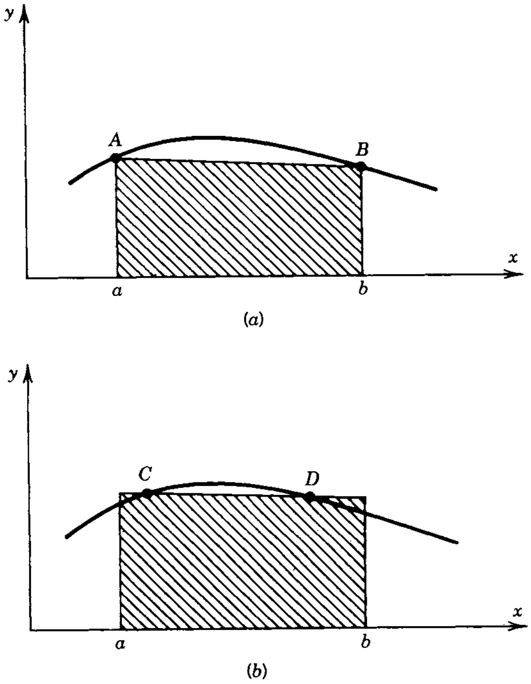
Many formulas have been developed to evaluate $\int_a^b f(x)\, dx$, most of them assuming the y_i to be known at equal intervals over the range (a, b). Among these are the *Euler–Maclaurin formula, Gregory's formula*, and the *Newton–Cotes formula*, which subdivides into the *Trapezoidal rule, Simpson's rule*, and *Weddle's rule* [13].

The *method of Gauss* or *Gauss quadrature* not only determines the $(n + 1)$ values of C_i but also fixes the $(n + 1)$ y_i of Eq. (10-26) in such a way as to make the difference between the two integrals given in Eq. (10-25) a minimum. As a result, there are in effect $(2n + 2)$ constants available so that if $f(x)$ is a polynomial of degree $\leqq (2n + 1)$ the method will give an exact result for the integral. Because of this characteristic, Gauss's method in general introduces less error than other quadrature methods for a given number of y_i.

The following elementary procedure is presented as a portal to the subject of Gauss quadrature. For complete mathematical processes, rigor, and preparation of computer software, the reader is referred to advanced texts like those listed at the end of this chapter. (See particularly [7, 11, 14, 15].)

An article by R. Barakat [15] discusses many of the mathematical procedures that are mentioned, but not explained, in this chapter. The article treats the procedures—which range from the trapezoidal rule and Simpson's rule to Gauss quadrature—and algorithms for applying the procedures to computer software from the viewpoint of optical design and analysis. Particular algorithms for a number of useful optical calculations are given.

Perhaps the simplest quadrature approach is illustrated in Fig. 10.5*a* where a trapezoid *abBA* has been constructed as a first approximation to the area under the curve $f(x)$ between the abscissa values a and b. As indicated point A on the curve is at $(a, f(a))$ and point B at $(b, f(b))$; and the trapezoid is com-



(*a*)

(*b*)

Figure 10.5. Trapezoid approximations of the area under a curve.

pleted by drawing the straight line AB connecting the two points. Since the width $w = (b - a)$,

$$\text{area} = (w/2)[f(a) + f(b)]. \tag{10-27}$$

This expression can be put in the form of Eq. (10-26):

$$\text{area} = (w/2)f(a) + (w/2)f(b). \tag{10-28}$$

It is obvious that this approximation of the area under the curve would be unsatisfactory as a general approach. It is true that the approximation could be improved by subdividing the interval (a, b) and adding the areas of the succession of trapezoids; but a far more effective way of improving the approximation, without requiring a large increase in the number of interpolation points, is to substitute points C and D (Fig. 10.5b) for A and B to define the approximation trapezoid. In fact, it is apparent that if these points are so chosen that the trapezoid area above the curve exactly equals the void under the curve, this method of calculating the area under the curve from a to b is exact. The Gauss method consists essentially of how to choose C and D to get the optimum approximation.

The Gauss method as developed requires that the integral be arbitrarily taken over the interval $(-1, +1)$ instead of the general interval (a, b) as previously indicated. However, this requirement can be handled by appropriate change of variable as indicated in the following example:

$$\begin{aligned}\int_0^\infty \exp(-t)\,dt &= \left[-\exp(-t)\right]_0^\infty \\ &= -\exp(-\infty) + \exp(0) \\ &= 0 + 1 = 1. \end{aligned} \tag{10-29}$$

To change the variable to realize the $(-1, +1)$ limits, let $t = (x + 1)/(x - 1)$. Then, when

$$t = 0, \qquad x = -1, \tag{10-30}$$

and when

$$t = \infty, \qquad x = +1, \tag{10-31}$$

and

$$dt = \left[(x-1)\,dx - (x+1)\,dx\right]/(x-1)^2. \tag{10-32}$$

So

$$\int_0^{\infty} \exp(-t)\,dt = -2\int_{-1}^{+1} \left\langle \left\{\exp[-(x+1)/(x-1)]\right\}/(x-1)^2 \right\rangle dx. \tag{10-33}$$

Having indicated with a particular example an approach for changing the integration limits to the required $(-1, +1)$, we turn to the general problem of evaluating the resulting definite integral, which is represented by Fig. 10.6. The trapezoid *EHGF* corresponds to the trapezoid of Fig. 10.5 discussed earlier. As indicated, point *C* on the curve is at $(x_1, f(x_1))$, and point *D* is at $(x_2, f(x_2))$. Our object is to develop a rationale for a Gauss quadrature formula based on the general type given as Eq. (10-26):

$$\int_{-1}^{+1} \phi(x)\,dx = C_0\phi(x_0) + C_1\phi(x_1). \tag{10-34}$$

As indicated earlier, most of the quadrature formulas work with equal intervals between $x_0, x_1, x_2, \ldots, x_n$; but the Gauss method requires specific inter-

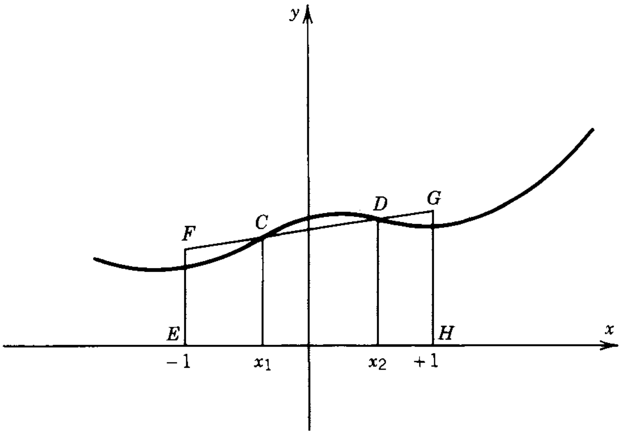


Figure 10.6. Optimizing the trapezoid area approximation by choosing x_1 and x_2 within the interval from -1 to $+1$.

vals, which are determined as part of developing the method. Therefore, in Eq. (10-34) there are four unknowns: C_0, C_1, x_0, and x_1. Also, one has to make some assumptions as to the nature of the ϕ functions. To do this, some philosophizing is in order as to how best to connect interpolation points to fit the unknown curve between the points. The easiest technique is simply to connect the points with straight lines; but we have already seen that the resulting trapezoids under the curve are likely, in most instances, to be unsatisfactory. Some sort of curved connecting line, whose shape is influenced by the interpolation point values, would seem to be an improvement. About as simple as we can go in this direction is to employ parabolas, that is, functions made up of powers of x. We thus can arbitrarily write the four independent equations needed to solve for the four unknowns in Eq. (10-34):

$$\begin{aligned} \phi_1(x) &= x^0 = 1, \\ \phi_2(x) &= x^1 = x, \\ \phi_3(x) &= x^2, \\ \phi_4(x) &= x^3. \end{aligned} \tag{10-35}$$

When these assumed functions are individually substituted in Eq. (10-34),

$$\begin{aligned} \int_{-1}^{+1} (1)\, dx &= [x]_{-1}^{+1} = 2 = C_0(1) + C_1(1), \\ \int_{-1}^{+1} x\, dx &= [x^2/2]_{-1}^{+1} = 0 = C_0x_0 + C_1x_1, \\ \int_{-1}^{+1} x^2\, dx &= [x^3/3]_{-1}^{+1} = 2/3 = C_0x_0^2 + C_1x_1^2, \\ \int_{-1}^{+1} x^3\, dx &= [x^4/4]_{-1}^{+1} = 0 = C_0x_0^3 + C_1x_1^3. \end{aligned} \tag{10-36}$$

From the first, second, and fourth of Eq. (10-36),

$$\begin{aligned} C_0 + C_1 &= 2, \\ C_0x_0 + C_1x_1 &= 0, \\ C_0x_0^3 + C_1x_1^3 &= 0, \end{aligned} \tag{10-37}$$

It is apparent that the following will satisfy the simultaneous equations in Eq. (10-37):

$$C_0 = C_1 = 1,$$
$$x_0 = -x_1. \tag{10-38}$$

When these values are substituted in the third of Eq. (10-36),

$$x_0^2 + x_1^2 = 2/3,$$
$$x_0^2 = 1/3,$$
$$x_0 = \pm 0.5773502692. \tag{10-39}$$

So, in general, from Eq. (10-34),

$$\int_{-1}^{+1} \phi(x)\, dx = \phi(-0.57735) + \phi(0.57735). \tag{10-40}$$

The Gauss method can be extended to three or more points instead of the two in the example above. As the number of points within the interval $(-1, +1)$ increases, accuracy increases; but the labor of solving simultaneous equations also increases. The latter job is considerably lightened by relations in-

Table 10.III Gauss Quadrature Coefficients and Abscissas

Number of Points	Coefficients C_i	Abscissas x_i
2	$C_0 = C_1 = 1.0$	$-x_0 = x_1 = 0.57735\ 02692$
3	$C_0 = C_2 = 0.555555\ldots$ $C_1 = 0.888888\ldots$	$-x_0 = x_2 = 0.77459\ 66692$ $-x_1 = 0.0$
4	$C_0 = C_3 = 0.34785\ 48451$ $C_1 = C_2 = 0.65214\ 51549$	$-x_0 = x_3 = 0.86113\ 63116$ $-x_1 = x_2 = 0.33998\ 10436$
5	$C_0 = C_4 = 0.23692\ 68851$ $C_1 = C_3 = 0.47862\ 86705$ $C_2 = 0.568888\ldots$	$-x_0 = x_4 = 0.90617\ 98459$ $-x_1 = x_3 = 0.53846\ 93101$ $-x_2 = 0.0$
6	$C_0 = C_5 = 0.17132\ 44924$ $C_1 = C_4 = 0.36076\ 15730$ $C_2 = C_3 = 0.46791\ 39346$	$-x_0 = x_5 = 0.93246\ 95142$ $-x_1 = x_4 = 0.66120\ 93865$ $-x_2 = x_3 = 0.23861\ 91861$

volving Legendre polynomials. When the Gauss method is extended to n points, the general equation corresponding to Eq. (10-34) is

$$\int_{-1}^{+1} \phi(x)\, dx = \sum_{i=0}^{n} C_i \phi(x_i), \qquad (10\text{-}41)$$

and the general equation corresponding to Eq. (10-36) is

$$\int_{-1}^{+1} x^m\, dx = \sum_{i=0}^{n} C_i x_i^m, \qquad (10\text{-}42)$$

where $m = 0, 1, 2, 3, \ldots, n$. Values of C_i and x_i are tabulated in Table 10.III for up to six interpolation points. Values for a greater number of points are available from various sources; for example, see the appendices of Kopal [7].

REFERENCES

1. R. E. Swing, The Case for the Pupil Function, *SPIE Proc.*, **46,** 104 (1974). (Please see the note following Ref. 2 of Chapter 1.)
2. J. DeVelis and G. B. Parrent, Transfer Function for Cascaded Optical Systems. *J. Opt. Soc. Am.* **57,** 1486(1967).
3. W. A. Minnick and J. D. Rancourt, Transfer Function Calculation Techniques for Real Optical Systems. *SPIE Proc.*, **13,** 87 (1968). (Please see the note following Ref. 2 of Chapter 1.)
4. A. M. Plight, The Rapid Calculation of the Optical Transfer Function for On-Axis Systems Using the Orthogonal Properties of Tchebycheff Polynomials. *Opt. Acta* **9,** 849(1978).
5. R. Barakat, Computation of the Transfer Function of an Optical System from the Design Data for Rotationally Symmetric Aberrations, I. Theory. *J. Opt. Soc. Am.* **52,** 985(1962).
6. K. W. T. Weierstrass, Uber die Analytische Functionen einer reelen Veränderlichen. *Sitzungeber. Akad. Berlin*, 633–789 (1885).
7. Z. Kopal, *Numerical Analysis*. Chapman & Hall, London, 1955.
8. P. W. Hawkes, The Diffraction Theory of the Aberration of Stigmatic Orthomorphotic Optical or Electronic Optical Systems Containing Toric Lenses or Quadropoles. *Opt. Acta* **11,** 237(1964).
9. B. Tatian, Aberration Balancing in Rotationally Symmetric Lenses. *J. Opt. Soc. Am.* **64,** 1083(1974).
10. M. De, L. N. Hazra, and P. K. Purkait, Walsh Functions in Lens Optimization, I.

FEE-Based Criterion. *Opt. Acta* **25,** 573(1978). II. Evaluation of the Diffraction-Based OTF for On-Axis Imagery. *Opt. Acta* **28,** 389(1981).

11. C. Lanczos, *Applied Analysis.* Prentice-Hall, Englewood Cliffs, N.J., 1956.
12. D. Jackson, *Fourier Series and Orthogonal Polynomials.* The Mathematical Association of America. The Collegiate Press, George Banta, Menasha, Wisc., 1941.
13. H. Margenau and G. M. Murphy, *The Mathematics of Physics and Chemistry,* 2d ed. Van Nostrand, New York, 1976.
14. M. J. Kidger, The Calculation of the Optical Transfer Function Using Gaussian Quadrature. *Opt. Acta* **25,** 665(1978).
15. R. Barakat, The Calculation of Integrals Encountered in Optical Diffraction Theory. In *The Computer in Optical Research: Methods and Applications,* B. R. Frieden (Ed.), Springer-Verlag, Berlin, 1980.

Appendix A

Calculated Optical Transfer Functions

INTRODUCTION

Optical Transfer Functions for optical systems having primary and secondary aberrations have been calculated for most of the classical Seidel aberrations. Published papers reporting these calculations have often been referenced in chapters of this book. Here, in Appendix A, we summarize results of the calculations by showing typical curves of the OTF and other pertinent results.

The reader will notice that certain works and parts of some works are not included. Omissions occur because we arbitrarily adopted the following criteria for including a set of calculated OTF curves:

1. A set of included curves extends only to small amounts of maximum wave aberration according to whether it is in either
   (a) a case where the MTF drops below 0.5 at the normalized spatial frequency of 0.1, or
   (b) a case of approximately 2 wavelengths of maximum wave aberration.
2. Calculations reported in a published paper must have been proved correct subsequent to the date of publication.
3. Geometrical approximations of the OTF are not included.
4. ''Heterochromic OTFs'' are not included.

Exceptions to these rules are allowed if additional curves show a trend when appropriate parameters vary.

DEFOCUSING

Figure A.1 shows the results obtained by H. H. Hopkins [1] for an optical system free of aberration but having a defect of focus. The method of calculation is discussed in Chapter 9 of the text. The curves have been replotted with an expanded scale along the abscissa so that they conform more closely to the

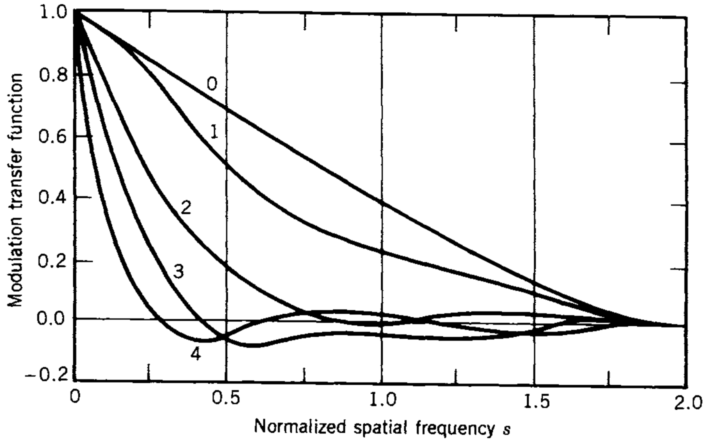


Figure A.1. MTF curves for a defocused optical system that is free of aberrations. The amount of defocusing is indicated by the number on each curve. $C_{20} = n\lambda/\pi$. (Reconstructed by changing the abscissa scale of Fig. 5.9, which is Fig. 5 of Ref. 1.)

style of the rest of the curves in this appendix. The number n on a curve is for (n/π) wavelengths of maximum wave aberration caused by defocusing. The aberration produced by defocusing is radially symmetrical; therefore, there is no phase transfer function.

PRIMARY SPHERICAL ABERRATION

Figures A.2–A.4 show OTFs for primary spherical aberration as calculated, using Simpson's rule to evaluate the integral, by Black and Linfoot [2]. The symbol C is the coefficient ${}_0C_{40}$ in the expression for spherical aberration as given in Chapter 4 of the text. It represents the amount of fourth power spherical wave aberration, and it would be given by

$$W(\rho, \varphi) = C_{40}\rho^4. \tag{A-1}$$

The aberration is radially symmetrical so that there is no phase transfer function. Calculations were made considering different image planes as denoted by the symbol B and shown in the figures. Black and Linfoot introduce a defocusing term as

$$W(\rho, \varphi) = C(\rho^4 + B\rho^2). \tag{A-2}$$

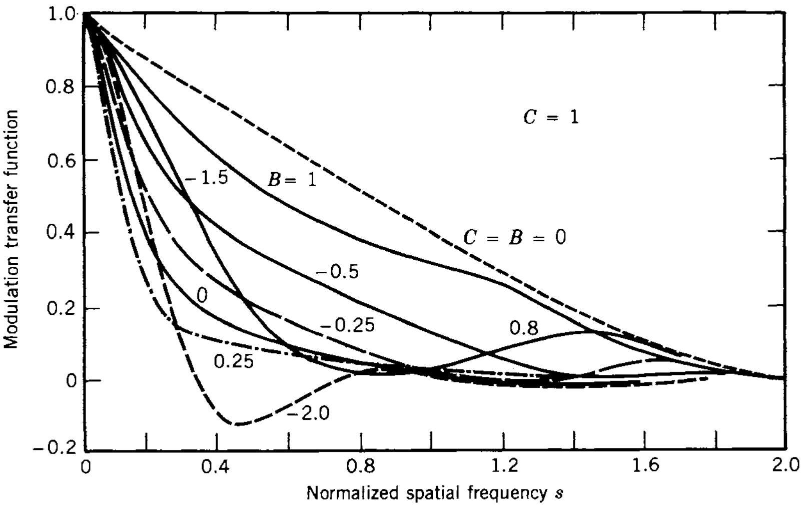


Figure A.2. MTF curves for an optical system having spherical aberration $C_{40} = 1$ at selected focal settings. See the text for an explanation of focal settings, which are indicated by B [2].

Then with the paraxial focus arbitrarily chosen as the "image" plane for study, they set $B = 0$ and $C = 1$. The maximum wave distortion is 1 wavelength and the maximum occurs at the edge of the pupil where $\rho = 1$. They allow B to take on other values denoting different image planes for study. At any image plane, the reference sphere and the wave front coincide on axis; for example,

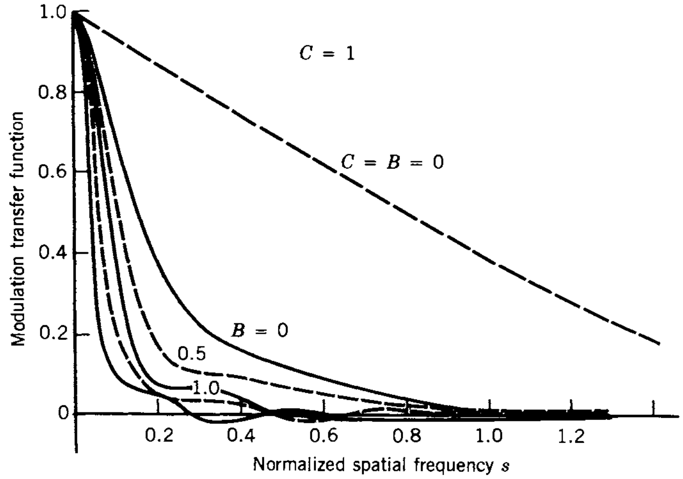


Figure A.3. Same as Fig. A.2 but with different focal settings [2].

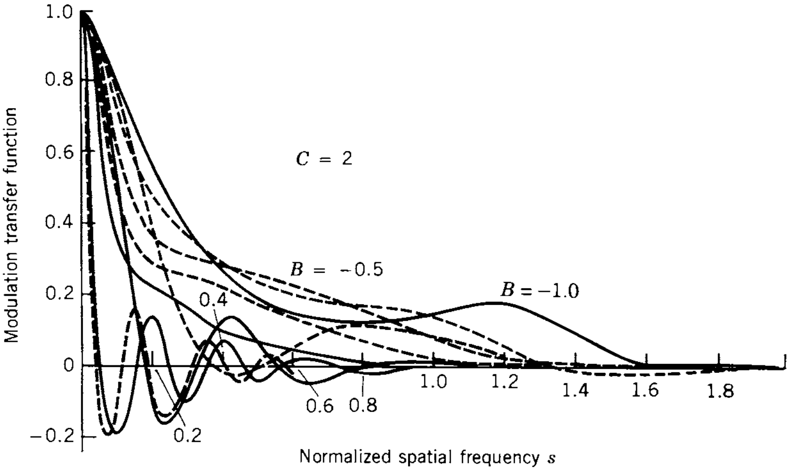


Figure A.4. Same as Fig. A.2 but with greater aberration equal to $C_{40} = 2$ [2].

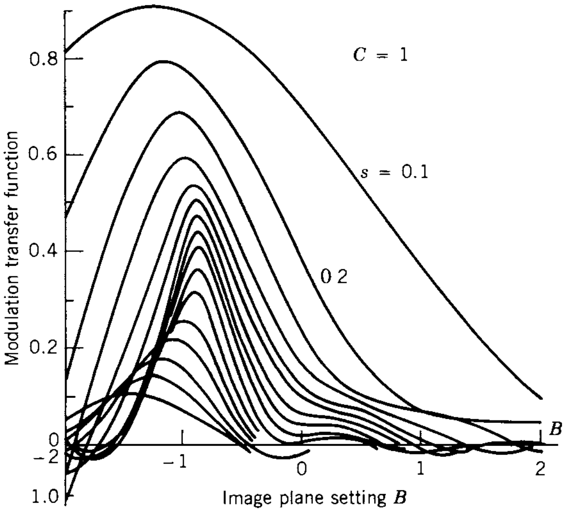


Figure A.5. Curves showing the dependence of the MTF on focal setting for a system having $C_{40} = 1$ at several normalized spatial frequencies s. The symbols have the same meaning as in Figs. A.2-A.4 [2].

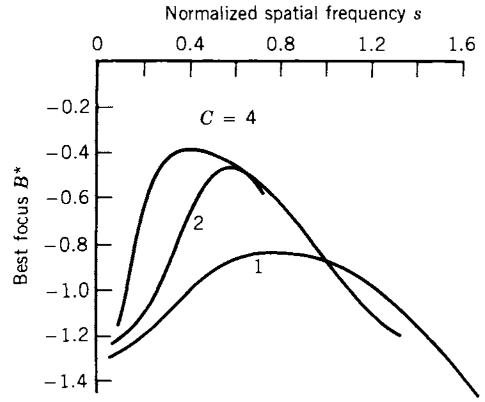


Figure A.6. Variation of best focus B^* with frequency for a system with selected amounts of aberration, $C_{40} = 1, 2$, and 4 [2].

when

$B = -2$, maximum distortion is -2 wavelengths, and the image plane is at marginal focus;

$B = -1$, distortion, at $\rho = 1$, is 0;

$B = 0$, image plane is at paraxial focus, and maximum distortion is $+1$; and when

$B = +2$, maximum wave distortion is 3 wavelengths, and the image plane is beyond the paraxial focus, that is, nearer the exit pupil.

Intermediate values of B are allowed as shown in the figures.

Figure A.5 is from Black and Linfoot showing the MTF as function of B, image plane setting, for fixed values of normalized spatial frequency. Maximum of each curve shows the image plane setting B to give peak MTF for $C = 1$, and the indicated value of normalized spatial frequency. There is no image plane setting to give peak MTF at all frequencies, which would be of interest with a heterochromic beam of light.

Figure A.6 shows the variation of the image plane position B^* giving best focus, that is, maximum MTF, with variation of normalized spatial frequency. Numbers on the curves are values of C.

PRIMARY WITH SECONDARY SPHERICAL ABERRATION

A. M. Goodbody [3] calculated OTFs for the cases of defocusing and primary spherical aberration each in the presence of secondary spherical aberra-

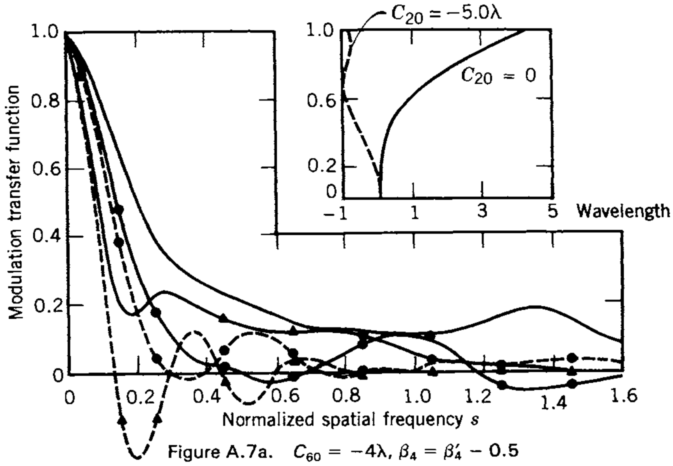


Figure A.7a. $C_{60} = -4\lambda$, $\beta_4 = \beta_4' - 0.5$

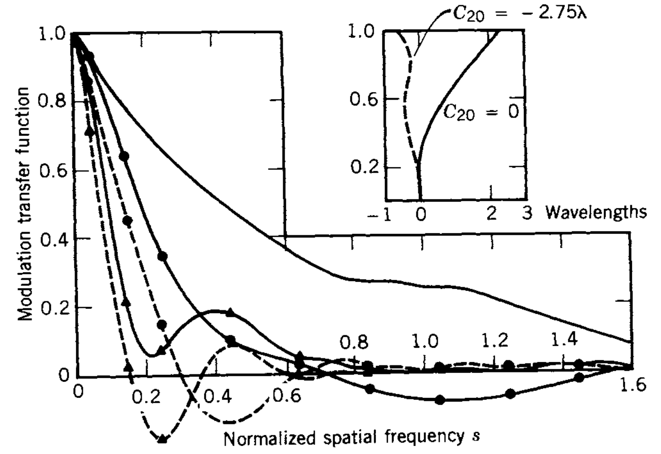


Figure A.7b. $C_{60} = -4\lambda$, $\beta_4 = \beta_4'$

Figure A.7. and Figure A.8. MTF curves in the presence of primary, C_{40}, and secondary, C_{60}, spherical aberration in various focal planes. In the upper diagram in each figure the shape of the wave fronts are compared, one (---) in the best focal plane and another (——) at paraxial focus. The ratio of defocusing to primary aberration is $(C_{20}/C_{40}) = \beta_2$, and the ratio of primary to secondary $(C_{40}/C_{60}) = \beta_4$. The *b* figure in each set of three represents the optimum balance of higher order aberration; when values of the β's optimize the response according to the tolerance theory of Hopkins, primed symbols are used. The legend for the MTF curves accompanies the c figure in each set [3]. (Reproduced by permission of The General Electric Co., Ltd., of England.)

tion. The method used was a numerical integration of the pupil function, which had been expressed in double integrals following an expansion in a Taylor's series by a method suggested by Hopkins [4]. Goodbody expressed the wave aberration function by

$$W(x, y) = C_{20}(x^2 + y^2) + C_{40}(x^2 + y^2)^2 + C_{60}(x^2 + y^2)^3, \quad \text{(A-3)}$$

which corresponds to our expression in Chapter 4 of the text. Calculations were made for different amounts of C_{20} and C_{40} and for different ratios of primary to

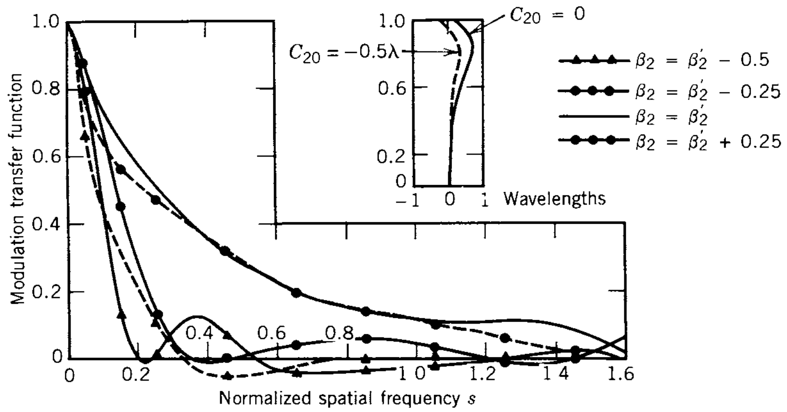


Figure A.7c. $C_{60} = -4\lambda$, $\beta_2 = \beta_2' + 0.5$

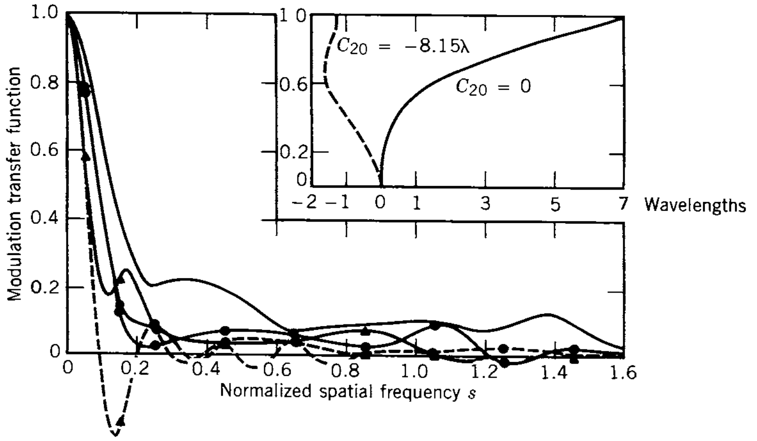


Figure A.8a. $C_{60} = -6\lambda$, $\beta_4 = \beta_4' - 0.5$

secondary aberrations, which he defined as β; that is,

$$\beta_2 = (C_{20})/(C_{40}); \qquad \beta_4 = (C_{40})/(C_{60}). \qquad \text{(A-4)}$$

When the ratio optimizes the response (MTF) according to a tolerance theory of Hopkins [5], primed symbols, β_2' and β_4', are used. A plot of wavelength distortion, in wavelengths, accompanies each set of MTF curves, which are presented in Figs. A.7 and A.8.

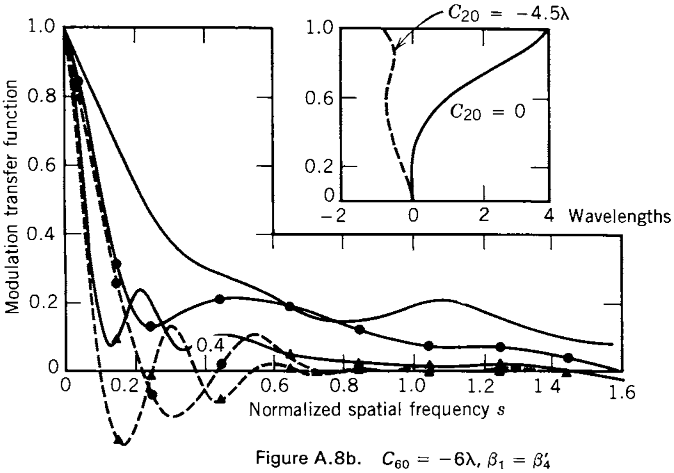


Figure A.8b. $C_{60} = -6\lambda$, $\beta_1 = \beta_4'$

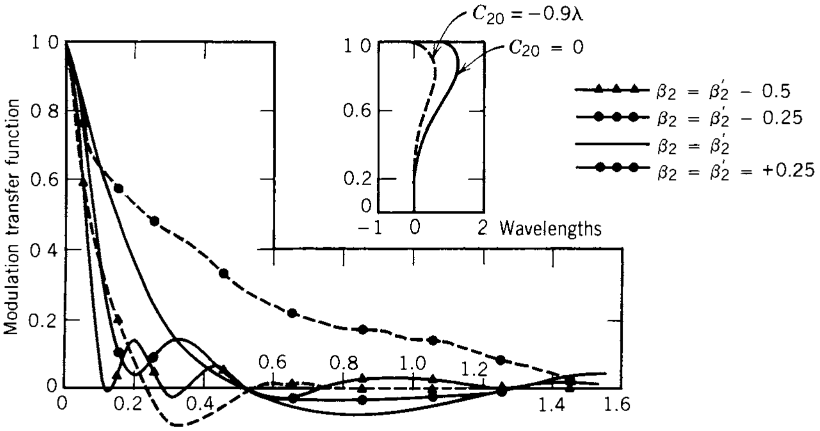


Figure A.8c. $C_{60} = -6\lambda$, $\beta_2 = \beta_2' + 0.5$

PRIMARY AND SECONDARY COMA WITH DEFOCUSING

In a second paper [6] Goodbody considers primary and secondary coma. Different positions of the image plane, the state of balance with higher order aberration, and different azimuths of the line structure in the object are considered. The wave aberration function that he used is

$$W(x, y) = C_{20}(x^2 + y^2) + C_{31}(x^2 + y^2)(y \cos \psi + x \sin \psi) + C_{51}(x^2 + y^2)^2(y \cos \psi + x \sin \psi), \qquad \text{(A-5)}$$

which corresponds only roughly to expressions given in Chapter 4. The symbol ψ in Eq. (A-5) is the angle of the spatial frequency line structure to the meridian plane of the optical system as shown in Fig. 5.4.

The results for primary coma, at different image positions (defocusing), and values 0 and $\pi/2$ for the azimuth are shown in Figs. 4.17–4.19 in the text.

Figures A.9*a*–*c* show the results that include secondary coma. Figure A.11 shows the OTF, which for coma is a complex quantity, on an Argand diagram.

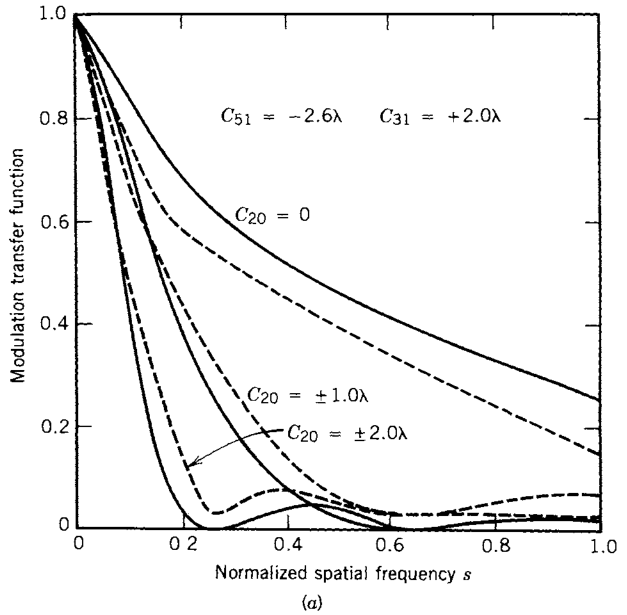
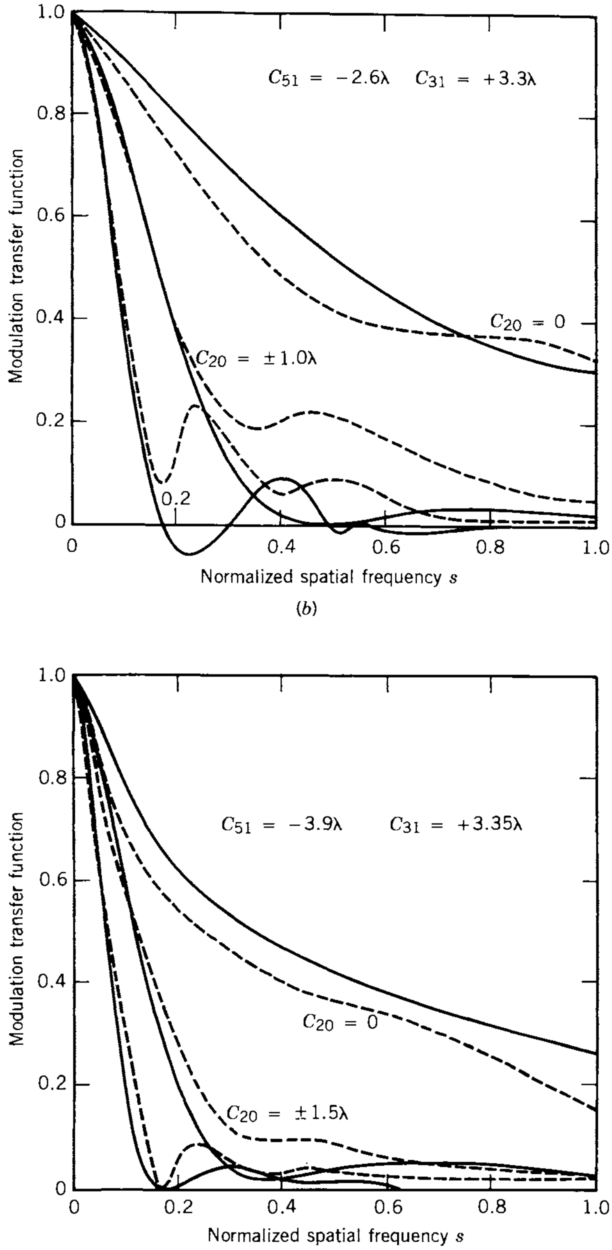


Figure A.9. MTF curves in the presence of primary, C_{31}, and secondary, C_{51}, coma in various focal planes. ——, Azimuth ψ equal to zero; - - -, $\psi = \pi/2$ [6]. (Reproduced by permission of The General Electric Co., Ltd., of England.)



(*b*)

(*c*)

Figure A.9. (*Continued*)

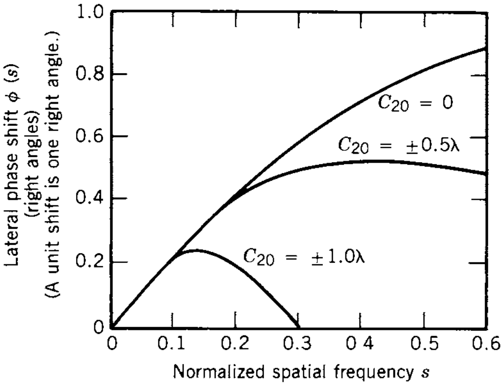


Figure A.10. Lateral phase shift for primary coma C_{31} = ±0.63λ in various focal planes plotted as function of the spatial frequency s [6]. (Reproduced by permission of The General Electric Co., Ltd., of England.)

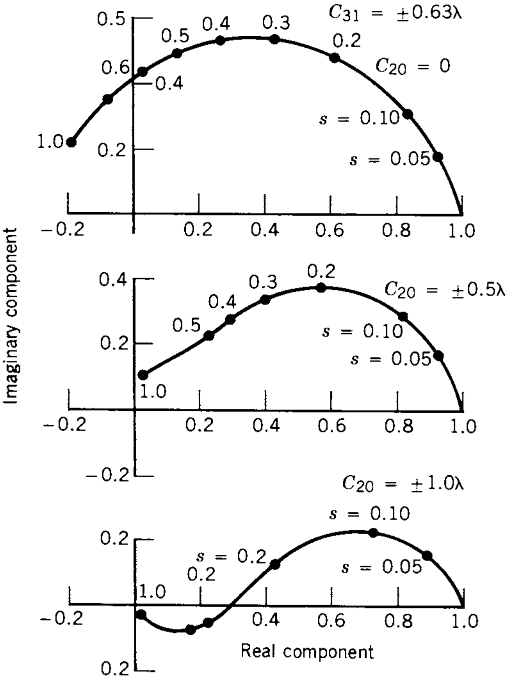


Figure A.11. MTF in the presence of primary coma plotted on an Argand diagram. Numbers around the curves indicate values of the normalized spatial frequency s at the points [6]. (Reproduced by permission of The General Electric Co., Ltd., of England.)

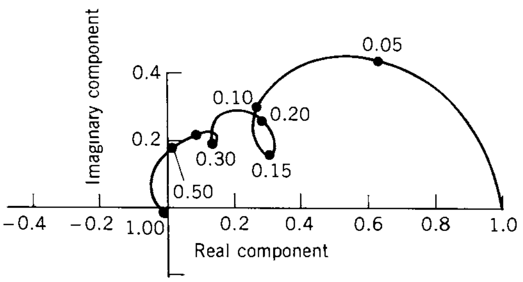


Figure A.12. MTF in the presence of primary and secondary coma plotted on an Argand diagram; $C_{31} = -4.6\lambda$, $C_{51} = 2.6\lambda$, $C_{20} = \pm 2\lambda$ [6]. (Reproduced by permission of The General Electric Co., Ltd., of England.)

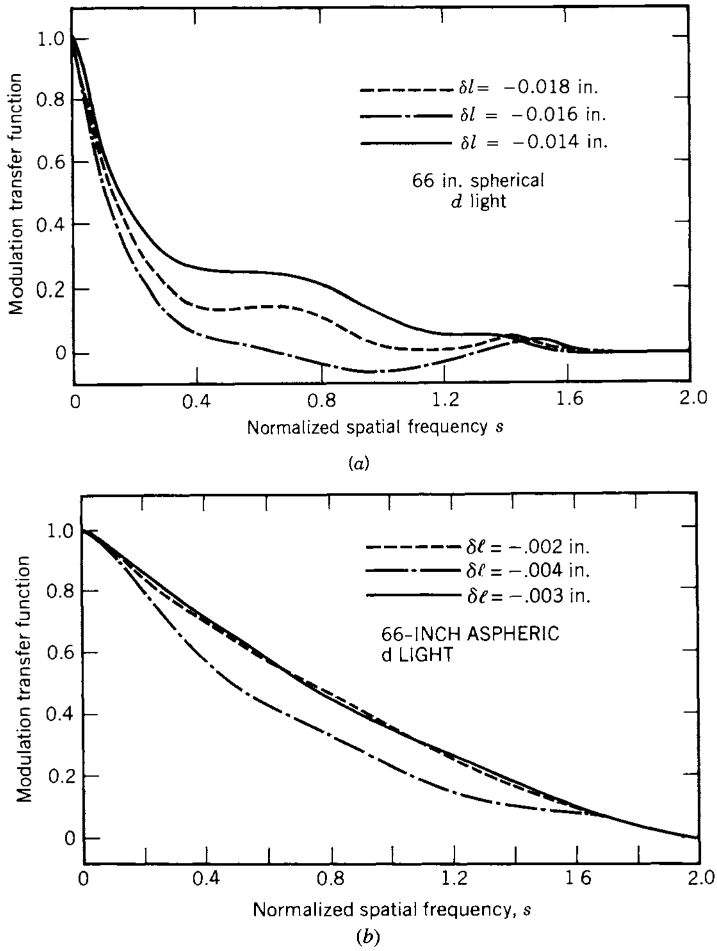
Figure A.10 shows the lateral phase shift, in right angles, for primary coma, $C_{31} = \pm 0.63\lambda$, in three different image positions. Figure A.12 is an Argand diagram for the OTF when $C_{20} = \pm 2\lambda$, $C_{31} = -4.6\lambda$, and $C_{51} = 2.6\lambda$. The figure illustrates how the phase transfer function can circulate, with relatively large amounts of aberration, as the spatial frequency varies from zero to 1.00; the PTF here is the angle ϕ of Eq. (2-19).

SPHERICAL ABERRATION WITH COLOR

In a series of papers during the 1960s Richard Barakat [7–12], working with several different collaborators, reported on extensive calculations of the OTF. He used optical path difference data obtained from the lens design data in a method that was discussed in Chapters 9 and 10. He then represented the aberration function in a series of Tschebyscheff polynomials and used the Gauss quadrature method of numerical integration to accomplish the convolution of the pupil function; these processes were discussed very briefly in Chapter 10.

Figures A.13*a*, A.14*a*, and A.15*a* show Barakat's results of computations (with M. V. Morrello) for a spherical doublet; and Figs. A.13*b*, A14*b*, and A.15*b* show the results for the same lens but with a tenth-order aspheric fitted (in *d*-light) on the last surface. Calculations were made at different wavelengths: *C*-light at 6562.8 Å, *d*-light at 5875.5 Å, and *e*-light at 5460.7 Å. The parameter δl is the displaced focal setting. When δl is zero, the setting is at the paraxial focus; a negative value for δl means a defocusing toward the marginal focus. All calculations were made for an $f/5$ system at focal lengths of 66, 33, and 13.2 in.

Considering spherical aberration as used in these calculations, we note the behavior of the MTF at low and high spatial frequencies to be almost independent of the order (i.e., 1st, 2nd, 3rd, 5th, etc.), when the coefficients for all orders have the same value. "The low order coefficients influence the MTF in



(*a*)

(*b*)

Figure A.13. MTF curves calculated from design data for a 66-in. spherical doublet in *d* light for different focal settings. The symbol $\delta l = 0$ is paraxial focus; δl is distance toward marginal focus [8].

the midrange the most since they deviate more from zero over the aperture than do the higher order" [8].

OPTIMUM BALANCED FIFTH-ORDER SPHERICAL ABERRATION

Barakat [9] calculated the OTF for examples having different amounts of, but a constant ratio between, third-order and fifth-order spherical aberration. He

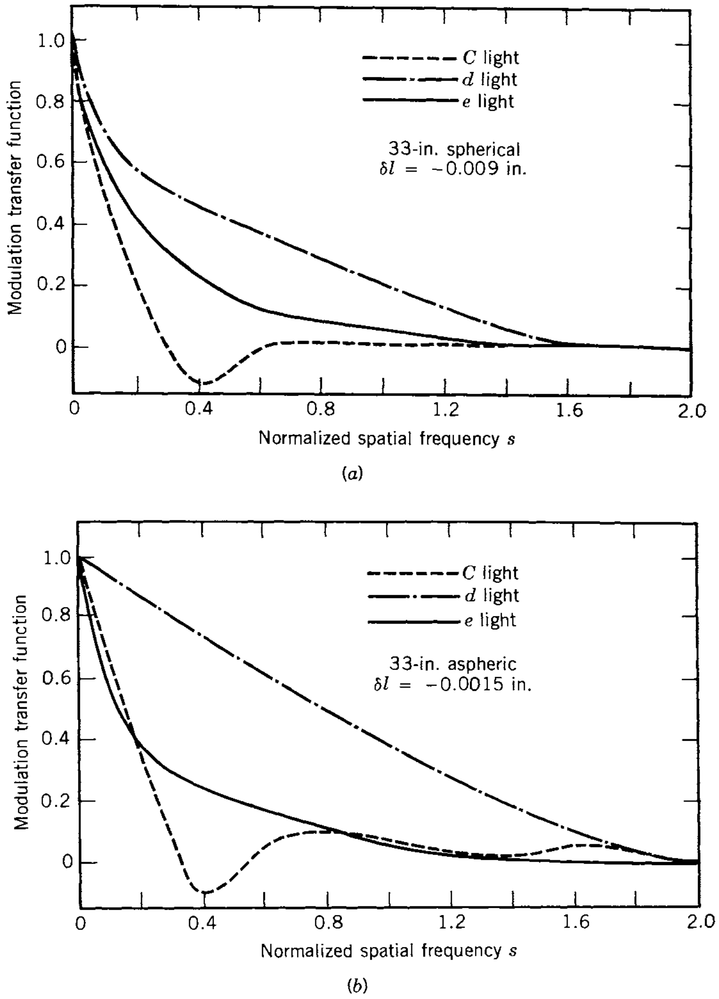


Figure A.14. MTF curves calculated from design data for a 33-in. spherical doublet with different amounts of defocusing and different wavelengths of light [8].

uses a formula for optimum balance,

$$W(\rho, \varphi) = C_{60}\left[\rho^6 - (3/2)\,\rho^4 + (3/5)\,\rho^2\right], \qquad \text{(A-6)}$$

which was published by W. Ta Hang in 1941. The formula thus gives $C_{40} = 1.5C_{60}$ and $C_{20} = 0.6C_{60}$. However, Barakat does use different image planes,

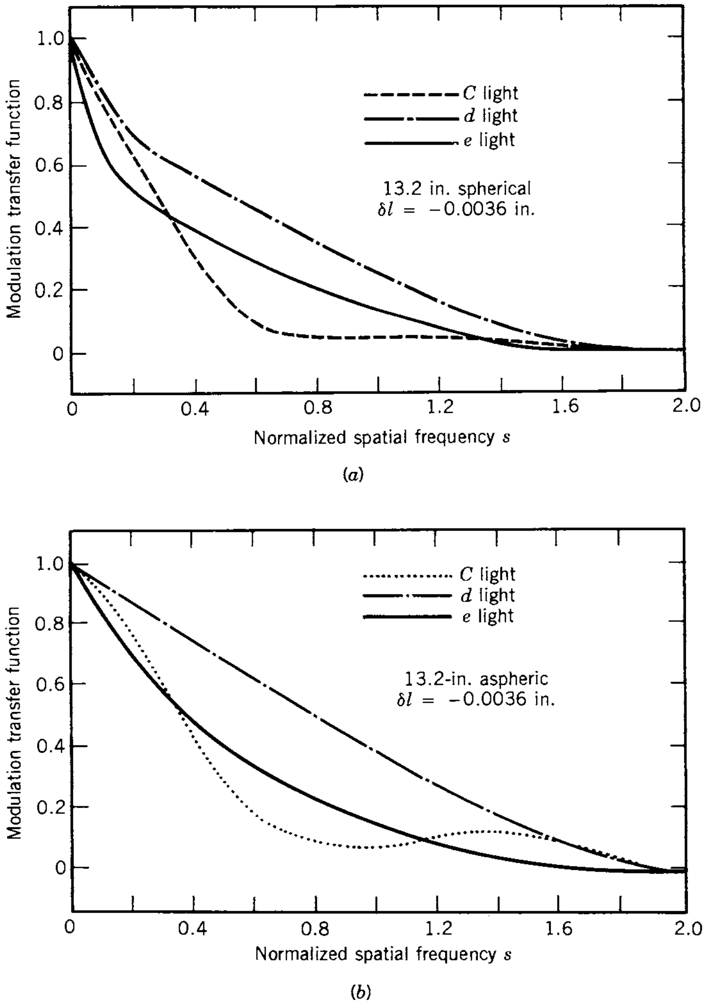


Figure A.15. MTF curves similar to the curves of Figs. A.13 and A.14 except for an aspheric doublet [8].

which would change C_{20}, the amount of defocusing, but not C_{40} and C_{60}. Typical MTF curves are shown in Figs. A.16 and A.17. The curves show that there is an image plane that gives best MTF and that any other image plane away from best focus degrades the MTF. Curve D in both sets of curves is the best; but greater amounts of aberration degrade all MTF curves although the optimum balance still produces the best MTF.

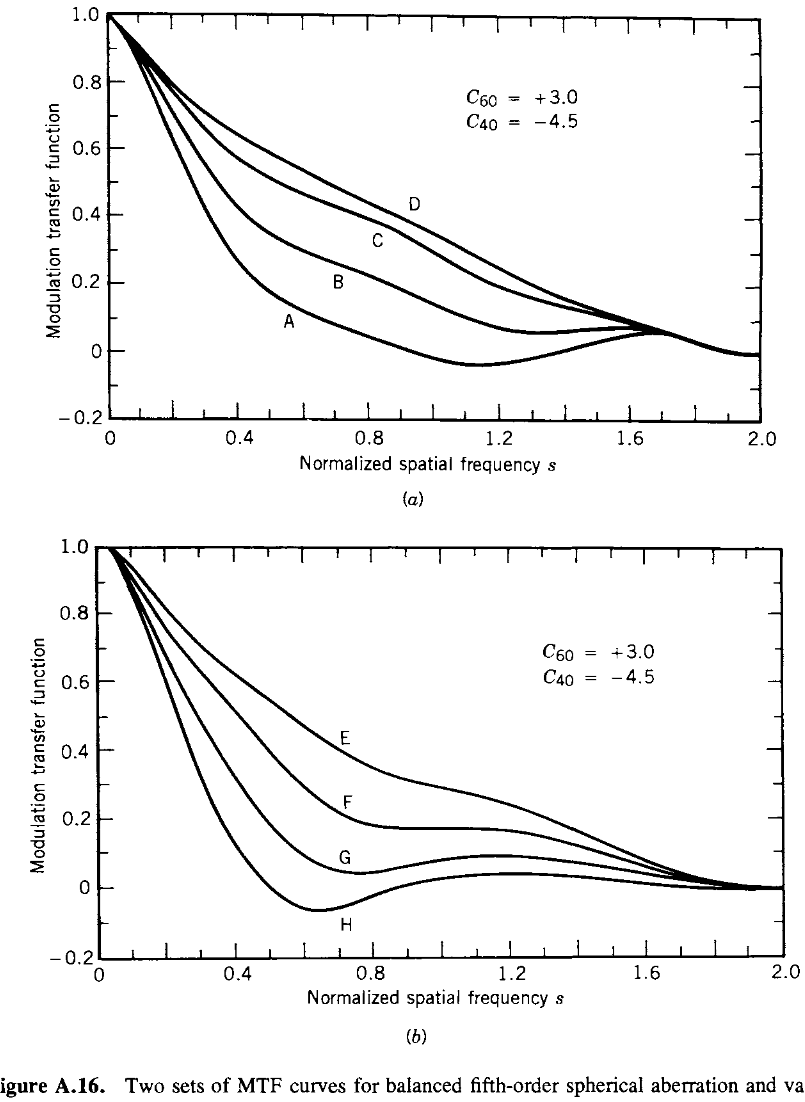
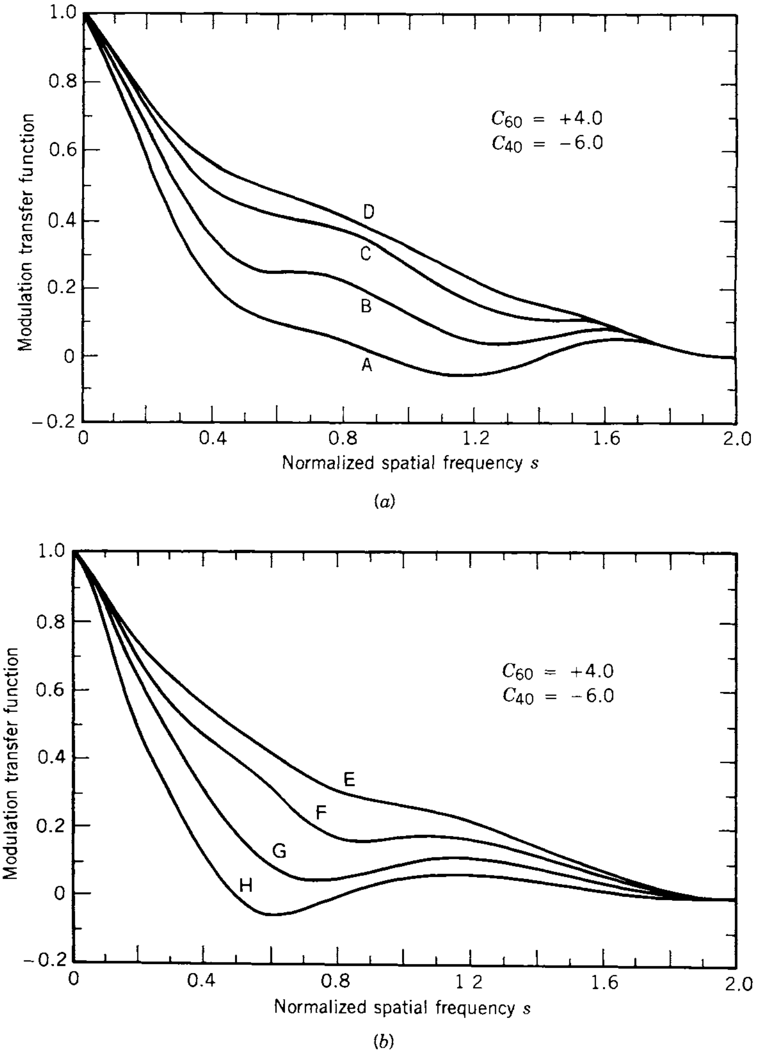


Figure A.16. Two sets of MTF curves for balanced fifth-order spherical aberration and various amounts of defocusing. Defocusing is measured in wavelengths from paraxial focus. Letter codes on the curves identify amounts of defocusing: A, 1.2; B, 1.4; C, 1.6; D, 1.8; E, 2.0; F, 2.2; G, 2.4; and H, 2.6. Curve D shows the MTF in the plane nearest best focus; the Strehl criterion test locates this same best focal plane [9].



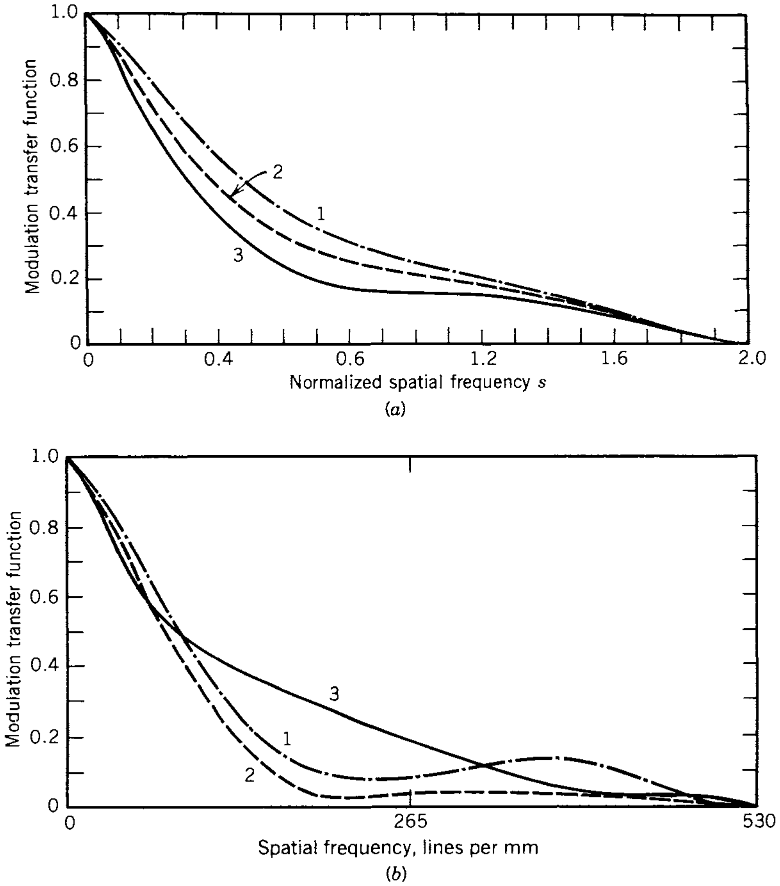
(*a*)

(*b*)

Figure A.17. Two sets of MTF curves similar to those of Fig. A.16 except for greater amounts of aberration and different amounts of defocusing. Correlation of defocusing with letters on the curves are A to H with defocusing from 1.8 to 3.2 wavelengths (in increments of 0.2), respectively. Curve D is again the setting nearest the best focus [9].

PRIMARY COMA AT DIFFERENT AZIMUTHS

It is difficult to make critical comparisons between results by different investigators; for example, Goodbody, whose work was reported in a previous section, calculated the OTF for coma at only two azimuths: $\psi = 0$ and $\psi = \pi/2$. When Barakat and Houston [10] calculated for coma, they added a calculation for $\psi = \pi/4$. Miyamoto [13] also included calculations at $\psi = \pi/4$. But there are differences between the calculations in the three papers that might be related to differences in their results.



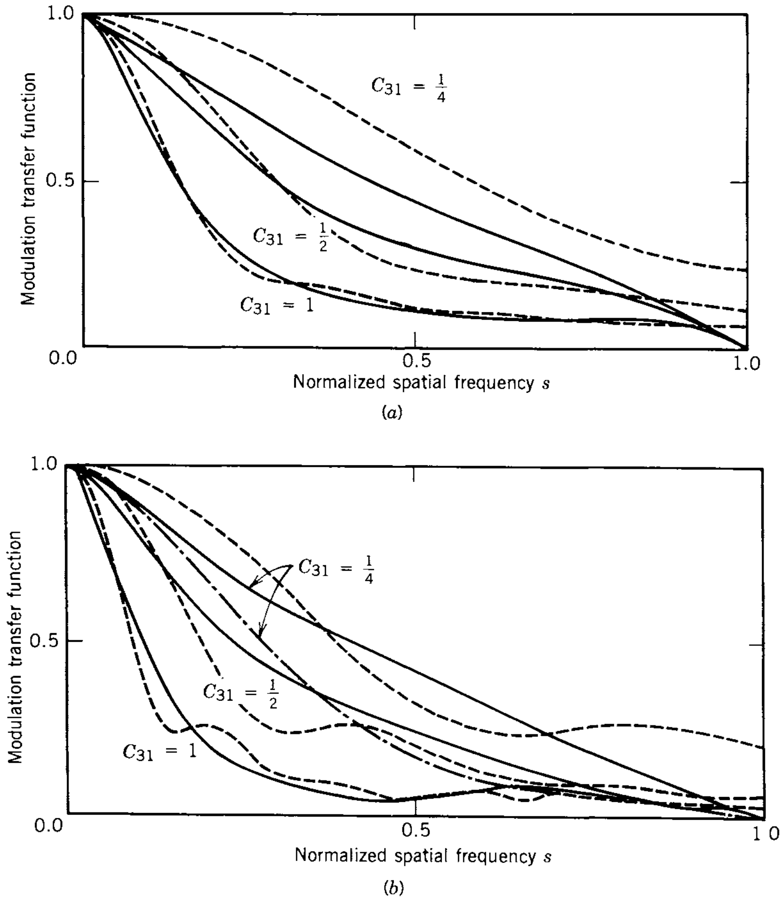
(a)

(b)

Figure A.18. (*a*) MTF curves for third-order coma, the author's $C_{131} = 1.0\lambda$, at paraxial focus, $C_{20} = 0$. Curve 1 is for $\psi = 0$; curve 2 is for $\psi = 45°$; and curve 3 is for $\psi = 90°$. (*b*) MTF curves similar to those of (*a*) except for an $f/3.5$ lens operating at 6° off axis with light at 5400 Å in the plane of best focus. Line code is the same as for (*a*) [10].

Barakat and Houston used Luneburg's modified Kirchhoff's diffraction theory, the Hamiltonian mixed characteristic which represents the optical path length of a ray from the object point x_0, y_0 to the exit pupil where the optical direction cosines of the normals to the converging wave front become the exit pupil coordinates. The wave aberration function is expressed in terms of the direction cosines. The Gauss quadrature method is used to accomplish the numerical evaluation of the OTFs.

Figure A.18*a* shows the MTF for three values of the azimuth angle—0, $\pi/4$, $\pi/2$—with the coefficient for third order coma—their C_{131}—equal to one wavelength of wave distortion. It would seem from Fig. A.18*a* that the OTF at $\psi = \pi/4 = 45°$ could be represented by a simple average of the OTFs at $\psi = 0$



(*a*)

(*b*)

Figure A.19. MTF curves for coma alone for azimuth angles (*a*) $\psi = 0$ and (*b*) $\psi = 90°$ [13].

and $\psi = 90°$. But the MTF curves of Fig. A.18*b* for a test lens (*f*/3.5; 6° off axis; light at 5400 Å; in the plane of best focus, $C_{020} = 0.7\lambda$) show that an appreciable error could be made by averaging the OTFs at 0° and 90° to obtain the OTF at 45°. Aberration coefficients other than third-order coma were included when the curves of Fig. A.18*b* were calculated.

The plane of best focus was the image plane in which the variance of the wave front is minimum; this is a criterion suggested by Maréchal [14] as a merit function in lens design and evaluation. The results of Barakat and Houston do indicate that the best MTF correlates with minimum variance.

Miyamoto [13] also calculated the MTF for coma at an azimuth of 45°; but his optical system has a square aperature. He shows us a form of nonrotationally symmetric optical systems in which the shape of the aperture has a bearing on the OTF. When the "square" pupil functions are sheared in a direction parallel

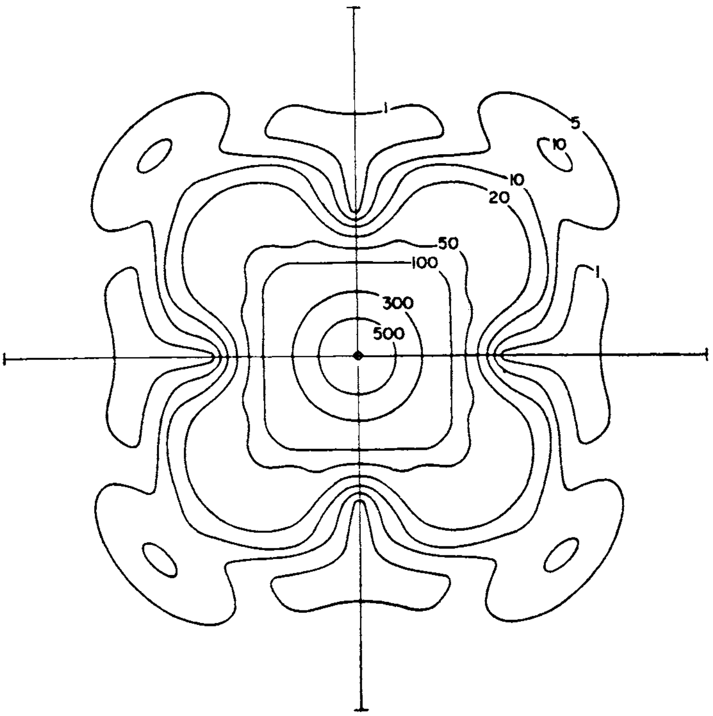


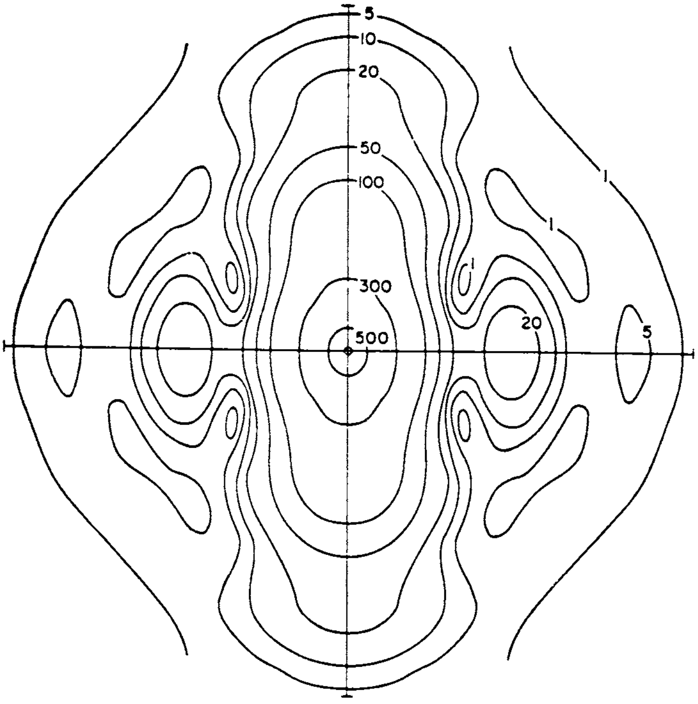
Figure A.20. Curves of constant illuminance for a circular aperture in the Fraunhofer receiving plane for cases of nonrotationally symmetric aberrations. The parameters *C*, *E*, *F*, and *G* are explained in the text. (*a*) $C = 0.5\lambda$. (*b*) $E = 0.5\lambda$. For $F = 0.5\lambda$, the figure would be rotated 90°. (*c*) $G = 0.5\lambda$ [11].

to a side of the aperture, the "perfect" MTF is a linearly decreasing function of s; thus, even the perfect MTF is higher in the midrange than the perfect MTF for a circular aperture, a side of the square being equal to the diameter of the circle. In any other direction of shear, the MTF varies nonlinearly with s, and the amount of nonlinearity varies with ψ. When shearing along a diameter, the scale of s differs by a factor of $\sqrt{2}$ from the scale when shearing parallel to a side of the square.

Figures A.19*a* and *b* show MTF curves calculated by Miyamoto; these curves also show that one cannot count on taking the average of the zero azimuth and the 90° azimuth for the MTF at 45°.

NONROTATIONALLY SYMMETRIC SYSTEMS

The work of Barakat and Houston [11] published in 1966 is an important departure from the comfort of thinking altogether in terms of the classical aber-



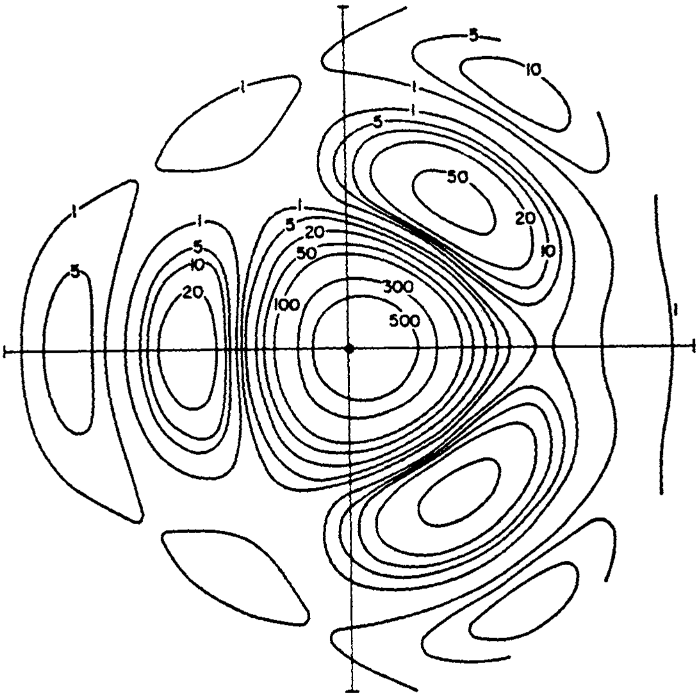
(b)

Figure A.20. (*Continued*)

rations and of rotationally symmetric optical systems. We even gamble somewhat our own professional stature by trying to simplify their explanation or to summarize the work. Nevertheless, because of the elegance of their analysis and generality of the treatment, we want their work to be mentioned here.

The most common optical system is one that can, at least in theory, be rotated about its axis with a change of neither its imaging properties nor the image position. Let us recall that even the square aperture used by Miyamoto, which was discussed briefly in a preceding section, ties certain imaging properties to the orientation of the optical system; and they must turn with the system, if it is rotated about its axis. However, Barakat and Houston deal with more subtle nonrotationally symmetric properties than those that might be produced by a noncircular aperture.

As in Ref. 10, these two authors use the Kirchhoff diffraction theory as expressed by Luneburg and the Hamilton mixed characteristic function. The mixed characteristic is expanded in a series of power polynomials; then terms and powers are chosen according to types of aberrations and number of planes of

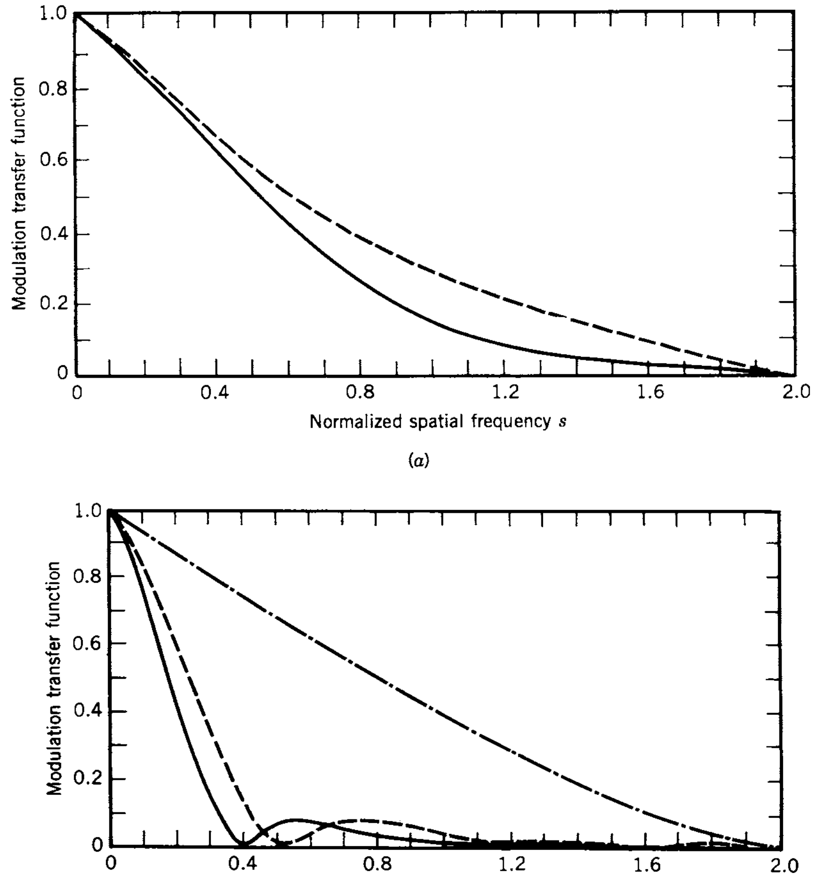


(c)

Figure A.20. (*Continued*)

symmetry. They mention as an example of a system with only one plane of symmetry a perfect refracting system with one lens element tilted with respect to either the tangential or the meridional plane. A system with two planes of symmetry would be an anamorphotic imaging system. No plane of symmetry is exemplified by a completely decentered objective.

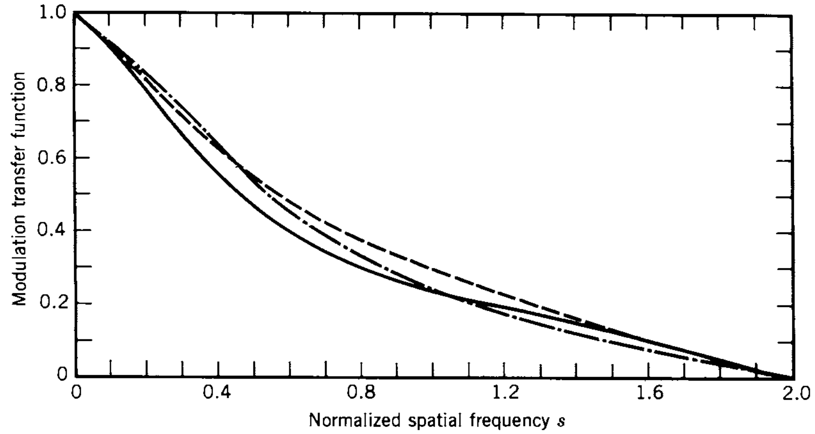
We show in Figs. A.20*a*–*c* examples of the calculated spread function in the



(*a*)

(*b*)

Figure A.21. Three sets of MTF curves. Calculations were made for nonrotationally symmetric aberrations. (*a*) The parameter $C = 1\lambda$. ——, $\psi = 0$ or 90°; - - -, $\psi = 45°$. (*b*) The parameter $E = 1\lambda$. - · - · -, $\psi = 0$; - - -, $\psi = 45$; ——, $\psi = 90°$. The same curves would hold for F if $\psi = 0$ and $\psi = 90°$ were interchanged. (*c*) The parameter $G = 1\lambda$. Line code is the same as in (*a*) and (*b*) [11].



(*c*)

Figure A.21. (*Continued*)

Fraunhofer receiving plane with maximum distortion in each example of 0.5 wavelength. The resulting MTF curves for the same three of their parameters—C, E, and G, which are tensors—are shown in Figures A.21*a*–*c*:

C is for defocusing due to a tilt of the wave front;

E and F are defocusing in directions of the direction cosines p and q, respectively; and

G is for a nonsymmetrical aberration which they characterize as comatic aberrations.

These parameters are for no plane of symmetry, and the maximum wave-front distortion for the MTF curves is in each example one wavelength.

REFERENCES

1. H. H. Hopkins, The Frequency Response of a Defocused Optical System. *Proc. R. Soc. London Ser. A* **231,** 91(1955).
2. G. Black and E. H. Linfoot, Spherical Aberration and the Information Content of Optical Images. *Proc. R. Soc. London Ser. A* **239,** 522(1957).
3. A. M. Goodbody, The Influence of Spherical Aberration on the Response Function of an Optical System. *Proc. Phys. Soc. (London) Ser. B* **72,** 411(1958).

4. H. H. Hopkins, The Numerical Evaluation of the Frequency Response of Optical Systems. *Proc. Phys. Soc. (London) Ser. B* **70,** 1002 (1957).
5. H. H. Hopkins, The Aberration Permissible in Optical Systems. *Proc. Phys. Soc. (London) Ser. B* **70,** 449 (1957).
6. A. M. Goodbody, The Influence of Coma on the Response Function of an Optical System. *Proc. Phys. Soc. (London) Ser. B* **75,** 677 (1960).
7. R. Barakat, Computation of the Transfer Function of an Optical System from the Design Data for Rotationally Symmetric Aberrations I. Theory. *J. Opt. Soc. Am.* **52,** 985 (1962).
8. R. Barakat and M. V. Morello, Computation of the Transfer Function of an Optical System from the Design Data for Rotationally Symmetric Aberrations, II. Programming and Numerical Results. *J. Opt. Soc. Am.* **52,** 992 (1962).
9. R. Barakat, Numerical Results Concerning the Transfer Functions and Total Illuminance for Optimum Balanced Fifth-Order Spherical Aberration. *J. Opt. Soc. Am.* **54,** 38 (1964).
10. R. Barakat and A. Houston, Transfer Function of an Optical System in the Presence of Off-Axis Aberrations. *J. Opt. Soc. Am.* **55,** 1142 (1965).
11. R. Barakat and A. Houston, The Aberrations of Non-Rotationally Symmetric Systems and Their Diffraction Effects. *Opt. Acta* **13,** 1 (1966).
12. R. Barakat, The Influence of Random Wavefront Errors on the Imaging Characteristics of an Optical System. *Opt. Acta* **18,** 683 (1971).
13. K. Miyamoto, Wave Optics and Geometrical Optics in Optical Design. In *Progress in Optics*, Vol. 1, E. Wolf (Ed.). North Holland, Amsterdam, 1961.
14. A. Maréchal, Study of the Combined Effect of Diffraction and Geometrical Aberrations on the Image of a Luminous Point. *Rev. d'Opt.* **26,** 257 (1947).

Appendix B

Some Mathematics

THE FOURIER TRANSFORM

The *Fourier transform* of a function $f(x)$ of a single variable x is defined in this book as

$$F(\omega) = \int_{-\infty}^{+\infty} f(x) \exp(-i2\pi\omega x)\, dx. \tag{B-1}$$

In our optics discussions, the parameter x is usually a space coordinate or its equivalent, and ω is spatial frequency.

The *inverse Fourier transform*, also called the *Fourier integral*, by which $f(x)$ can be recovered from $F(\omega)$ is defined by

$$f(x) = \int_{-\infty}^{+\infty} F(\omega) \exp(i2\pi\omega x)\, d\omega. \tag{B-2}$$

The conditions, known as the *Dirichlet conditions*, under which a Fourier integral representation for a given function is possible are adequately summarized for our purposes by the expression

$$\int_{-\infty}^{+\infty} |f(x)|\, dx \quad \text{shall be finite.} \tag{B-3}$$

In our applications, $f(x)$ usually represents the flow of radiant energy in units appropriate to flux density, incidance, exitance, intensity, radiance, or reflectance. The integral of Eq. (B-3) would, therefore, represent the totalizing of some kind of energy in an optical system, and we know that such a total must, in an actual system, remain finite.

References in this book to Eqs. (B-1) and (B-2) are usually shortened to *transform* and *inverse transform*, respectively.

In general, the functions $f(x)$ and $F(\omega)$ are complex. Usually we will depend upon the context to indicate whether quantities are real or complex; however, when we wish to indicate explicitly that functions are complex, carets will be added to the symbols as, for example, $\hat{f}(x)$ and $\hat{F}(\omega)$. In our discussion of the transforms, we use the subscripts 1 and 2 to label the real and imaginary parts of the functions:

$$\hat{f}(x) = f_1(x) + if_2(x), \tag{B-4}$$

$$\hat{F}(\omega) = F_1(\omega) + iF_2(\omega). \tag{B-5}$$

If the expanded form of $f(x)$ given in Eq. (B-4) is substituted in the transform expression of Eq. (B-1) and the trigonometric equivalent is substituted for the exponential factor in the transform according to the identity, $\exp(-2\pi i\omega x) = \cos 2\pi\omega x - i \sin 2\pi\omega x$, the following form of the transform results:

$$\begin{aligned}\hat{F}(\omega) &= \int_{-\infty}^{+\infty} [f_1(x)\cos 2\pi\omega x + f_2(x)\sin 2\pi\omega x]\, dx \\ &\quad - i\int_{-\infty}^{+\infty} [f_1(x)\sin 2\pi\omega x - f_2(x)\cos 2\pi\omega x]\, dx.\end{aligned} \tag{B-6}$$

Then, according to Eq. (B-5), the real and imaginary parts of $\hat{F}(\omega)$ are

$$F_1(\omega) = \int_{-\infty}^{+\infty} [f_1(\omega)\cos 2\pi\omega x + f_2(x)\sin 2\pi\omega x]\, dx, \tag{B-7}$$

$$F_2(\omega) = \int_{-\infty}^{+\infty} [-f_1(x)\sin 2\pi\omega x + f_2(x)\cos 2\pi\omega x]\, dx. \tag{B-8}$$

If the steps leading to Eqs. (B-6)–(B-8) are repeated for the inverse transform, the real and imaginary parts of $\hat{f}(x)$ are found to be

$$f_1(x) = \int_{-\infty}^{+\infty} [F_1(\omega)\cos 2\pi\omega x - F_2(\omega)\sin 2\pi\omega x]\, d\omega, \tag{B-9}$$

$$f_2(x) = \int_{-\infty}^{+\infty} [F_1(\omega)\sin 2\pi\omega x + F_2(\omega)\cos 2\pi\omega x]\, d\omega. \tag{B-10}$$

If $\hat{\mathrm{f}}(x)$ is a real function, that is, if it is equal to $f_1(x)$ because $f_2(x)$, the imaginary part, is equal to zero, the real and imaginary parts of $\hat{F}(\omega)$ are

$$F_1(\omega) = \int_{-\infty}^{+\infty} f(x) \cos 2\pi\omega x \, dx, \tag{B-11}$$

$$F_2(\omega) = \int_{-\infty}^{+\infty} -f(x) \sin 2\pi\omega x \, dx. \tag{B-12}$$

Similarly, when $\hat{F}(\omega)$ is real,

$$f_1(x) = \int_{-\infty}^{+\infty} F(\omega) \cos 2\pi\omega x \, d\omega, \tag{B-13}$$

$$f_2(x) = \int_{-\infty}^{+\infty} F(\omega) \sin 2\pi\omega x \, d\omega. \tag{B-14}$$

Because we can often choose where to place the origin in the coordinate system for $f(x)$, we find that we can then simplify the mathematics of the Fourier transform operations by making the appropriate choice. For instance, if $f(x)$ is real and has a symmetry that makes it an *even* function, that is, $f(x) = f(-x)$, with proper choice of origin, certain simplifying benefits result. The fact that $f(x)$ is real, as has already been shown, reduces the real and imaginary parts of the transform to Eqs. (B-11) and (B-12). Because the sine function is *odd*, that is, $\sin 2\pi\omega x = -\sin 2\pi\omega(-x)$, and $f(x)$ has been assumed even, the integrand in Eq. (B-12) is an odd function. The indicated integral of an odd function is always zero, so $F_2(\omega)$ is zero. This means that the transform of a real even function is real. Furthermore, because $f(x)$ is real, making $f_2(x)$ in Eq. (B-10) equal to zero, the remaining integrand, $F_1(\omega) \sin 2\pi\omega x$, must be an odd function. Since the sine factor is odd, $F_1(\omega)$ then has to be even to make the product odd. To summarize, if $f(x)$ is both real and even, its transform $F(\omega)$ must also be real and even.

When the transform and its inverse are each a function of two dimensions, Eqs. (B-1) and (B-2) are replaced by

$$F(\omega_x, \omega_y) = \iint_{-\infty}^{+\infty} f(x, y) \exp[-i2\pi(\omega_x x + \omega_y y)] \, dx \, dy, \tag{B-15}$$

$$f(x, y) = \iint_{-\infty}^{+\infty} F(\omega_x, \omega_y) \exp[i2\pi(\omega_x x + \omega_y y)] \, d\omega_x \, d\omega_y. \tag{B-16}$$

When $\hat{F}(\omega)$ is the transform of $\hat{f}(x)$, a common shorthand to show their relation is

$$\hat{f}(x) \leftrightarrow \hat{F}(\omega). \tag{B-17}$$

A frequently encountered relation in transform theory, which will not be derived here, is *Parseval's formula* expressed as

$$\int_{-\infty}^{+\infty} |f(x)|^2 \, dx = \int_{-\infty}^{+\infty} A^2(\omega) \, d\omega, \tag{B-18}$$

where $\hat{f}(x) \leftrightarrow \hat{F}(\omega)$ and $\hat{F}(\omega) = A(\omega) \exp[i\phi(\omega)]$. According to the usual convention, $A^2(\omega)$ is the product of $\hat{F}(\omega)$ times its complex conjugate. A more general form of Parseval's formula is

$$\int_{-\infty}^{+\infty} [\hat{f}_1(x)][\hat{f}_2(x)] \, dx = \int_{-\infty}^{+\infty} [\hat{F}_1(-\omega)][\hat{F}_2(\omega)] \, d\omega, \tag{B-19}$$

where $\hat{f}_1(x) \leftrightarrow \hat{F}_1(\omega)$ and $\hat{\mathrm{f}}_2(x) \leftrightarrow \hat{F}_2(\omega)$.

THE DELTA FUNCTION

The *Dirac delta function*, $\delta(x)$ or $\delta(x, y)$, is useful in optics to represent point sources of light and arrays of apertures and slits. Here we only outline the properties of the delta function and refer the reader to standard texts [1–3] for a more complete treatment.

The delta function (more correctly called a distribution) is zero except when the argument is zero:

$$\delta(x) = 0 \quad \text{when} \quad x \neq 0, \tag{B-20}$$

but in the immediate vicinity of $x = 0$, the delta function has the peculiar property to make the integral

$$\int_{-\infty}^{+\infty} \delta(x) \, dx = 1. \tag{B-21}$$

Similarly,

$$\delta(x - a) = 0 \quad \text{when} \quad x \neq a, \tag{B-22}$$

but in the immediate vicinity of where the argument is zero, that is, where $x = a$, the delta function takes on a value to make the integral

$$\int_{-\infty}^{+\infty} \delta(x - a)\, dx = 1. \tag{B-23}$$

Therefore, when this delta function is multiplied by a function $f(x)$, the integral becomes

$$\int_{-\infty}^{+\infty} f(x)\, \delta(x - a)\, dx = f(a), \tag{B-24}$$

where the parameter a may be a constant or it may be an independent variable that ranges, for instance, over the same values allowed for x. When the delta function is applied in this way, it is part of a coordinate shifting mechanism.

In Eq. (B-24) if $a = 0$, the integral, of course, becomes $f(0)$. Thus, the transform of a constant times the delta function is simply the constant:

$$\int_{-\infty}^{+\infty} A\delta(x) \exp(-i2\pi\omega x)\, dx = A, \tag{B-25}$$

which is an application of Eq. (B-1). The inverse Fourier transform according to Eq. (B-2) is then

$$\int_{-\infty}^{+\infty} A \exp(i2\pi\omega x)\, d\omega = A\delta(x). \tag{B-26}$$

It is at once evident that the reverse sequence would work out the same, that is, the inverse transform of $A\delta(x)$ would be A; and the transform A would be $A\delta(x)$. If in $\hat{f}(x) \leftrightarrow \hat{F}(\omega)$ we choose $A\delta(x)$ to be $\hat{f}(x)$, then $\hat{F}(\omega)$ becomes A, which tells us that the spectrum of the delta function includes all optical frequencies of equal amplitudes and at zero phase.

From our discussion of the delta function, it is evident that

$$\delta(x) = \delta(-x), \tag{B-27}$$

and

$$\delta(x - a) = \delta(a - x). \tag{B-28}$$

So the delta function is recognized as even. This property can be extended in two dimensions to

$$\delta(x, y) = \delta(-x, y) = \delta(x, -y) = \delta(-x, -y), \qquad \text{(B-29)}$$

which indicates that the delta function has a radial symmetry about the zero of its argument.

When the transform of the shifted delta function is taken and Eq. (B-24) is applied

$$\int_{-\infty}^{+\infty} \delta(x - a) \exp(-i2\pi\omega x)\, dx = \exp(-i2\pi\omega a). \qquad \text{(B-30)}$$

As in our interpretation of the transform of the unshifted delta function, all frequencies are present at a constant amplitude; but because the exponent is proportional to the amount of the shift a, we can add that the phase is proportional to the shift as well as the frequency ω. In other words, the phase varies linearly with the product ωa.

THE CONVOLUTION INTEGRAL

In the following equation, the integral $\hat{f}_3(x)$ is said to be the convolution of the two functions, $\hat{f}_1(x)$ with $\hat{f}_2(x)$:

$$\hat{f}_3(x) = \int_{-\infty}^{+\infty} \hat{f}_1(\alpha)\hat{f}_2(x - \alpha)\, d\alpha. \qquad \text{(B-31)}$$

This relation between three functions occurs when their transforms are related as

$$\hat{F}_3(\omega) = [\hat{F}_1(\omega)][\hat{F}_2(\omega)], \qquad \text{(B-32)}$$

where

$$\hat{f}_1(x) \leftrightarrow \hat{F}_1(\omega), \qquad \hat{f}_2(x) \leftrightarrow \hat{F}_2(\omega), \qquad \hat{f}_3(x) \leftrightarrow \hat{F}_3(\omega). \qquad \text{(B-33)}$$

The consistency between Eqs. (B-31) and (B-32) can be demonstrated by successive substitution in Eq. (B-2):

$$\begin{aligned}\hat{f}_3(x) &= \int_{-\infty}^{+\infty} \hat{F}_3(\omega) \exp(i2\pi\omega x)\, d\omega \\ &= \int_{-\infty}^{+\infty} [\hat{F}_1(\omega)\, \hat{F}_2(\omega)] \exp(i2\pi\omega x)\, d\omega. \end{aligned} \qquad \text{(B-34)}$$

Writing out the first two shorthand expressions of Eq. (B-33), we have

$$\hat{F}_1(\omega) = \int_{-\infty}^{+\infty} \hat{f}_1(x) \exp(-i2\pi\omega x)\, dx, \tag{B-35}$$

$$\hat{F}_2(\omega) = \int_{-\infty}^{+\infty} \hat{f}_2(x) \exp(-i2\pi\omega x)\, dx. \tag{B-36}$$

For convenience in later substitutions, the variable of integration in Eqs. (B-35) and (B-36) is changed as follows (without, of course, affecting the sense of the two equations):

$$\hat{F}_1(\omega) = \int_{-\infty}^{+\infty} \hat{f}_1(\alpha) \exp(-i2\pi\omega\alpha)\, d\alpha, \tag{B-37}$$

$$\hat{F}_2(\omega) = \int_{-\infty}^{+\infty} \hat{f}_2(\gamma) \exp(-i2\pi\omega\gamma)\, d\gamma. \tag{B-38}$$

These two expressions for $\hat{F}_1(\omega)$ and $\hat{F}_2(\omega)$ are substituted in Eq. (B-34):

$$\hat{f}_3(x) = \int_{-\infty}^{+\infty} \left\{ \left[\int_{-\infty}^{+\infty} \hat{f}_1(\alpha) \exp(-i2\pi\omega\alpha)\, d\alpha \right] \right.$$
$$\left. \times \left[\int_{-\infty}^{+\infty} \hat{f}_2(\gamma) \exp(-i2\pi\omega\gamma)\, d\gamma \right] \right\} \exp(i2\pi\omega x)\, d\omega. \tag{B-39}$$

By combining the exponential factors and rearranging other terms, Eq. (B-39) becomes

$$\hat{f}_3(x) = \iint_{-\infty}^{+\infty} \hat{f}_1(\alpha) \hat{f}_2(\gamma) \left\{ \int_{-\infty}^{+\infty} \exp[i2\pi\omega(x - \alpha - \gamma)]\, d\omega \right\} d\gamma\, d\alpha. \tag{B-40}$$

By a substitution of variable, $x_1 = (x - \alpha - \gamma)$, the integral within braces is recognized as the one in Eq. (B-26) with $A = 1$, so

$$\int_{-\infty}^{+\infty} 1 \exp(i2\pi\omega x_1)\, d\omega = 1\ \delta(x_1). \tag{B-41}$$

When this delta function is substituted in Eq. (B-40),

$$\hat{f}_3(x) = \int_{-\infty}^{+\infty} \hat{f}_1(\alpha) \left\{ \int_{-\infty}^{+\infty} \hat{f}_2(\gamma)\, \delta(x - \alpha - \gamma)\, d\gamma \right\} d\alpha. \quad \text{(B-42)}$$

With change of variable, $x_2 = (x - \alpha)$, the integral within braces is the same as the one in the identity of Eq. (B-24), so

$$\int_{-\infty}^{+\infty} \hat{f}_2(\gamma)\, \delta(x_2 - \gamma)\, d\gamma = \hat{f}_2(x_2). \quad \text{(B-43)}$$

When this substitution is made in Eq. (B-42),

$$\hat{f}_3(x) = \int_{-\infty}^{+\infty} \hat{f}_1(\alpha)\hat{f}_2(x - \alpha)\, d\alpha, \quad \text{(B-44)}$$

we have shown that Eq. (B-31) holds if Eq. (B-32) is assumed. A commonly used shorthand for Eq. (B-44) is

$$\hat{f}_3(x) = [\hat{f}_1(x)] * [\hat{f}_2(x)]. \quad \text{(B-45)}$$

The proposition demonstrated above, that the inverse Fourier transform (or Fourier integral) of the product of two functions is the convolution of the inverse transforms (or Fourier integrals) of the two functions, is known as the *convolution theorem.*

When successive convolutions are made, that is, when the function resulting from one convolution is convolved with another function and this sequence is continued indefinitely, the shorthand notation is

$$f(x) = [f_1(x)] * [f_2(x)] * \cdots * [f_n(x)], \quad \text{(B-46)}$$

where $f(x)$ is the result of the $(n - 1)$th convolution.

CONVOLUTION IDENTITIES

It is sometimes convenient to set up the convolution integral in a form differing from Eq. (B-31); so the following identities are offered, without proof, to provide alternatives:

$$\hat{f}(x) = \int_{-\infty}^{+\infty} \hat{f}_1(\alpha)\hat{f}_2(x - \alpha)\, d\alpha \tag{B-47}$$

$$= \int_{-\infty}^{+\infty} \hat{f}_1(x - \alpha)\hat{f}_2(\alpha)\, d\alpha \tag{B-48}$$

$$= \int_{-\infty}^{+\infty} \hat{f}_1(x - \alpha/2)\hat{f}_2(x + \alpha/2)\, d\alpha \tag{B-49}$$

$$= \int_{-\infty}^{+\infty} \hat{f}_1(x + \alpha/2)\hat{f}_2(x - \alpha/2)\, d\alpha. \tag{B-50}$$

CONVOLUTION INTEGRAL WHEN ONE FUNCTION IS SINUSOIDAL

When either $\hat{f}_1(x)$ or $\hat{f}_2(x)$ is a sinusoidal function, their convolution integral, $\hat{f}_3(x)$, is also sinusoidal provided the various integrals in the following demonstration development exist. The functions $\hat{f}_1(x)$ and $\hat{f}_2(x)$ are more explicitly represented as

$$\hat{f}_1(x) = f_1^{\mathrm{r}}(x) + if_1^{\mathrm{i}}(x), \tag{B-51}$$

$$f_2(x) = a + b\cos 2\pi\omega x, \tag{B-52}$$

where a and b are real and $b < a$. It follows that $f_2(x)$ is then also real. The functions $f_1^{\mathrm{r}}(x)$ and $f_1^{\mathrm{i}}(x)$ are assumed real with the superscripts r and i indicating the real part and the imaginary part, respectively. (The superscripts are not exponents.) The convolution of $\hat{f}_1(x)$ with $f_2(x)$ is

$$\hat{f}_3(x') = \int_{-\infty}^{+\infty} [f_1^{\mathrm{r}}(x) + if_1^{\mathrm{i}}(x)][a + b\cos 2\pi\omega(x' - x)]\, dx \tag{B-53}$$

$$\begin{aligned} &= \int_{-\infty}^{+\infty} [af_1^{\mathrm{r}}(x) + bf_1^{\mathrm{r}}(x)\cos 2\pi\omega(x' - x)]\, dx \\ &\quad + i\int_{-\infty}^{+\infty} [af_1^{\mathrm{i}}(x) + bf_1^{\mathrm{i}}(x)\cos 2\pi\omega(x' - x)]\, dx. \end{aligned} \tag{B-54}$$

Note that, unlike previous handling of the convolution integral, the variable x has been retained as the variable of integration necessitating introduction of x' as the independent variable for f_3. For convenience, parts of the two integrals in Eq. (B-54) will be represented by the symbols

$$I_1^{\mathrm{r}} = \int_{-\infty}^{+\infty} af_1^{\mathrm{r}}(x)\, dx, \tag{B-55}$$

$$I_1^{\mathrm{i}} = \int_{-\infty}^{+\infty} af_1^{\mathrm{i}}(x)\, dx. \tag{B-56}$$

The real and imaginary parts of the right-hand expression in Eq. (B-54) can then be written as follows after a trigonometric identity is substituted for the cosine function:

$$\begin{aligned} \mathrm{Re}\,[f_3(x')] = I_1^{\mathrm{r}} &+ b \cos 2\pi\omega x' \int_{-\infty}^{+\infty} f_1^{\mathrm{r}}(x) \cos 2\pi\omega x\, dx \\ &+ b \sin 2\pi\omega x' \int_{-\infty}^{+\infty} f_1^{\mathrm{r}}(x) \sin 2\pi\omega x\, dx, \end{aligned} \tag{B-57}$$

and

$$\begin{aligned} \mathrm{Im}\,[f_3(x')] = I_1^{\mathrm{i}} &+ b \cos 2\pi\omega x' \int_{-\infty}^{+\infty} f_1^{\mathrm{i}}(x) \cos 2\pi\omega x\, dx \\ &+ b \sin 2\pi\omega x' \int_{-\infty}^{+\infty} f_1^{\mathrm{i}}(x) \sin 2\pi\omega x\, dx. \end{aligned} \tag{B-58}$$

The integrals in Eqs. (B-57) and (B-58) are recognized as Fourier cosine and sine transforms, which will be represented by the symbols

$$F_{\mathrm{c}}^{\mathrm{r}}(\omega) = \int_{-\infty}^{+\infty} f_1^{\mathrm{r}}(x) \cos 2\pi\omega x\, dx, \tag{B-59}$$

$$F_{\mathrm{s}}^{\mathrm{r}}(\omega) = \int_{-\infty}^{+\infty} f_1^{\mathrm{r}}(x) \sin 2\pi\omega x\, dx, \tag{B-60}$$

$$F_{\mathrm{c}}^{\mathrm{i}}(\omega) = \int_{-\infty}^{+\infty} f_1^{\mathrm{i}}(x) \cos 2\pi\omega x\, dx, \tag{B-61}$$

$$F_{\mathrm{s}}^{\mathrm{i}}(\omega) = \int_{-\infty}^{+\infty} f_1^{\mathrm{i}}(x) \sin 2\pi\omega x\, dx. \tag{B-62}$$

When the various symbols are substituted in Eq. (B-54),

$$\hat{f}_3(x') = I_1^r + bF_c^r(\omega)\cos 2\pi\omega x' + bF_s^r(\omega)\sin 2\pi\omega x'$$
$$+ iI_1^i + ibF_c^i(\omega)\cos 2\pi\omega x' + ibF_s^i(\omega)\sin 2\pi\omega x'. \quad \text{(B-63)}$$

If a further assumption is made that $f_1(x)$ is real, that is, $f_1^i = 0$, all imaginary terms in Eq. (B-36) become zero, and the equation is simplified to

$$f_3(x') = I_1^r + b[F_c^r(\omega)\cos 2\pi\omega x' + F_s^r(\omega)\sin 2\pi\omega x']. \quad \text{(B-64)}$$

The expression in brackets can be simplified by defining ψ, a new variable, in terms of trigonometric functions as

$$\cos\psi = F_c^r(\omega)/\left\{[F_c^r(\omega)]^2 + [F_s^r(\omega)]^2\right\}^{1/2}, \quad \text{(B-65)}$$
$$\sin\psi = F_s^r(\omega)/\left\{[F_c^r(\omega)]^2 + [F_s^r(\omega)]^2\right\}^{1/2}. \quad \text{(B-66)}$$

When these two equations are solved for $F_c^r(\omega)$ and $F_s^r(\omega)$, respectively, and substitution is made in Eq. (B-64),

$$f_3(x') = I_1^r + b\left\{[F_c^r(\omega)]^2 + [F_s^r(\omega)]^2\right\}^{1/2}$$
$$\times (\cos\psi\cos 2\pi\omega x' + \sin\psi\sin 2\pi\omega x'). \quad \text{(B-67)}$$

The transform expression will be assigned the symbol $F^r(\omega)$:

$$F^r(\omega) = \left\{[F_c^r(\omega)]^2 + [F_s^r(\omega)]^2\right\}^{1/2}. \quad \text{(B-68)}$$

By substituting this symbol and substituting an identity for the trigonometric expression, Eq. (B-67) becomes

$$f_3(x') = I_1^r + bF^r(\omega)\cos(2\pi\omega x' - \psi), \quad \text{(B-69)}$$

which is similar to the assumption for $f_2(x)$, Eq. (B-52), a constant term plus a constant times a cosine function—however, here the cosine is shifted ψ radians.

SIGNIFICANCE OF THE CONVOLUTION INTEGRAL

Because the convolution integral shares importance with the Fourier transforms in the mathematics associated with the optical transfer function, it deserves elucidation beyond just the bare demonstration of its validity and the indication of where it fits in OTF analysis.

Our method for developing concepts of what the convolution integral is all about is to show graphically the nature of certain functions and especially how the functions relate to each other.

The triangular function $f_2(x)$ shown in Fig. B.1*a* can be described as

$$\begin{aligned} f_2(x) &= x + 2 && \text{when} \quad -2 \le x \le -1, \\ f_2(x) &= -x/2 + 1/2 && \text{when} \quad -1 \le x \le +1, \\ f_2(x) &= 0 && \text{when} \quad x \le -2 \text{ and } x \ge +1. \end{aligned} \tag{B-70}$$

The same function is plotted in Fig. B.1b except that $-x$ is substituted everywhere for x:

$$\begin{aligned} f_2(-x) &= -x + 2 && \text{when} \quad +1 \le x \le +2, \\ f_2(-x) &= x/2 + 1/2 && \text{when} \quad -1 \le x \le +1, \\ f_2(-x) &= 0 && \text{when} \quad x \le -1 \text{ and } x \ge +2. \end{aligned} \tag{B-71}$$

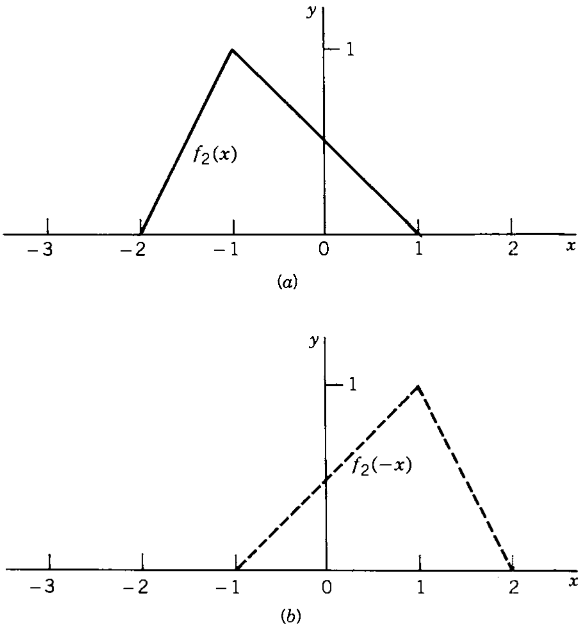


Figure B.1. Effect of reversing the sign of the independent variable in a particular triangular function.

Comparison of the two plots in Fig. B.1 shows that reversing the sign of the independent variable produces a "reflection" of the function in the vertical axis ($x = 0$). Another way of expressing this relation is to say the sign reversal folds back or rotates the function about the vertical axis, a sort of right-handed to left-handed transformation. This transformation is always performed on one of the two functions in a convolution.

If the triangular function is altered further by adding a constant x_n to the independent variable, the resulting function, $f_2(x_n - x)$, plots as shown in Fig. B.2. When this figure is compared with Fig. B.1*b*, it is apparent that the effect of adding a positive number to the independent variable is to shift the whole plot to the right by the amount of that number. So, by assigning different values to x_n, we can slide the function along the x-axis as we will. Experimentation with various values of x and x_n, positive and negative, will show that positive increments of x_n shift the plot in a positive direction by virtue of our having chosen to plot $f_2(-x)$ rather than $f_2(x)$ as the basic function subject to the shifting.

In Fig. B.3 two functions, $f_1(x)$ and $f_2(x)$, are plotted against x in (*a*) and (*b*), respectively. The value of each function is assumed to be zero for values of x at which no curve is shown. As was done for the triangular function in Fig. B.1, the sign of the independent variable for the second function is reversed and the result plotted as a broken line in Fig. B.3*b*. (In continuing the analogy with the triangular function, the added constant x_n is considered here to be zero.) A third function, $f_3(x)$, is defined as the product of $f_1(x)$ and $f_2(-x)$:

$$f_3(x) = f_1(x) f_2(-x). \tag{B-72}$$

The function curve in Fig. B.3*c* was plotted by measuring the function values of $f_1(x)$ and $f_2(-x)$ at each of a number of x values, calculating the product, and plotting the result as a point on the $f_3(x)$ curve. The shaded area under the curve in Fig. B.3*c*, whose area is A_0, represents the integral of $f_3(x)$. The

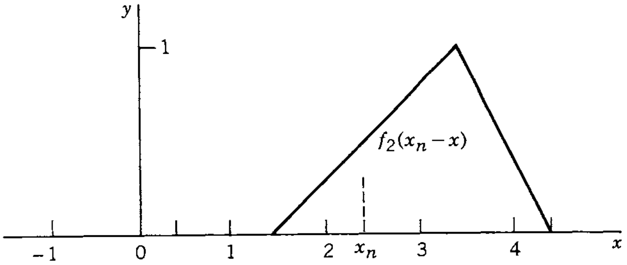


Figure B.2. Shifting a function by adding a constant to the independent variable.

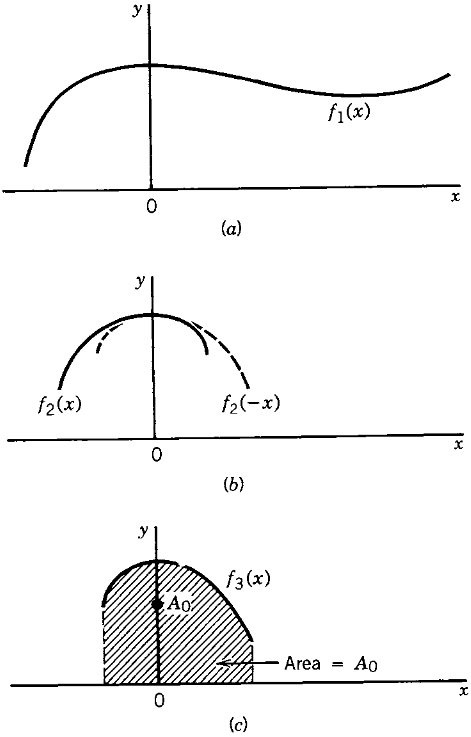


Figure B.3. Multiplying and integrating functions, $x_n = 0$.

graphical sequence described for Fig. B.3 can be stated mathematically as

$$A_0 = \int_{-\infty}^{+\infty} f_3(x)\, dx = \int_{-\infty}^{+\infty} f_1(x) f_2(-x)\, dx. \tag{B-73}$$

Because the independent variable for the second function can be regarded as $x_n - x$ with $x_n = 0$, the area value A_0 is plotted in Fig. B.3*c* at $x = 0$.

In Fig. B.4 the graphical sequence described for Fig. B.3 is repeated except that x_n has been given a positive, nonzero value, which moves the second factor function curve, $f_2(x_n - x)$, x_n units to the right. Because of the sag of the $f_1(x)$ curve in this region, the product curve, $f_3'(x)$, plots lower and has a slightly different shape compared with the corresponding curve in Fig. B.3. Consequently, the area A_n is smaller than A_0 as indicated by its plotted value at $x = x_n$. The graphical sequence for Fig. B.4 can be stated mathematically as

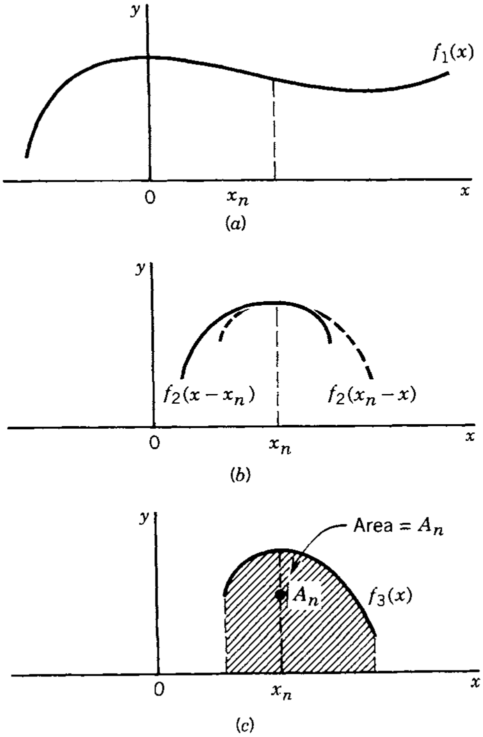


Figure B.4. Multiplying and integrating functions, $x_n \neq 0$.

$$A_n = \int_{-\infty}^{+\infty} f_3'(x)\, dx = \int_{-\infty}^{+\infty} f_1(x) f_2(x_n - x)\, dx. \qquad \text{(B-74)}$$

In the procedure described for finding A_n, it is obvious that this value is a function of x_n, so it can be designated $f_4(x_n)$:

$$f_4(x_n) = \int_{-\infty}^{+\infty} f_1(x) f_2(x_n - x)\, dx, \qquad \text{(B-75)}$$

which, except for the choice of symbols, is the convolution integral defined in Eq. (B-31).

One question that seems to cause some confusion in writing convolution expressions and in writing shifting Dirac delta functions is the order of terms in the argument or independent variable binomial; that is, should the expression be $(x_n - x)$ or $(x - x_n)$? One who has not learned the answer by rote sometimes finds reasons for one or the other quite elusive. One approach in connection with convolution is to make one of the two functions involved a delta function and then extend an answer thus gained to more general combinations. This we pursue in the following discussion.

From the earlier discussion of the Dirac delta function in this appendix, we know that it integrates to unity, and its value is zero except where its argument is zero. From these properties it follows that

$$\int_{-\infty}^{+\infty} f_1(x)\,\delta(x - x_n)\,dx = f_1(x_n), \tag{B-76}$$

because the delta function acts as a factor of unity only where $x - x_n = 0$ or $x = x_n$. At all other values of x, the delta function is zero. It is also true that

$$\int_{-\infty}^{+\infty} f_1(x)\,\delta(x_n - x)\,dx = f_1(x_n). \tag{B-77}$$

The integrals of Eqs. (B-76) and (B-77) are identical because $x - x_n = 0$ and $x_n - x = 0$ give the same answer: $x = x_n$. Thus, we know that the delta function is an even function; that is, its value is unchanged when the sign of its argument is reversed. However,

$$\int_{-\infty}^{+\infty} \delta(x) f_2(x - x_n)\,dx = f_2(-x_n), \tag{B-78}$$

but

$$\int_{-\infty}^{+\infty} \delta(x) f_2(x_n - x)\,dx = f_2(x_n). \tag{B-79}$$

So, unless the second function in the integrand happens to be an even function, the integral is changed when x and x_n are reversed in the binomial argument. When the integral of the product of two functions is to be the convolution of the two functions, the order of terms in the binomial argument of Eq. (B-79) is followed. The advantage seems obvious when one of the integrand factors is the delta function: The convolution function will vary with increasing values of the independent variable exactly as the convolved function except that it will be shifted along the axis. Since in OTF theory the delta function is the ideal limit of many of the functions that are convolved, it follows that a desirable treatment when the delta function is involved is likely to be preferred for the more general combinations as well [5].

CONVOLUTION AND SPREAD FUNCTIONS

From the properties of the delta function already discussed, we know that graphically it is a high spike of negligible width so proportioned that its area is unity. These idealized dimensions make the delta function a suitable representation of an ideal point or line source in optics. In the course of transferring energy from the source to the image, the optical system in one or more steps operates on the dimensions of the source to produce an image usually resembling but not identical to the source in size and energy distribution. In the convolution integral, the first function describes the source or object; and the second function is the transmission characteristic of the optical system. (Actually the roles of the two functions could be reversed without affecting the result because the order of functions in the convolution integral is commutative. See Eqs. (B-47) and (B-48).) Since the optical system tends to spread out the light from each point in the object, the second function is known as a *spread function*. Therefore, insofar as the object of an optical system is a true point or line (negligible width), that is, fairly represented by a delta function, a plot of the light distribution in the image is a graph of the optical system spread function as indicated by Eq. (B-79).

An actual source, of course, unlike a delta function, does have appreciable width as, for example, $f_1(x)$ of Fig. B.5. In this figure, $f_2(x)$ represents the optical system spread function. By mentally reviewing the graphical significance of the convolution process described in connection with Figs. B.1–B.4, one realizes that the convolution of $f_1(x)$ and $f_2(x)$ (which represents the image) resembles but is not exactly the same shape as the spread function. Also, it is evident that the convolution (image) function will always be broader than either of the other two functions.

So far, in showing how the convolution process represents the physical behavior of an optical imaging system, we have used relatively narrow objects. It is our purpose now to apply some of the concepts gained thus far to show how an extended object is transferred through an optical system to produce an image.

In Fig. B.6, $f_1(x)$ represents the object. Our approach is to consider this

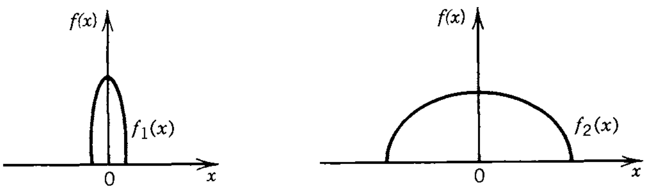


Figure B.5. Source and spread functions.

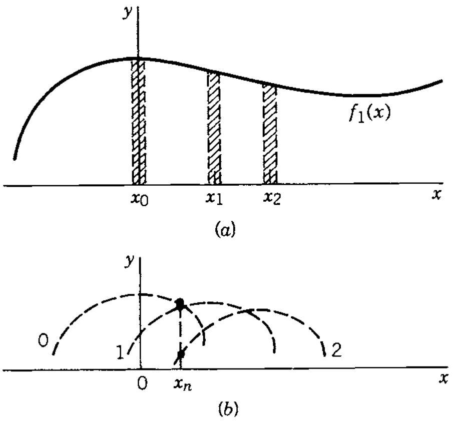


Figure B.6. Forming the image of an extended object by summing a large number of spread functions.

function as the sum of a large number of extremely narrow incremental functions, three of which are shown at x_0, x_1, and x_2. By visualizing how each increment would be imaged and then superimposing the incremental images, we can see how the total image is made up and possibly comprehend qualitatively how the shape of the spread function influences the object-to-image transfer.

Each of the incremental functions making up $f_1(x)$ resembles a delta function in being extremely narrow; but instead of having a uniform value of unity, each incremental area is proportional to the value of the function at its x. Like the delta function, each incremental function in combination with the spread function produces a convolution function shaped like the spread function; however, these images are located along the x-axis according to the respective x values of their originating incremental functions, and their corresponding ordinates are proportional to the heights of their originating incremental functions. Three convolution (image) functions corresponding to the three incremental functions shown are plotted in Fig. B.6*b*. The ordinate for the complete extended image at some $x = x_n$ would be the sum of all incremental x_n ordinates, only three of which are indicated in the figure. A complete diagram in Fig. B.6, of course, would require that the area under $f_1(x)$ be solidly packed with incremental functions, each infinitesimally wide, resulting in a corresponding number of infinitesimally spaced image components in Fig. B.6*b*.

OTHER CONVOLUTION INTEGRALS

If $f_1(x)$ is the same function as $f_2(x)$ for the convolution integral of Eq. (B-75), the equation can be written

$$f_4(x_n) = \int_{-\infty}^{+\infty} f_1(x) f_1(x_n - x)\, dx, \tag{B-80}$$

and the expression is called the *self-convolution* integral.

Although all the discussion about convolution thus far in this appendix has involved only real functions, the convolution integral can also be written in terms of complex functions:

$$\hat{f}_4(x_n) = \int_{-\infty}^{+\infty} \hat{f}_1(x) \hat{f}_2(x_n - x)\, dx. \tag{B-81}$$

The family of correlation functions, discussed in the next section, resembles in a number of respects the group of convolution functions discussed in this and the previous two sections; but important differences also exist, which will be taken up in the next section.

THE CORRELATION FUNCTION

In the mathematics associated with the optical transfer function, one often encounters the *correlation integral*, which is expressed as

$$g(x_n) = \int_{-\infty}^{+\infty} g_1(x)\, g_2(x - x_n)\, dx. \tag{B-82}$$

The parameter x_n typically has a range over both positive and negative values. The form of Eq. (B-82) is identical to that of the convolution integral, Eq. (B-75), except that the sign of the binomial argument of the second function is reversed. Except for this reversal the graphical representation to illustrate the evaluation of the correlation integral would be similar to the one described earlier in detail for the convolution integral. Again, the graphical effect of the binomial argument is to slide the second function along the x-axis relative to the first function with which it is being correlated. Each of the relative positions, of course, represents a particular value of x_n.

Like the convolution integral, the correlation integral can be expressed in terms of complex functions:

$$\hat{g}(x_n) = \int_{-\infty}^{+\infty} \hat{g}_1(x)\, \hat{g}_2^*(x - x_n)\, dx. \tag{B-83}$$

However, as indicated by the asterisk, the complex conjugate of $g_2(x)$ is substituted for $g_2(x)$ in the integrand. When $g_1(x)$ and $g_2(x)$ are the same function, the integral given in real and complex form in Eqs. (B-82) and (B-83), respectively, becomes the autocorrelation function:

$$\hat{g}(x_n) = \int_{-\infty}^{+\infty} \hat{g}_1(x)\, \hat{g}_1^*(x - x_n)\, dx. \tag{B-84}$$

In both the convolution and the correlation integrals, if the second function is even, that is, $f_2(x) = f_2(-x)$ or $g_2(x) = g_2(-x)$, the sign of the binomial argument has no significance—the integral has the same value whether the argument is written $(x_n - x)$ or $(x - x_n)$. So, if one of two functions is both real and even, the convolution and the correlation integrals formed from the two functions are identical.

When the self-convolution is found of a complex function,

$$\hat{f}(x_n) = \int_{-\infty}^{+\infty} \hat{f}_1(x)\, \hat{f}_1(x_n - x)\, dx, \tag{B-85}$$

the phase information in $\hat{f}_1(x)$ is retained in the self-convolution. On the other hand, when the autocorrelation function is found for a complex function as in Eq. (B-84), the phase information in $\hat{g}_1(x)$ is lost. In general, therefore, one would expect the self-convolution $\hat{f}(x_n)$ and the autocorrelation $\hat{g}(x_n)$ of a given complex function to differ. There is one notable exception. A Fourier transform of a real function, that is, $\hat{F}(\omega) \leftrightarrow f(x)$, is found in Fourier integral theory to have *complex symmetry*, which means

$$\hat{F}(-\omega) = \hat{F}^*(\omega). \tag{B-86}$$

Comparison of Eq. (B-85) with Eq. (B-84) shows that if $\hat{F}(\omega)$ were substituted for both $\hat{f}_1(x)$ and $\hat{g}_1(x)$, $\hat{f}(x_n)$ and $\hat{g}(x_n)$ would be identical functions.

Although the scope of the present work does not permit further developments in these fields, the topics already touched on in convolution theory and correlation theory probably indicate to the reader that a complete treatment would have considerable depth. References are cited at the end of this appendix to aid in broader and more thorough study of these interesting areas.

EXAMPLES

Example 1. This example applies the convolution integral of Eq. (B-75) to the two identical functions, $f_1(x)$ and $f_2(x)$, shown in Fig. B.7. (The exercise is

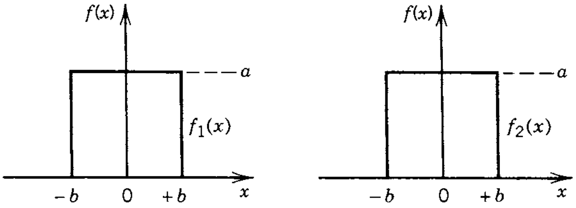


Figure B.7. Two identical rectangular functions.

the same as finding the self-convolution of either function.) Equations for the two functions are

$$f_1(x) = f_2(x) = a \quad \text{when} \quad -b \le x \le +b,$$

$$f_1(x) = f_2(x) = 0 \quad \text{when} \quad x < -b \quad \text{and} \quad x > +b. \qquad \text{(B-87)}$$

Since the functions are even, no pains have to be taken to reverse one of the functions to accommodate the sign of the binomial argument in the integrand of Eq. (B-75). As in our earlier discussion of the significance of the convolution integral, the graphical portrayal of this convolution has to convey the sliding of one of the functions, say $f_2(x)$, along the x-axis in accordance with the value of the shifting parameter x_n. As the positions of $f_2(x_n - x)$ relative to $f_1(x)$ are noted, it is apparent that for $|x_n| > 2b$ the two functions do not overlap; and the convolution integral is, consequently, zero. However, for all values of x_n in the range where $|x_n| < 2b$, the two functions do overlap, resulting in nonzero values for the integral. Since each function has the value a over its nonzero range, the product in the integrand through the overlapping range is a^2, and the value of the integral is a^2 times the overlap interval. In Fig. B.8, the overlap

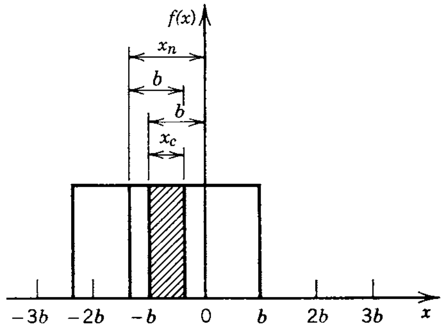


Figure B.8. Overlap of two rectangular functions being convolved.

interval is shown as x_c, and the following relations can be written from the geometry of the figure:

$$x_c = 2b + x_n \quad \text{when} \quad -2b \le x_n \le 0,$$
$$x_c = 2b - x_n \quad \text{when} \quad 0 \le x_n \le +2b. \tag{B-88}$$

Because the convolution function (integral) $f_4(x_n)$ is equal to a^2x_c, the relations of Eq. (B-88) can be substituted to give

$$\begin{aligned} f_4(x_n) &= a^2x_c = 2a^2b + a^2x_n && \text{when} \quad -2b \le x_n \le 0, \\ f_4(x_n) &= a^2x_c = 2a^2b - a^2x_n && \text{when} \quad 0 \le x_n \le +2b, \\ f_4(x_n) &= 0 && \text{when} \quad x_n \le -2b \quad \text{and} \quad x_n \ge +2b. \end{aligned} \tag{B-89}$$

When these equations are plotted in Fig. B.9, the convolution is found to be a triangular function with its base extending from $x_n = -2b$ to $x_n = +2b$.

Example 2. A convolution that is of particular value in optics is two dimensional and involves the overlapping area of two circles:

$$f_4(x_n, y_m) = \iint_{\mathcal{Q}} f_1(x, y) f_2(x_n - x, y_m - y)\, dx\, dy, \tag{B-90}$$

where the single symbol $\mathcal{Q}$ takes the place of the negative to positive infinity limits usually placed on each of the two variables of integration, x and y. Like the previous example, the two functions being convolved have a constant, nonzero value only in a defined region (in this instance a circle of radius r for each function), and their value in the rest of space is zero. It is, therefore, immaterial whether one indicates that integration is to take place in all of space or just in

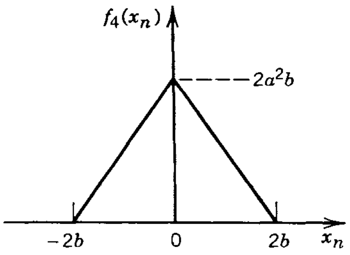


Figure B.9. Convolution function of two rectangular functions.

the region where both functions simultaneously have nonzero values. This region in the present example is the overlapping area $\mathfrak{A}$ of the two circles. The integrand in all the rest of space is zero because at least one of the factor functions is zero.

If the value of each function within its circle is assumed to be A, the two functions can be expressed as

$$\begin{aligned} f_1(x, y) &= f_2(x, y) = A \quad \text{when} \quad x^2 + y^2 \leq r^2, \\ f_1(x, y) &= f_2(x, y) = 0 \quad \text{when} \quad x^2 + y^2 > r^2. \end{aligned} \tag{B-91}$$

Graphically the convolution of $f_1(x, y)$ and $f_2(x, y)$ is represented by Fig. B.10 where $f_2(x_n - x, y_m - y)$ is shown being shifted with respect to $f_1(x, y)$ in two dimensions, the angle φ made by the line between centers and the x-axis being determined by the relative values of x_n and y_m. For all combinations of these shifting variables where $(x_n^2 + y_m^2)^{1/2} < 2r$ there will be an overlapping area, and the convolution will have a nonzero value. For all other combinations, the convolution is zero.

In Example 1, being one-dimensional, one axis in the graphical representations, Figs. B.7 and B.8, could be used to show the value of the function. In the present example in Fig. B.10, both axes have to be used for the independent variables; so the constant function value A within the circles would have to be shown on a third axis perpendicular to the diagram (z-axis) making the three-dimensional representation a pair of right cylinders sharing a common volume where the cross hatching is shown in Fig. B.10.

The present example can be reduced to a one-dimensional problem by rotating the x- and y-axes through the angle φ so that the x-axis coincides with the line between circle centers as in Fig. B.11. Then the convolution function becomes

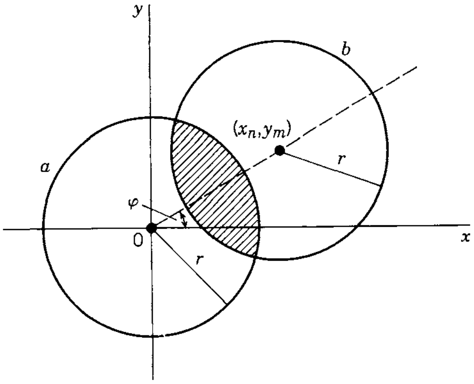


Figure B.10. Two-dimensional convolution—overlap of the nonzero regions of two functions.

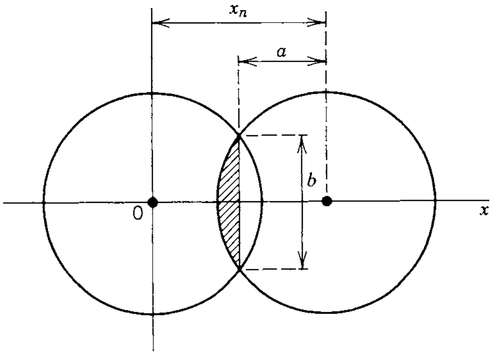


Figure B.11. Rotation of the axes in Fig. B.10.

$$f_4(x_n) = \iint_{\alpha} f_1(x, y) f_2(x_n - x, y)\, dx\, dy. \qquad \text{(B-92)}$$

Within the overlapping areas of the two circles, the product of the two integrand functions is A^2. Evaluating the integral, that is, the convolution function $f_4(x_n)$, consists of finding an expression for the overlapping area in terms of x_n and multiplying it by the constant A^2. Half of the overlapping area, which is the segment of one circle, is shown cross hatched in Fig. B.11. Applying a standard mensuration formula, we can write

$$\text{Segment area} = r^2 \cos^{-1}(a/r) - a\sqrt{r^2 - a^2}. \qquad \text{(B-93)}$$

From Fig. B.11 it is apparent that $a = x_n/2$; so the convolution function is

$$f_4(x_n) = 2A^2\left\{r^2 \cos^{-1}(x_n/2r) - (x_n/2)\left[r^2 - (x_n^2/4)\right]^{1/2}\right\}, \qquad \text{(B-94)}$$

where the normal range of x_n is from zero to $2r$.

If we had not rotated the axes and had derived an expression for $f_4(x_n, y_m)$ based on Fig. B.10, the overlapping area would have had to be written in terms of the distance D between centers instead of in terms of x_n. This would require only that D be substituted for x_n on the right side of Eq. (B-94). Its value in terms of the old x_n and y_m corresponding to Fig. B.10 would be

$$D = (x_n^2 + y_m^2)^{1/2}. \qquad \text{(B-95)}$$

A plot of $f_4(x_n)$ in accordance with Eq. (B-94) is shown in Fig. B.12. As indicated, the scales on both axes have been normalized. On the vertical scale, $f_4(x_n)$ has been divided by $\pi r^2 A^2$:

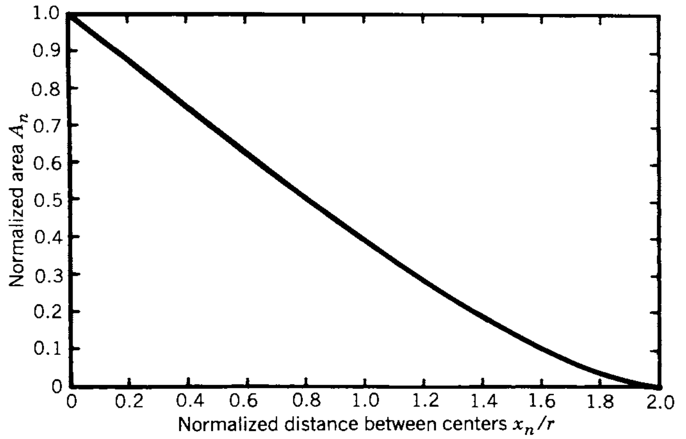


Figure B.12. Convolution of two functions having two independent variables.

$$\begin{aligned} A_n &= \left[1/(\pi r^2 A^2)\right] f_4(x_n) \\ &= 2/\pi \left\{\cos^{-1}(x_n/2r) - (x_n/2r)\left[1 - (x_n/2r)^2\right]^{1/2}\right\}. \qquad \text{(B-96)} \end{aligned}$$

As one considers ranges of values, it is evident that the quantity in braces goes from $\pi/2$ when $x_n = 0$ to zero when $x_n = 2r$ or when the normalized distance $x_n/r = 2$. The corresponding range of the normalized area A_n is from unity to zero.

REFERENCES

1. M. Born and E. Wolf, *Principles of Optics*, 3rd ed. Pergamon, Oxford, 1965, Appendix IV.
2. A. Papoulis, *The Fourier Integral and Its Application*. McGraw-Hill, New York, 1962, pp. 36–39, 270–283.
3. E. C. Titchmarsh, *Introduction to the Theory of Fourier Integrals*, 2d ed. McGraw-Hill, New York, 1962.
4. R. Bracewell, *The Fourier Transform and Its Applications*, McGraw-Hill, New York, 1965.
5. T. P. Sheahen, Importance of Proper Phase Analysis in Using Fourier Transforms. *Am. J. Phys.* **44**, 22(1976).
6. W. T. Cathey, *Optical Information Processing and Holography*. Wiley, New York, 1974, pp. 24–29.

Appendix C

Diffraction Integral Fundamentals

INTRODUCTION

The development of the diffraction integral, Eq. (4-22), in Chapter 4, ''Diffraction Integral and Wave-Front Aberration Function,'' assumes that the reader has some familiarity with traveling wave theory. However, for those who would like a little more exposure to the fundamentals underlying the diffraction integral, this appendix has been added.

THE TRAVELING WAVE EQUATION

Light has long been recognized as a wavelength region in the electromagnetic spectrum, which also includes, among others, radio, radar, infrared, ultraviolet, X, and gamma rays. Some of these, including light, have a complementary particle nature, which does not bear directly on diffraction phenomena and, therefore, will not be discussed further in this appendix.

In the theory of electromagnetism, Maxwell's equations have become so basic and so firmly established that the study of wave phenomena is started with them. They are, in vector notation,

$$\nabla \times \mathbf{E} = -\partial \mathbf{B}/\partial t, \tag{C-1}$$

$$\nabla \times \mathbf{H} = \mathbf{j} + \partial \mathbf{D}/\partial t. \tag{C-2}$$

Two additional equations, which are often included with Maxwell's equations, may be derived from these by assuming that electric charge is conserved:

$$\nabla \cdot \mathbf{B} = 0, \tag{C-3}$$

$$\nabla \cdot \mathbf{D} = q_{\mathrm{d}}. \tag{C-4}$$

In these equations, **E**, **D**, **H**, and **B** are the field vectors (electric field strength, electric induction, magnetic field strength, and magnetic induction, respectively), $\mathbf{j}$ is current density, and q_{d} is electric charge density. We assume the equations to be valid, that is, that they form a self-consistent set whose predictions concerning static and slowly varying (nonrelativistic) fields are in agreement with experimental data.

The four field vectors can also be related by using three material parameters: ϵ, μ, and σ, which are called permittivity, permeability, and conductivity, respectively. These quantities, which characterize the medium, are point functions constant in time. They may be scalars or, for anisotropic materials, tensors. Their values in a given medium depend upon the frequency components constituting the wave disturbance. The relations between the field vectors established by the material parameters may be linear or nonlinear, the latter occurring particularly in connection with lasers and their high-intensity, coherent beams. In the present text we assume linearity and isotropy; so the relations are

$$\mathbf{D} = \epsilon\mathbf{E}, \tag{C-5}$$

$$\mathbf{B} = \mu\mathbf{H}, \tag{C-6}$$

$$\mathbf{j} = \sigma\mathbf{E}. \tag{C-7}$$

If we assume that a solid material has no free charges (i.e., no electrons are free to move throughout the body of material so that σ is everywhere zero) the material is called a dielectric. If σ is zero, the current density $\mathbf{j}$ also is zero. Any accumulated excess charge q will be a static charge and will neither contribute to nor respond to changes in field. Then the four "Maxwell's equations," by our assuming isotropy and homogeneity, are

$$\operatorname{curl}\mathbf{E} \equiv \nabla \times \mathbf{E} = -\mu(\partial\mathbf{H}/\partial t), \tag{C-8}$$

$$\operatorname{curl}\mathbf{H} \equiv \nabla \times \mathbf{H} = \epsilon(\partial\mathbf{E}/\partial t), \tag{C-9}$$

$$\operatorname{div}\mathbf{H} \equiv \nabla \cdot \mathbf{H} = 0, \tag{C-10}$$

$$\operatorname{div}\mathbf{E} \equiv \nabla \cdot \mathbf{E} = 0, \tag{C-11}$$

where the symbol "$\equiv$" signifies "equals by definition." To derive *wave equations* from Maxwell's equations, the following vector identity can be applied:

$$\operatorname{curl}\operatorname{curl}\mathbf{V} = \operatorname{grad}\operatorname{div}\mathbf{V} - \nabla^2\mathbf{V}, \tag{C-12}$$

which can be verified by expanding both sides according to the vector definitions of the symbols (curl, grad, div, and ∇) applied to the general vector $\mathbf{V}$. Because of Eqs. (C-10) and (C-11), Eq. (C-12) simplifies to the following for $\mathbf{E}$ and $\mathbf{H}$:

$$\text{curl curl } \mathbf{E} = -\nabla^2\mathbf{E}, \tag{C-13}$$

$$\text{curl curl } \mathbf{H} = -\nabla^2\mathbf{H}. \tag{C-14}$$

By taking the curl of both sides of Eqs. (C-8) and (C-9), we obtain

$$\text{curl curl } \mathbf{E} = -\mu\ \partial/\partial t(\text{curl } \mathbf{H}), \tag{C-15}$$

$$\text{curl curl } \mathbf{H} = \epsilon\ \partial/\partial t(\text{curl } \mathbf{E}). \tag{C-16}$$

Substituting from Eqs. (C-13) and (C-14) in these equations, we find

$$\nabla^2\mathbf{E} = \mu\ \partial/\partial t(\text{curl } \mathbf{H}), \tag{C-17}$$

$$\nabla^2\mathbf{H} = -\epsilon\ \partial/\partial t(\textit{curl } \mathbf{E}). \tag{C-18}$$

By taking the partial derivatives with respect to time of both sides of Eqs. (C-8) and (C-9) we obtain

$$\partial/\partial t(\text{curl } \mathbf{E}) = -\mu(\partial^2\mathbf{H}/\partial t^2), \tag{C-19}$$

$$\partial/\partial t(\text{curl } \mathbf{H}) = \epsilon(\partial^2\mathbf{E}/\partial t^2). \tag{C-20}$$

By substituting from Eqs. (C-17) and (C-18) and rearranging terms in (C-19) and (C-20), the following wave equations result:

$$\nabla^2\mathbf{H} = \epsilon\mu(\partial^2\mathbf{H}/\partial t^2), \tag{C-21}$$

$$\nabla^2\mathbf{E} = \epsilon\mu(\partial^2\mathbf{E}/\partial t^2). \tag{C-22}$$

In a three-dimensional coordinate system, six scalar wave equations are represented by Eqs. (C-21) and (C-22) and are of the form

$$\nabla^2 U = \epsilon\mu(\partial^2 U/\partial t^2), \tag{C-23}$$

where U, in a rectangular coordinate system, for instance, is any one of the field components E_x, E_y, E_z, H_x, H_y, and H_z. The general solution of Eq. (C-23) is

$$U(\mathbf{r}, t) = U_a[k\zeta(\mathbf{r}) + vt] + U_b[k\zeta(\mathbf{r}) - vt], \tag{C-24}$$

where U_a and U_b are arbitrary functions of the indicated arguments, $\mathbf{r}$ is a position vector from the origin to a point (x, y, z), k is a *propagation constant*, $\zeta(\mathbf{r})$ is a real scalar quantity, a function of position, and v is a speed.

To comprehend the traveling wave nature of Eq. (C-24), each of the arguments is considered at some time t_1 and a later time t_2. If the argument is to remain constant so that U_a or U_b, remains unchanged, the $\mathbf{r}_1$, corresponding to t_1, must change to a compensating $\mathbf{r}_2$. The displacement represented by $\Delta\mathbf{r} = (\mathbf{r}_2 - \mathbf{r}_1)$ during the time interval $\Delta t = (t_2 - t_1)$ indicates that the assumed fixed field configuration of U_a or U_b is a *traveling wave*. The difference between the two functions U_a and U_b is that in one instance the wave is traveling in a direction to increase $\mathbf{r}$ and in the other the wave is traveling in a direction to reduce $\mathbf{r}$. To define the functions U_a, U_b, and $\zeta(\mathbf{r})$, *initial* or *boundary* conditions and the physics of the space medium have to be utilized. The solution requires ingenuity rather than knowledge of some routine approach.

Rather than attacking Eq. (C-24) directly as outlined above, it is usually more productive to recognize that OTF problems generally involve steady-state beams that can be resolved into sinusoidal components, that is, sine or cosine functions. Although these trigonometric functions can be introduced directly as real functions into the solutions of Eq. (C-23), common practice is to use the complex identity,

$$e^{\pm i\theta} \equiv \exp \pm i\theta = \cos\theta \pm i \sin\theta, \tag{C-25}$$

and to express the solution of Eq. (C-23) as the sum of two particular complex solutions,

$$\begin{aligned}\hat{U} &= U_a \exp(i\mathbf{k}\cdot\mathbf{r} + i2\pi\nu t) + U_b \exp(-i\mathbf{k}\cdot\mathbf{r} + i2\pi\nu t)\\ &= \hat{U}_a \exp(i2\pi\nu t) + \hat{U}_b \exp(i2\pi\nu t)\\ &= (\hat{U}_a + \hat{U}_b)\exp(i2\pi\nu t),\end{aligned} \tag{C-26}$$

where ν is the time frequency of the light and

$$\hat{U}_a = U_a \exp(+i\mathbf{k}\cdot\mathbf{r}), \tag{C-27}$$

$$\hat{U}_b = U_b \exp(-i\mathbf{k}\cdot\mathbf{r}). \tag{C-28}$$

U_a and U_b are real scalar space functions determined by initial or boundary conditions of each particular problem. The propagation vector $\mathbf{k}$ has a magnitude k and the direction of wave propagation, for example, in a rectangular coordinate system:

$$\begin{aligned}\mathbf{k} &= \mathbf{I}_x k_x + \mathbf{I}_y k_y + \mathbf{I}_z k_z,\\ k^2 &= k_x^2 + k_y^2 + k_z^2.\end{aligned} \tag{C-29}$$

$\mathbf{I}_x$, $\mathbf{I}_y$, and $\mathbf{I}_z$ are unit vectors parallel to the coordinate axes. Substitution of Eq. (C-26) into Eq. (C-23) gives

$$\nabla^2(\hat{U}_a + \hat{U}_b) + \epsilon\mu(2\pi\nu)^2(\hat{U}_a + \hat{U}_b) = 0. \qquad \text{(C-30)}$$

This is known as the Helmholtz equation. Again, as in the discussion of Eq. (C-24), $\hat{U}_a$ and $\hat{U}_b$ describe waves that are traveling in opposite directions. One or the other or both have to be chosen according to the conditions of the problem. A wave function that satisfies the wave equation must be both a space and a time function. Whenever the time function can be separated out as in Eq. (C-26), the remaining space function need satisfy only the Helmholtz equation. *A separable time function is usually omitted (suppressed) in wave theory problems.*

SPHERICAL WAVE FRONTS

Our discussion of diffraction involves the image space between the exit pupil and the image plane. The space is assumed isotropic and homogeneous. The wave fronts considered are spherical. Functions are sinusoidal and for a single frequency; and the complete notation discussed in the previous section, with the time function suppressed, is used to describe the traveling light wave. First, a function is set up consistent with the assumed boundary conditions, and then its legitimacy as a traveling wave is checked by seeing whether it satisfies Eq. (C-30).

The simplest concept of a spherical wave is one originating at a point source radiating uniformly in all directions of three-dimensional space (Fig. C.1). As in the previous discussion of traveling waves, this wave has a counterpart, which is a spherical wave traveling toward, rather than away from, the center (Fig. C.2). This concept, modified by limiting the extent of the wave front, corresponds to the idealized wave emerging from the exit pupil and converging to the focal point of an optical system.

Having suppressed the time function of the traveling wave, we retain the part that describes the distribution of the wave in space. When the whole sphere is considered, symmetry about the center dictates that field values depend only on the radial distance r; no variation in these values can be expected when either the azimuthal angle ϕ or the polar angle θ (Fig. C.3) is varied. From the definition of the wavelength λ, the angular argument of the sinusoidal functions (sine and cosine) must go through 2π radians for each λ increment of r in free space. In general for a transmission medium with an index of refraction n', the actual wavelength is $\lambda' = \lambda/n'$; so the argument becomes $2\pi(r/\lambda')$ or

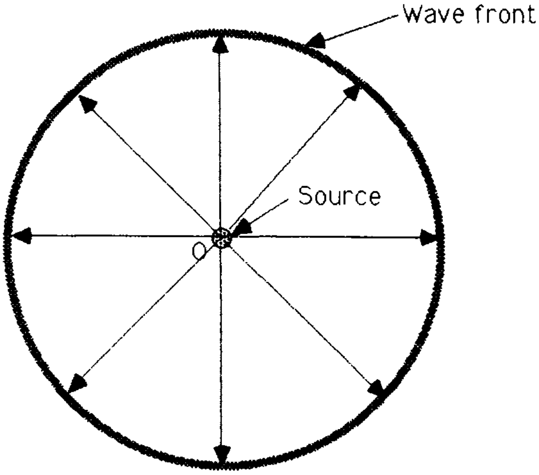


Figure C.1. Cross section through the center of a spherical wave system, energy originating at a point source at the center.

$2\pi(n'r/\lambda)$. Throughout this text, this expression has been shortened by defining $k \equiv 2\pi/\lambda$, which allows the argument to be expressed as $kn'r$. The complex expression to describe the sinusoidal distribution in space then becomes $\exp(\pm ikn'r)$. The choice of sign depends upon the variation of the argument relative to the argument of the suppressed time function, $\exp(2\pi\nu t)$. Tempo-

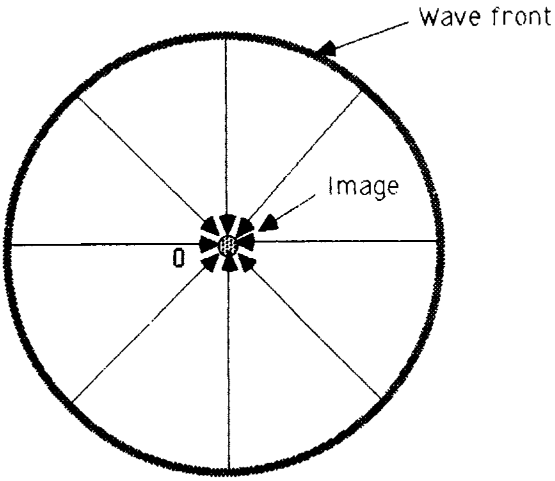


Figure C.2. Spherical wave system in which external energy is uniformly directed toward the center.

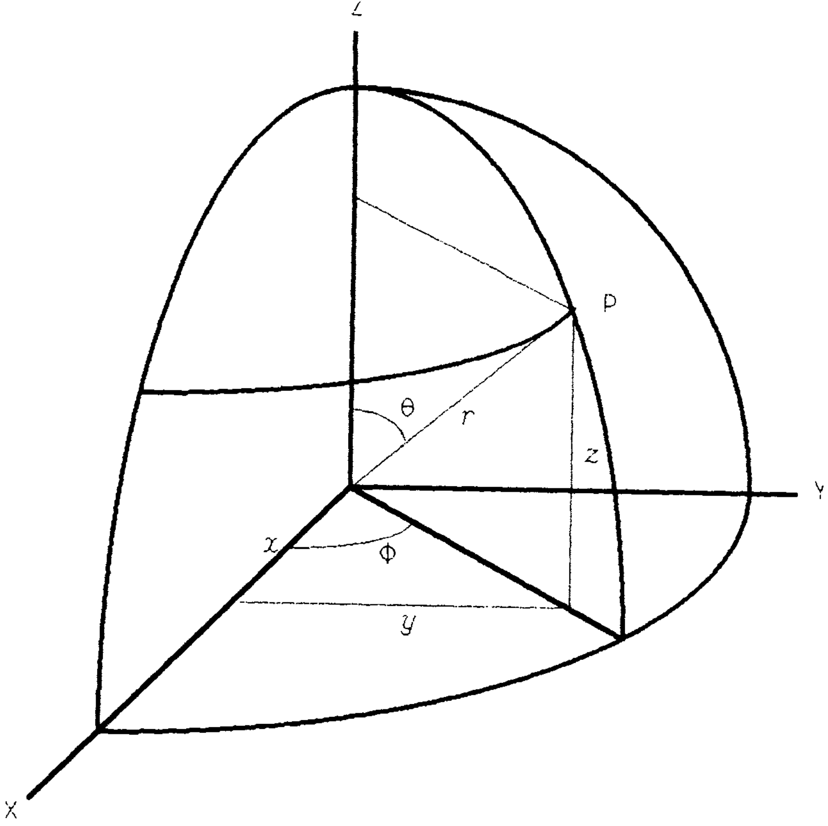


Figure C.3. Spherical coordinate system.

rarily restoring the time function gives the complete expression $\exp(2\pi\nu t \pm ikn'r)$. When the traveling wave reasoning of the previous section is applied to this argument, it becomes apparent that the positive sign applies to a spherical wave traveling toward the center; and the negative sign applies to a spherical wave traveling away from the center.

Having determined cyclic variation of the field value with the space variable r, we still have to find the coefficient of the sinusoidal function, that is, the amplitude variation of the field value with r.

When energy is conserved in the traveling wave, the flux density on a spherical wave front varies inversely as the total area of the spherical surface. This area varies directly with the radius squared, r^2, so we are affirming the "inverse square law," which states that the flux density in this configuration is inversely proportional to r^2. However, the desired expression is for the amplitude function $\hat{U}(r)$, which is proportional to the square root of the flux density; hence the amplitude varies inversely with the radius r. The coefficient, then, of the complex sinusoidal function is the ratio G/r where G is the amplitude at unit distance from the center. Therefore, the complete expression for a spherical wave traveling toward the center is

$$\hat{U} = \hat{U}(r) = \left[G \exp(ikn'r)\right]/r, \tag{C-31}$$

which corresponds, with appropriate changes in nomenclature, to Eq. (4-1) of Chapter 4.

The question remains whether Eq. (C-31) is a legitimate expression for a traveling wave according to Eq. (C-30); that is, does the expression for $\hat{U}$ satisfy Eq. (C-32)?

$$\nabla^2 \hat{U} + \epsilon\mu(2\pi\nu)^2 \hat{U} = 0. \tag{C-32}$$

From vector analysis, ∇^2, known generally as the *Laplacian*, is expressed in rectangular and polar coordinate systems as

$$\begin{aligned}\nabla^2 &= \partial^2/\partial x^2 + \partial^2/\partial y^2 + \partial^2/\partial z^2 \\ &= \left\{1/(r^2 \sin\phi)\right\}\left\{(\sin\phi)(\partial/\partial r)\left[r^2(\partial/\partial r)\right] + (\partial/\partial\theta)\left[\sin\theta)(\partial/\partial\theta)\right] \right. \\ &\quad \left. + (1/\sin\theta)(\partial^2/\partial\phi^2)\right\}. \end{aligned} \tag{C-33}$$

Because the functions of interest in the present discussion are independent of the polar coordinates ϕ and θ, the Laplacian reduces to

$$\nabla^2 = \left\{1/r^2\right\}\left\{(\partial/\partial r)\left[r^2(\partial/\partial r)\right]\right\}. \tag{C-34}$$

Applying this operator to the $\hat{U}$ of Eq. (C-31), we find

$$\nabla^2 \hat{U} = -(G/r)(kn')^2 \exp(ikn'r). \tag{C-35}$$

When the second term of Eq. (C-32) is added to the above expression for $\nabla^2 \hat{U}$, and the complete expression is algebraically simplified:

$$-(kn')^2 + \epsilon\mu(2\pi\nu)^2 = 0. \tag{C-36}$$

By substituting the defined ratio $2\pi/\lambda$ for k, and simplifying the equation further, we obtain

$$(\eta'/\lambda)^2 = \epsilon\mu\nu^2. \tag{C-37}$$

Since a traveling wave moves one actual wavelength $\lambda' = \lambda/\eta'$ during the time period $1/\nu$, the speed of the wave in the direction of propagation (radially for the assumed spherical waves) is $c' = (\lambda/n')/(1/\nu)$. From Eq. (C-37), this ratio is

$$c' = \sqrt{1/\epsilon\mu}. \qquad \text{(C-38)}$$

We conclude that our expression for a spherical wave traveling toward the center, Eq. (C-31), does indeed turn out to be a solution of Helmholtz equation, Eq. (C-32). Similarly, the following equation for a spherical wave traveling *away* from the center can be shown also to be a solution of Helmholtz equation:

$$\hat{U} = \hat{U}(r) = [G\exp(-ikn'r)]/r. \qquad \text{(C-39)}$$

APPLICATION OF THE HUYGENS–FRESNEL PRINCIPLE TO A SPHERICAL WAVE FRONT

In 1690 Huygens, the first proponent of the wave theory of light, published a rule for the construction of a set of surfaces that are "optically parallel" to each other. It states that each element of a wave front may be regarded as the center of a secondary disturbance that gives rise to spherical wavelets; the position of the wave front at subsequent times is the envelope of all such wavelets. By illustrating a few of the unlimited number of wavelets on a wave front, Fig. C.4 indicates Huygens' construction.

Utilization of Huygens' construction to explain the puzzling phenomenon of diffraction did not occur until about 1818 when Fresnel showed that Huygens' concept, combined with the principle of interference, provided the means for understanding diffraction. Later (1882) Kirchhoff put Fresnel's analysis on a sound mathematical basis.

With reference to Fig. C.5, Fresnel's concern was how a secondary wavelet originating at a point Q on the primary wave front contributed to the light field at point P in what we would call image space. Huygens' wavelet being approximately a hemisphere, it could be expected to differ in properties from the spherical waves that are discussed in the previous section. In particular, the amplitude may well be a function of the angle χ in Fig. C.5 instead of uniform about the center as in spherical waves. In the wavelet, Fresnel assumed that the *inclination factor* $K(\chi)$ applied to the amplitude of the corresponding spherical wave would be unity at $\chi = 0$ and would decline to zero at $\chi = \pi/2$ radians. To check his assumptions for a spherical wave front, Fresnel used a construction similar to Fig. C.6. He divided the spherical wave front into zones Z_1, Z_2, Z_3,

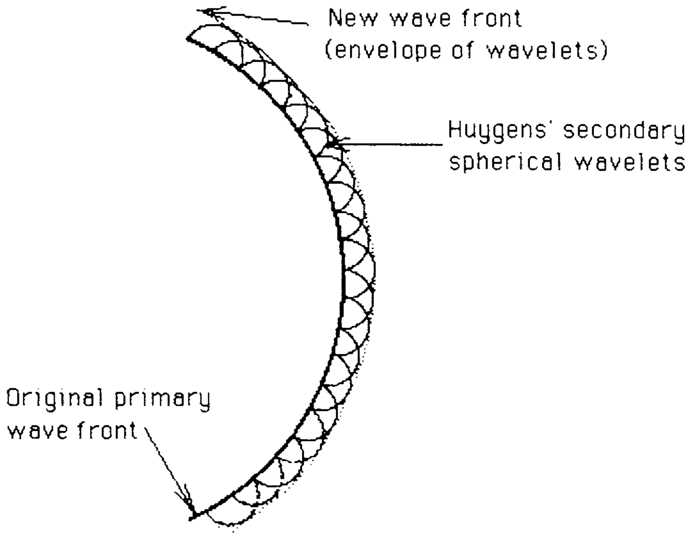


Figure C.4. Huygens' construction.

$Z_4, \ldots$ and rationalized a series of values for $K(\chi)$: $K_1, K_2, K_3, K_4, \ldots,$ corresponding to the sequence of zones. He wrote an expression for the amplitude at the point Q (see Eq. (C-39)):

$$\hat{U}(Q) = \left[G \exp(-ikn'r)\right]/r. \tag{C-40}$$

Then he applied the wavelet concept to find the amplitude at P due to the wavelet from each zone:

$$d\hat{U}(P) = \left\{\left[\hat{U}(Q)\, K(\chi) \exp(-ikn'l)\right]/l\right\} d\sigma. \tag{C-41}$$

When Eqs. (C-40) and (C-41) are combined,

$$\hat{U}(P) = \left[G \exp(-ikn'r)\right]/r \iint_{\alpha} \left\{\left[\exp(-ikn'l)\right]/l\right\} K(\chi)\, d\sigma. \tag{C-42}$$

Since no analytical expression was available for $K(\chi)$, the integral was evaluated by a special summation (involving a number of rationalizing assumptions of only historical interest) [1], which gave the following result:

$$\hat{U}(P) = -i\lambda K_1 \left\{G \exp\left[-ikn'(r + r')\right]\right\}/(r + r'), \tag{C-43}$$

where $K(\chi) = K_1$, the inclination factor of the first zone, Z_1.

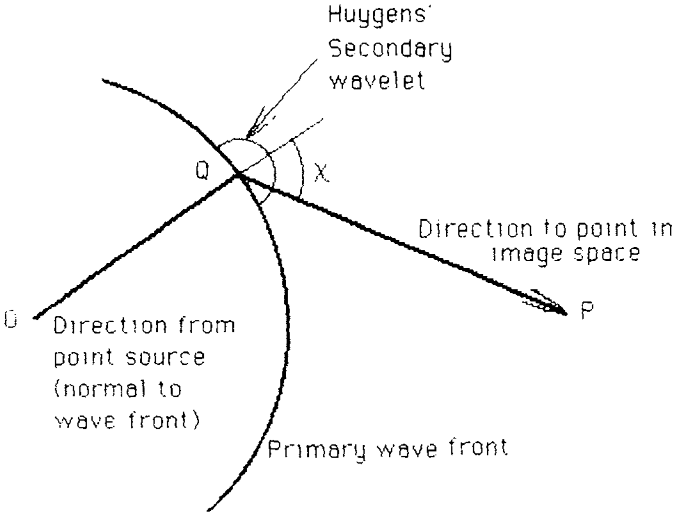

Huygens' Secondary wavelet
Q
X
Direction to point in image space
P
O
Direction from point source (normal to wave front)
Primary wave front

Figure C.5. Single secondary wavelet showing the inclination angle χ.

To verify the above wavelet approach, $\hat{U}(P)$ was evaluated also as a simple spherical wave-front problem (Eq. (C-39)) with the center at 0 and the spherical wave front through the point P:

$$\hat{U}(P) = \left\{G\exp\left[-ikn'(r + r')\right]\right\}/(r + r'). \qquad \text{(C-44)}$$

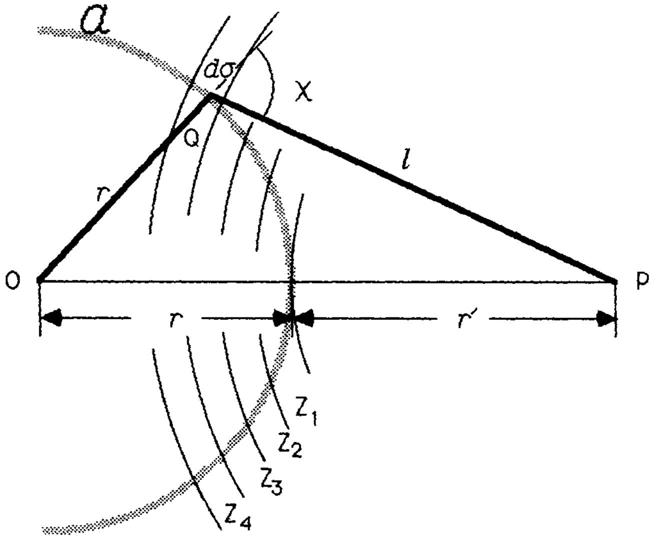


Figure C.6. Fresnel's zone construction.

Comparing Eq. (C-44) with Eq. (C-43), we note that they are equivalent provided that

$$-i\lambda K_1 = 1, \quad \text{or} \quad K_1 = -1/(i\lambda) = i/\lambda = \left[\exp(\pi/2)\right]/\lambda. \qquad \text{(C-45)}$$

The interpretation of this value for K_1 is that the secondary wavelets oscillate a quarter of a period out of phase with the primary wave and that the amplitude of the secondary wave is to the amplitude of the primary wave as 1 is to λ. About sixty years after Fresnel reached these conclusions, Kirchhoff showed that

$$K(\chi) = \left[i/(2\lambda)\right]\left[1 + \cos\chi\right], \qquad \text{(C-46)}$$

which verified Fresnel's K_1: $K(0) = i/\lambda$; but it was not true, as Fresnel assumed, that $K(\pi/2) = 0$.

APPLICATION OF THE HUYGENS-FRESNEL PRINCIPLE TO CHAPTER 4

The optical system discussed in Chapter 4 is shown schematically in Fig. C.7 (duplicate of Fig. 4.1). The point source $\bar{Q}$ is imaged at $\bar{Q}'$ on the image plane. The region of interest in our study of diffraction is the image space, which extends from the exit pupil at $\bar{E}'$ to the image plane at O'. This region is shown in Fig. C.8 (duplicate of Fig. 4.2). The arc at the pupil point $\bar{E}'$ represents the reference spherical surface, which coincides with an aberration-free wave front at the exit pupil. The spherical surface is centered at $\bar{Q}'$ and has a radius $\bar{R}'$ shown from the center to a general point B' on the surface. The general point

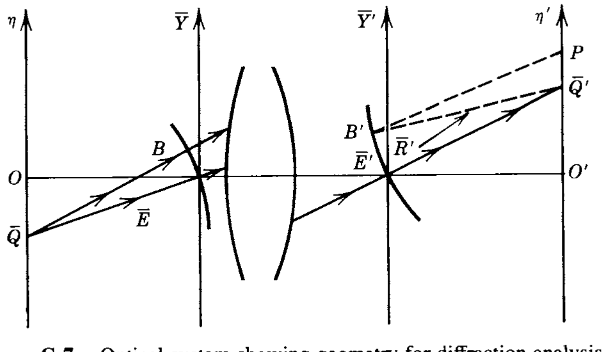


Figure C.7. Optical system showing geometry for diffraction analysis.

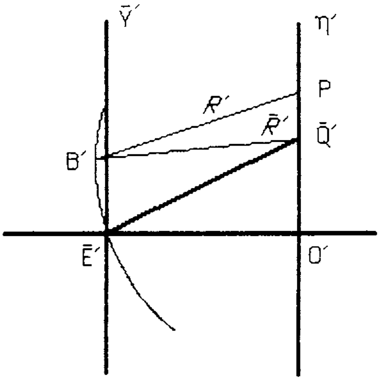


Figure C.8. Optical system image space.

P, at a distance R' from B' and having the coordinates (ξ_0, η_0) with reference to $\bar{Q}'$ as the origin, is in the diffraction pattern. Our present purpose is to develop the expression in Eq. (4-7) for the amplitude $\hat{U}_0(\xi_0, \eta_0)$ at the general point P on the image plane.

The sequence of steps to evaluate $\hat{U}_0$ is similar to the described approach to determine $\hat{U}(P)$ in the previous section: The amplitude is first evaluated at a general point on a spherical wave front; then, by using the wavelet concept, the total contribution of all such point amplitudes to the amplitude at a designated point on the image plane is evaluated. An important difference, however, is that in the present instance the spherical wave front is propagating toward its center; so the amplitude at point B' on its surface is according to Eq. (C-31), which has a positive rather than a negative exponent. With the nomenclature indicated in Fig. C.8, the amplitude expression becomes

$$\hat{U}_0(B') = [G \exp(ikn'\bar{R}')]/\bar{R}', \qquad \text{(C-47)}$$

which is Eq. (4-1) in Chapter 4. Then, proceeding with the steps that correspond to Eqs. (C-41) and (C-42), we find

$$d\hat{U}(P) = \left\{[\hat{U}_0(B')\, K(\chi) \exp(-ikn'R')]/R'\right\} d\sigma, \qquad \text{(C-48)}$$

where $d\sigma$ is an element of the total wave-front area $\mathcal{A}$. Because ξ_0 and η_0 are each extremely small compared with R', the angle χ between the line $B'P$ and the radius line $B'\bar{Q}'$ is practically zero; hence, according to Eq. (C-46),

$$K(0) = i/\lambda. \qquad \text{(C-49)}$$

Then,

$$\hat{U}_0(\xi_0, \eta_0) = \iint_{\mathfrak{a}} \hat{U}(P)\, d\sigma$$

$$= (i/\lambda) \iint_{\mathfrak{a}} \left\{ \left[\hat{U}_0(B') \exp(-ikn'R') \right] / R' \right\} d\sigma. \quad \text{(C-50)}$$

This is the same as Eq. (4-7) in Chapter 4, where considerable manipulation finally produces Eq. (4-22), which is recognized as the *diffraction integral.*

REFERENCES

1. M. Born and E. Wolf, *Principles of Optics.* Pergamon, New York, 1965.
2. C. S. Williams and O. A. Becklund, *Optics: A Short Course for Engineers and Scientists.* Krieger, Malabar, Fla., 1984.
3. S. G. Lipson and H. Lipson, *Optical Physics.* Cambridge Univ. Press, London, 1969.
4. S. Monk, *Light Principles and Experiments*, Dover, New York, 1963.

Appendix D

Updated Calculations

The following tables represent calculations performed by Dr. David F. Edwards, formerly of Lawrence Livermore National Laboratory, as referenced in the Preface to the Reprinted Edition (p. vii).

Table D.1 Optical Transfer Function Values for the Direction $\psi = \pi / 6$

Normalized Frequency	Perfect	For $p = 3s$, $q = 3s$	For $p = 5s$, $q = 5s$
s	Lens	MTF	MTF
0.0	1.0000	1.0000	1.0000
0.1	0.936365	0.864463	0.849491
0.2	0.872889	0.734759	0.703933
0.3	0.809733	0.632737	0.583414
0.4	0.74706	0.559287	0.463263
0.5	0.685038	0.501847	0.338613
0.6	0.623838	0.450883	0.22291
0.7	0.563639	0.404508	0.159816
0.8	0.504632	0.359639	0.190057
0.9	0.447014	0.308483	0.198298
1.0	0.391002	0.249433	0.138789
1.1	0.33683	0.196026	0.0708125
1.2	0.284757	0.163287	0.0541181
1.3	0.235075	0.142588	0.0735196
1.4	0.18812	0.117652	0.0764514
1.5	0.144294	0.0909744	0.052712
1.6	0.104088	0.0713678	0.0314071
1.7	0.0681474	0.0573227	0.0278218
1.8	0.0373861	0.0460639	0.0201757
1.9	0.01332	0.0337247	0.0106097
2.0	0.0000	0.0000	0.0000

Table D.2 Optical Transfer Function Values for the Direction $\psi = \pi / 3$

Normalized Frequency s	Perfect Lens	For $p = 6.5s$, q $= 3s$ MTF	For $p = 10.8s$, $q = 5s$ MTF
0.0	1.0000	1.0000	1.0000
0.1	0.936365	0.83634	0.774362
0.2	0.872889	0.666959	0.519545
0.3	0.809733	0.543534	0.299992
0.4	0.74706	0.458088	0.108394
0.5	0.685038	0.375121	0.188769
0.6	0.563639	0.253577	0.1979
0.7	0.504632	0.107993	0.838718
0.8	0.447014	0.0785467	0.0661051
0.9	0.391002	0.0794094	0.0743255
1.0	0.33683	0.0240422	0.0213139
1.1	0.284757	0.00548988	0.024274
1.2	0.235075	0.0317163	0.0312558
1.3	0.18812	0.0419184	0.01192
1.4	0.144294	0.0141535	0.0366746
1.5	0.104088	0.0303124	0.0192608
1.6	0.0681474	0.0164231	0.0311978
1.7	0.0373861	0.0160886	0.011905
1.8	0.01332	0.026357	0.00171976
1.9	0.0000	0.0181265	0.00852818
2.0	0.0000	0.0000	0.0000

Index